STATISTICAL ISSUES AND DEVELOPMENTS IN TIME SERIES ANALYSIS AND EDUCATIONAL MEASUREMENT

BY

ZHEWEN FAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Jeff A. Douglas, Chair
Assistant Professor Xiaofeng Shao, Co-Director of Research
Assistant Professor Jinmin Zhang
Associate Professor Annie Qu
Professor Emeritus William F. Stout

Abstract

Chapter 1 is concerned with confidence interval construction for the mean of a long-range dependent time series. It is well known that the moving block bootstrap method produces an inconsistent estimator of the distribution of the normalized sample mean when its limiting distribution is not normal. The subsampling method of Hall, Lahiri and Jing (1998) produces a consistent estimator but involves consistent estimation of the variance of the normalized sample mean using one seemingly arbitrary tuning parameter. By adopting a self-normalization idea, we modify the subsampling procedure of Hall *et al.*(1998) and the resulting procedure does not require consistent variance estimation. The modified subsampling procedure only involves the choice of the subsampling widow width, which can be addressed by using some existing data driven selection methods. Simulations studies are conducted to compare the finite sample performances.

The behavior of cluster analysis under different distance measures is explored in Chapter 2, using some of the most common models in educational testing for data generation. Theoretical results on clustering accuracy are given for distance measures used in minimum diameter partitioning and hierarchical agglomerative cluster analysis with complete linkage for data from unidimensional item response models, restricted latent class models for cognitive diagnosis, and the linear factor analysis model. An aim is to identify distance measures that work well for a variety of models, explore how much knowledge of the underlying model is needed to construct a distance measure that leads to a consistent solution, and provide theoretical justifications for using them. Clustering consistency is defined on the space of the latent trait, and consistency and inconsistency results are given for competing distance measures.

We study response times in computerized adaptive testing in Chapter 3. We propose a semi-parametric model for response times that arises in educational assessment data. Algorithms for item selection that use the response time information are proposed and studied for their efficiency and how well they distribute item exposure.

To Mother and Father.

Acknowledgments

This project would not have been possible without the support of many people. Many thanks to my advisers, Jeff A. Douglas and Xiaofeng Shao, who read my numerous revisions and helped make some sense of the confusion. Also thanks to my committee members, Hua-Hua Chang, Jinming Zhang, Annie Qu and William F. Stout who offered guidance and support. Thanks to the Department of Statistics, University of Illinois Graduate College for providing me with the financial means to complete this project. And finally, thanks to my numerous friends who endured this long process with me, always offering support and love.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

SW Subsampling window method.

MSW Modified subsampling window method.

LRD Long Range dependence.

IRT Item Response Theory.

CAT Computer Adaptive Testing. item[HACA] Hierarchical Agglomerative Cluster Analysis.

MDP Minimum Diameter Partitioning.

CDM Cognitive diagnosis model.

DINA Deterministic Input, Noisy Output "AND" gate model.

LFA Linear Factor Analysis.

RT Response Times.

MIC Maximum Information Criterion.

MICT Maximum Information Criterion per Time Unit.

ASB a-Stratification with b-Blocking.

Chapter 1

On the Modified Subsampling Method for Long-Range Dependent Data

1.1 Introduction

Assume the data $(X_1, \cdots, X_n)$ represents a realization from a strictly stationary process $\{X_t\}_{t\in\mathbf{Z}}$, with mean μ and autocovariance function $\gamma_X(k) = \mathrm{cov}(X_t, X_{t+k})$, $k \in \mathbb{Z}$. The process $\{X_t\}$ is said to be long-range dependent (LRD) if $\gamma_X(k)$ satisfies:

$$\gamma_X(k) = k^{-\alpha}L(k), \quad k \to \infty, \tag{1.1}$$

where $0 < \alpha < 1$ and L is slowly varying at infinity, that is $\lim_{x\to\infty} L(\lambda x)/L(x) = 1,$ for any $\lambda > 1$. (See Bingham, Goldie and Teugels (1987)). Alternatively the strength of dependence can be described in terms of the local behavior of the integrable spectral density function $\{f(\lambda),\ |\lambda| \le \pi\}$ around the origin, i.e.,

$$f(\lambda) \sim C_d|\lambda|^{-2d}, \quad \text{as } \lambda \to 0, \tag{1.2}$$

where $d = 1/2 - \alpha/2 \in (0, 1/2)$, $C_d > 0$, and the symbol "$\sim$" means that the ratio of the terms on the two sides converges to one. The long memory phenomenon has been found for times series in various fields [see Beran (1994)]. The study of LRD time series has become a rapidly developing subject and it has diverse applications. In theoretical research, two major categories of LRD processes: transformed Gaussian LRD processes and linear LRD processes have been discussed intensively; See for example Davydov (1970), Taqqu (1975, 1979) and Dobrushin and Major (1979) for the early works on these processes.

Inference of the mean μ is often the first and an important step in the analysis of stationary time series. When the data is weakly dependent, the moving block bootstrap approach developed by Künsch (1989) and Liu and Singh (1992) provides consistent nonparametric estimate for the distribution of the sample mean. When the time series is LRD, Lahiri (1993) showed that the block bootstrap method fails to provide a consistent estimator for the class of transformed-Gaussian LRD process, whose normalized sample mean can have a non-normal limiting distribution. To remedy this problem, Hall, Jing and Lahiri (1998) developed

a block based sampling window procedure (i.e., subsampling; Politis and Romano (1994)) and proved its consistency for the transformed Gaussian LRD processes. Nordman and Lahiri (2005) further extended the theoretical validity of the subsampling method to linear LRD processes.

In Hall *et al.* (1998) [also see Nordman and Lahiri (2005)], the procedure involves consistent estimation of asymptotic variance of the normalized sampling mean, which depends on several unknown parameters. A nonparametric method was used and a tuning parameter needs to be selected. No sound guidance seems provided as to the choice of this particular tuning parameter. In this paper, we propose to extend the self-normalization idea in Lobato (2001) to this setting. We first form a self-normalized statistic, whose asymptotic distribution is free of the asymptotic variance of the sample mean, then apply the subsampling method. One advantage of our approach is that we do not need to consistently estimate the variance of sample mean and therefore we reduce the number of the tuning parameters involved in the subsampling procedure. The only tuning parameter in our modified scheme is the subsampling window width, the choice of which can be addressed using the existing techniques; see e.g., Politis, Romano and Wolf (1999, Chapter 9). We do not address the consistency of the data driven subsampling widow width but will investigate its finite sample performance in Section 1.3.2.

The rest of the chapter is organized as follows. Sections 2.1 and 2.2 detail some theoretical results of transformed Gaussian processes and linear processes, respectively. Section 2.3 reviews Hall *et al.* (1998)'s subsampling method and then proposes our modified subsampling window procedure. In Section 3, we report several simulation studies on the coverage accuracies as well as the interval widths of our modified subsampling procedure for the LRD Gaussian and linear process mean and compare them with the subsampling method of Hall *et al.* (1998). Section 4 concludes the paper.

1.2 Methodology

1.2.1 Transformed Gaussian LRD Processes

Let $\{Z_t,\ t \in \mathbb{Z}\}$ be a strictly stationary Gaussian process with $EZ_1 = 0$, $EZ_1^2 = 1$ and $\gamma_Z(k) = \mathrm{cov}(Z_1, Z_{1+k}), k \in \mathbb{Z}$. The stationary transformed Gaussian process $\{X_t\}_{t\in\mathbb{Z}}$ is defined as $X_t = G(Z_t)$, where $G : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function that satisfies $E\{G(Z_1)^2\} < \infty$. In order to make this case distinct from that of the linear LRD process considered in Section 1.2.2 below, we suppose that G is not an affine function. Let $\mu = EX_1$ be the parameter of interest and let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ denote the sample mean. To construct a confidence interval for μ, it is natural to consider the asymptotic behavior of $(\bar{X}_n - \mu)$. To this end, we

introduce the k^{th} Hermite polynomial $H_k(x)$, which is defined by

$$H_k(x) = (-1)^k \exp(x^2/2)(d^k/dx^k)(\exp(-x^2/2)), \quad x \in \mathbb{R}. \tag{1.3}$$

Denote by $h_k = E[G(Z)H_k(Z)]$ the Hermite coefficients. Then the Hermite rank q of $G(\cdot)$ is defined as $q = \inf\{k \geq 1 : h_k \neq 0\}$. The asymptotic distribution of $\bar{X}_n - \mu$ depends on q and the autocovariance function $\gamma_X(k)$, which is assumed to be slowly varying at infinity, i.e., it satisfies (1.1). The following theorem by Taqqu (1975, 1979), Dobrushin and Major (1979) characterizes the asymptotic distribution of $\bar{X}_n - \mu$:

Theorem 1.2.1. *Assume that γ_Z admits the representation at (1.1) and that G has Hermite rank q, where $0 < \alpha < q^{-1}$. Then,*

$$n(\bar{X}_n - \mu)/d_n \xrightarrow{d} W_q, \text{ as } n \to \infty, \tag{1.4}$$

where $d_n = \{n^{2-q\alpha}L^q(n)\}^{1/2}$ and "$\xrightarrow{d}$" denotes convergence in distribution. The limiting distribution W_q is defined in terms of a multiple Wiener-Ito integral with respect to the random spectral measure W of the Gaussian white-noise process as

$$W_q = \frac{C_q}{A^{q/2}} \int \frac{exp\{i(x_1 + \cdots + x_q)\} - 1}{i(x_1 + \cdots + x_q)} \prod_{k=1}^{q} |x_k|^{\alpha/2} dW(x_1) \cdots dW(x_q), \tag{1.5}$$

where $A = 2\Gamma(\alpha)\cos(\alpha\pi/2)$ and $C_q = E\{H_q(Z_1)G(Z_1)\}/q!$.

Note that W_q has a normal distribution with mean zero and variance $2C_1^2/\{(1-\alpha)(2-\alpha)\}$ when $q = 1$, and W_q has a non-normal distribution when $q \geq 2$.

1.2.2 Linear LRD Processes

Another widely used class of processes that allows for long-range dependence is the so-called linear processes. Let $\{\epsilon_t,\ t \in \mathbb{Z}\}$ be a sequence of stationary innovations with zero mean and finite second moment, and $\{a_k,\ k \in \mathbb{Z}\}$ be a sequence satisfying $\sum_{k\in\mathbb{Z}} a_k^2 < \infty$. Let

$$X_t = \mu + \sum_{k=-\infty}^{\infty} a_{t-k}\epsilon_k. \tag{1.6}$$

Then the autocovariance function of X_t also admits the representation as in (1.1) for certain $\{a_k\}$. In practice, it is usually not plausible that X_t should depend on future values of ϵ_t. The summation in (3.13)

is therefore often restricted to $k \geq 0$. That is, we assume the casuality of X_t:

$$X_t = \mu + \sum_{k=0}^{\infty} a_{t-k}\epsilon_k. \tag{1.7}$$

The well-known fractional autoregressive integrated moving average (FARIMA) model [c.f Adenstedt (1974), Granger and Joyeux (1980), and Hosking (1981)] has the representation (1.7).

Under suitable regularity conditions, Davydov (1970) showed that

$$n(\bar{X}_n - \mu)/d_n \xrightarrow{d} Z, \text{ as } n \to \infty, \tag{1.8}$$

where Z is a standard normal random variable. For linear LRD processes, the correct scaling $d_n = \{n^{2-\alpha}L(n)\}^{1/2}$ also depends on the unknown quantities α and $L(n)$.

1.2.3 Subsampling and a Modified Version

To construct a confidence interval for μ, a standard approach is to resort to asymptotic distribution and use the corresponding critical values. As seen from the results in the above discussions, the asymptotic distribution of the studentized sample mean can admit a complicated form and/or involve unknown nuisance parameters, the estimation of which is nontrivial. One way out is to use resampling methods, such as the moving block bootstrap. However, for transformed Gaussian processes, Lahiri (1993) showed the inconsistency for the moving block bootstrap when W_q is nonnormal. As a remedy, Hall *et al.* (1998) proposed to use the so-called subsampling window method (SW). Specifically, let $T_n = n(\bar{X}_n - \mu)/d_n$ be the normalized sample mean and denote its corresponding cumulative distribution function by

$$J_n(x) = P(T_n \leq x), \quad x \in \mathbb{R}.$$

The subsampling method consists of the following steps:

Step 1 Partition the time series into $N = n - b + 1$ consecutive overlapping blocks of length b, $\mathbf{B_k} = \{X_k, \cdots, X_{k+b-1}\}$, $k = 1, \cdots, N$. Each data block $\mathbf{B_k}, k = 1, \cdots, N$ is treated as a scaled-down replicate of the original data $\{X_1, \cdots, X_n\}$.

Step 2 For each block $\mathbf{B_k}$, we calculate the subsampled version of T_n and denote it as

$$T_{b,k} = \frac{\sum_{i=k}^{k+b-1} X_i - b\bar{X}_n}{d_b}, \ k = 1, \cdots, N.$$

Step 3 The subsampling estimator of $J_n(x)$ is formed as

$$\hat{J}_{n,b}(x) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}(T_{b,k} \le x).$$

where $\mathbf{1}(\cdot)$ stands for the indication function.

Note that the scaling parameters d_n and d_b depend on the unknown quantities α and $L(\cdot)$. To implement the above procedure, Hall *et al.*(1998) proposed to obtain consistent estimates of d_n and d_b, denoted by $\hat{d}_n$ and $\hat{d}_b$ respectively, by a nonparametric approach. Specifically, let $m_{1n}, m_{2n} \in [1, n]$ denote integers such that for some $\theta \in (0,1)$, $m_{1n} = n^{(1+\theta)/2}$ and $m_{2n} = n^{\theta}$. Define $\tilde{d}_m^2 = (n-m+1)^{-1}\sum_{i=1}^{n-m+1}(\sum_{j=i}^{i+m-1} X_j - m\bar{X}_n)^2$, for $m \in [1, n]$. Then $\hat{d}_n^2 = \tilde{d}_{m_{1n}}^4/\tilde{d}_{m_{2n}}^2$. Similarly, $\hat{d}_b^2$ is calculated analogously based on each subsample. Let $\hat{T}_n = n(\bar{X}_n - \mu)/\hat{d}_n$ and $\hat{T}_{b,k} = (\sum_{i=k}^{k+b-1} X_i - b\bar{X}_n)/\hat{d}_b$. The subsampling estimator of $J_{1,n}(x) = P(\hat{T}_n \le x)$ is given by

$$\hat{J}_{1,n,b}(x) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}(\hat{T}_{b,k} \le x).$$

In practice, the empirical coverage can be sensitive to the choice of θ, but no guidance seems available about the choice of θ. In this article, we propose a modified subsampling window procedure that does not involve consistent estimates of d_n and d_b. Instead of using a consistent estimator of the scale as the studentizer [Hall et al. 1998], we use an inconsistent estimator, which does not involve any tuning parameter. It is worth noting that the idea of using inconsistent studentizer in the subsampling method has been justified in Theorem 11.3.1. of Politis et al. (1999) under some general conditions.

The key ingredient of our proposal is to extend the self-normalization method in Lobato (2001) to the confidence interval construction for μ. The self-normalization idea has been used in Lobato (2001) and Shao (2009, 2010) to construct confidence intervals for quantities associated with a weakly dependent time series. The extension to the LRD time series seems new. For the sake of readership, we illustrate the self-normalization idea in the short-range dependence case. Let $\{Y_t,\ t = 1, \cdots, n\}$ be a weakly dependent stationary process with mean μ_Y and denote by $\bar{Y}_n = n^{-1}\sum_{t=1}^{n} Y_t$ the sample mean. Under some regularity conditions, we have

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor nr \rfloor}(Y_t - \mu_Y) \Rightarrow \sigma B(r), \text{ for some } \sigma > 0,$$

where $\lfloor nr \rfloor$ denotes the integer part of nr and " $\Rightarrow$ " stands for the weak convergence in $\mathcal{D}[0,1]$. Here $\mathcal{D}[0,1]$ deontes the space of functions on $[0,1]$ which are right continuous and have left-hand limits, endowed with the Skorohod topology [Billingsley (1968)], $B(\cdot)$ stands for Brownian motion. Then by the continuous

mapping theorem,

$$\frac{\sqrt{n}(\bar{Y}_n - \mu_Y)}{\sqrt{n^{-2}\sum_{t=1}^{n}\{\sum_{j=1}^{t}(Y_j - \bar{Y}_n)\}^2}} \xrightarrow{d} \frac{B(1)}{\sqrt{\int_0^1 \{B(r) - rB(1)\}^2 dr}} \equiv U_0. \tag{1.9}$$

The limiting distribution U_0 is non-standard, but its critical values have been tabulated by Lobato (2001) via simulations. As mentioned in Shao (2010), the self-normalization method is a special case of the so-called fixed-b approach [Kiefer and Vogelsang (2005)], since the normalizer $n^{-2}\sum_{t=1}^{n}(\sum_{j=1}^{t} Y_j - t\bar{Y}_n)^2$ is the lag-window estimator of the long run variance σ^2 when $K(\cdot)$ is taken to be the Bartlett Kernel and the bandwidth is equal to the sample size (i.e., $b = 1$). It was brought to our attention by a referee that McElroy and Politis (2009) have extended the fixed-b approach to the long memory case. Their results show that the fixed-b limiting distribution of the studentized sample mean depends on the kernel, the magnitude of memory, and the taper. However, in contrast to the paper at hand, they do not consider a subsampling distribution estimator and also do not provide a practical method for forming confidence intervals.

To extend the self-normalization idea to the long memory case, the key tool we need is the functional central limit theorem. For the transformed Gaussian processes, it has been proved by Taqqu (1975). Under certain assumptions [cf. Theorems 2.1 and 4.1 of Taqqu (1975)], we have that by reduction principle $\frac{1}{d_n}\sum_{j=1}^{\lfloor nr \rfloor}(X_j - \mu)$ and $\frac{h_q}{q! d_n}\sum_{i=1}^{\lfloor nr \rfloor} H_q(X_i)$ converge to the same limiting process $E_q(r)$, i.e.,

$$\frac{\sum_{j=1}^{\lfloor nr \rfloor}(X_j - \mu)}{d_n} \Rightarrow E_q(r), \quad r \in [0,1]. \tag{1.10}$$

Note that $E_q(r)$ is the fractional Brownian motion process when $q = 1$ and that $E_q(1)$ has the same distribution as W_q. When $q = 2$, $E_q(r)$ is the so-called non-Gaussian Rosenblatt process.

Let $S_{i,j} = \sum_{t=i}^{j} X_t$, and 0 if $i > j$. By (1.10) and the continuous mapping theorem, we have

$$K_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{n^{-2}\sum_{t=1}^{n}(S_{1,t} - t\bar{X}_n)^2}} \xrightarrow{d} \frac{E_q(1)}{\sqrt{\int_0^1 \{E_q(r) - rE_q(1)\}^2 dr}} =: M_q. \tag{1.11}$$

In general, the limiting distribution M_q is unknown, so we cannot directly use (1.11) to construct a confidence interval for μ. Instead, we propose to utilize the subsampling approach to approximate the sampling distribution of K_n. Let $F_n(x) = P(K_n \leq x)$, $x \in \mathbb{R}$ and $\hat{F}_n(x)$ be its subsampling based estimator. For the i^{th} block $\mathbf{B}_i$, $i = 1, 2, ...N$, we calculate the within block mean $\bar{X}_{i,b} = b^{-1}\sum_{j=i}^{i+b-1} X_j$ and define the

subsampling counterpart of K_n as

$$K_{b,i} = \frac{\sqrt{b}(\bar{X}_{i,b} - \bar{X}_n)}{\sqrt{b^{-2}\sum_{j=i}^{i+b-1}\{S_{i,j} - (j-i+1)\bar{X}_{i,b}\}^2}}, \quad i = 1, \cdots, N. \tag{1.12}$$

Then the subsampling estimator of F_n is given by

$$\hat{F}_n(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{K_{b,i} \le x\}. \tag{1.13}$$

Let $V_n = n^{-3/2}\{\sum_{j=1}^{n}(S_{1,j} - j\bar{X}_n)^2\}^{1/2}$ and $\hat{t}_{\beta,n}$ be the $\lfloor N\beta \rfloor^{\text{th}}$ order statistic of $K_{b,k},\ 1 \le k \le N$. The two-sided lower and upper $100(1-\beta)\%$ confidence bounds for μ are constructed as $L_{\beta/2,n} = \bar{X}_n - V_n\hat{t}_{1-\beta/2,n}$ and $U_{1-\beta/2,n} = \bar{X}_n - V_n\hat{t}_{\beta/2,n}$ respectively.

For linear processes (1.7), the following functional central limit theorem has been established under appropriate moment and weakly dependent conditions on $\{\epsilon_t\}$ [see Wu and Shao (2006) and the references therein]:

$$\frac{\sum_{j=1}^{\lfloor nr \rfloor}(X_j - \mu)}{d_n} \Rightarrow \sigma B_d(r), \quad r \in [0,1], \tag{1.14}$$

where $B_d(\cdot)$ is the fractional Brownian motion and σ is a constant. By the continuous mapping theorem, the distribution of K_n in (1.11) now becomes

$$K_n \xrightarrow{d} \frac{B_d(1)}{\sqrt{\int_0^1 \{B_d(r) - rB_d(1)\}^2 dr}} \equiv U_d, \quad d \in (0, 1/2). \tag{1.15}$$

The limiting distribution U_d depends on the unknown long memory parameter d. We shall also use the modified subsampling method to approximate the sampling distribution of K_n.

Nordman *et al.* (2007) developed a blockwise empirical likelihood method (EL) to construct a confidence interval for the mean of the linear LRD process. The EL method also requires a consistent estimation of long memory parameter d. Consequently, both the EL and SW methods require two tuning parameters: subsampling width or block size b and a user-dependent parameter related to variance estimation or estimation of d. By contrast, our method only involves the choice of the subsampling width. In practice, given a long memory time series, it is hard to judge if the series is from a LRD linear process or from a transformed Gaussian process. Thus it seems desirable that our method does not rely on the assumption of linear processes.

It would be interesting to provide a consistency result for the modified subsampling method proposed here, but the proofs seem very nontrivial for either transformed Gaussian processes or linear processes. In

the case of Gaussian processes, it is actually possible to mimic the argument presented in Hall et al. (1998) to show that the modified subsampling scheme is consistent. However, as pointed out by a referee, the completely regularity assumption assumed for the LRD Gaussian processes, can not hold, as a completely regular Gaussian process can not have a pole in the spectral density. For linear processes, Nordman and Lahiri (2005) proved the consistency of the subsampling window method by Hall *et al.* (1998). It can be expected that our modified subsampling window method is consistent for linear processes. But a rigorous proof of the consistency seems difficult and is left for future research. Instead, we investigate the finite sample performance through simulations.

1.3 Simulation Studies

To evaluate the performance of the modified subsampling method, we divide our simulation studies into two parts according to the data generating processes: transformed Gaussian processes and linear processes. To generate the data $\mathbf{X}_n \equiv \{X_1, \cdots, X_n\}$ that is a realization of a transformed Gaussian process with self similarity parameter α, (or Hurst parameter $H = 1/2(2-\alpha)$) and Hermite rank q, we follow the procedure described in Hall *et al.* (1998). Let $\mathbf{Z}_{n0} = \{Z_{10}, \cdots, Z_{n0}\}$ be size n iid standard normal random variables, and $R = (r_{ij})$ be the correlation matrix with $(i,j)^{th}$ entry defined as

$$r_{ij} = \frac{1}{2}\{(|j-i|+1)^{2H} + ||j-i|-1|^{2H} - 2|j-i|^{2H}\}, \quad i,j = 1, \cdots, n \tag{1.16}$$

for $\frac{1}{2} < H < 1$. Then $\mathbf{Z}_n \equiv \{Z_1, \cdots, Z_n\} = U^T \mathbf{Z}_{n0}$ is a LRD process with $\alpha = 2 - 2H \in (0,1)$ (cf. Beran (1994), p.50), where U is obtained by the Cholesky decomposition, i.e., $R = U^T U$. We then take $X_i = H_q(Z_i),\ i = 1, \cdots, N$ to get the transformed Gaussian process with Hermite rank q.

To generate data $\mathbf{X}_n \equiv \{X_1, \cdots, X_n\}$ that is a realization of a LRD linear process, we consider the commonly used FARIMA model and let $\tilde{\mathbf{X}}_n \equiv \{\tilde{X}_1, \cdots, \tilde{X}_n\}$ represent an FARIMA$(0,d,0)$ process, where $d = 1/2(1-\alpha) \in (0, 1/2)$. Then the series $\mathbf{X_n}$ is generated via

$$X_t = \varphi X_{t-1} + \tilde{X}_t + \vartheta \tilde{X}_{t-1},\ t = 1, \cdots, n \tag{1.17}$$

by combining one of the following ARMA filters, α values and innovation distributions:

- $\varphi = 0.7, \vartheta = -0.3$ (Filter1); $\varphi = -0.7, \vartheta = 0.3$ (Filter2); $\varphi = \vartheta = 0$ (Filter3);
- $d = 0.45, 0.25, 0.05$ (i.e., $\alpha = 0.1, 0.5, 0.9$).

- $\{\epsilon_t\}$ as standard normal; $\chi_1^2 - 1$; or t_3.

The experimental design and the results are described next.

1.3.1 Modified subsampling vs. subsampling window method for both transformed Gaussian and linear processes

To facilitate a comparison between our modified subsampling method and the subsampling window method, we report coverage probabilities of 90% two-sided confidence intervals as well as the average lengths of the confidence intervals for both the transformed Gaussian processes and linear processes. For the transformed Gaussian processes, we use block sizes $b = Cn^{1/2}$, $C \in \{0.5, 1, 3, 6\}$, $n \in \{400, 1000\}$, $q \in \{1, 2, 3\}$, $H \in \{0.95, 0.75, 0.55\}$ and $\theta = 0.8$. The subsampling widths b are overall shorter than those considered in Hall *et al.* (1998), where the coverage probabilities are significantly lower with large C values (e.g., 6,9) than with small C values. Tables 1 and 2 provide the empirical coverage probabilities of the 90% two-sided confidence intervals for transformed Gaussian processes, with the average lengths of the confidence intervals over $1,000$ simulation runs for each configuration reported in parenthesis. For linear LRD series, Tables 3, 4 and 5 correspond to FARIMA series with normal, $\chi^2 - 1$ and t_3 innovations with block sizes $b = Cn^{1/2}$, $C \in \{0.5, 1, 2\}$, $n \in \{400, 1000\}$, $q \in \{1, 2, 3\}$, filter $\in 1, 2, 3$, $H \in \{0.95, 0.75, 0.55\}$ and $\theta = 0.8$.

We summarize the major findings as follows:

1. Overall our method (MSW) is comparable to Hall *et al* (1998)'s method (SW) in terms of coverage accuracies. It can be seen that MSW method yields lower empirical coverage probabilities but shorter confidence intervals than SW. With big block sizes (e.g., $C = 3$ or 6) and strong degree of dependence (e.g., $\alpha = 0.1$), both SW and MSW experience severe under-coverage. It is not surprising because the data-driven optimal block selection study (see Tables 6 and 7) shows that the optimal bandwidths never exceed $\sqrt{n}$, where n denotes the sample size.

2. For linear LRD processes, both SW and MSW perform similarly across the different types of innovations with the same sample sizes and block sizes. Generally, the coverage accuracy improves as the sample size increases or the strength of dependence decreases (i.e., when α increases), although we see that in a few cases the coverage can get worse, e.g. with more severe overcoverage. This could be attributed to the randomness in our simulation based on 1000 replications.

Figures 1(a) - 1(d) compare the empirical coverage probabilities between SW and MSW methods for both transformed Gaussian and linear processes with various innovations across a range of b. For the transformed Gaussian process, we simulate a size $n = 400$, $H = 0.75$ and $q = 1$ sequence and subsampling window sizes

vary from 3 to 40. For linear process, we generate a size $n = 400$, $H = 0.75$, $q = 1$ sequence and pass through Filter 2 and subsampling window sizes vary from 3 to 40. We report both the coverage probability and the average lengths of the confidence interval. As seen from Figures 1(a)-1(d), the coverage probabilities of both SW and MSW methods decrease when the block size increases. The widths of the confidence intervals also decrease when block size increases. MSW method tends to have lower coverage probabilities than SW method, but has shorter confidence interval widths.

1.3.2 Data driven block size selection method

As shown in Tables 1-5, the coverage probabilities of SW and MSW heavily depend on the choice of the block size b, so the selection of the block size b is a crucial issue. Here we employ a data driven block size selection procedure proposed in Bickel and Sakov (2008) for the m out of n bootstrap, which is closely related to the subsampling method. The use of Bickel and Sakov's automatic bandwidth selection in the subsampling context seems unexplored before. Note that there are other available methods, such as those in Politis, Romano and Wolf (1999) and Götze and Račkauskas (2001). The procedure consists of the following steps:

Step 1 Consider a sequence of b's of the form

$$b_j = [g^j n], \text{ for } j = 1, 2, \cdots, J, \quad 0 < g < 1, \tag{1.18}$$

where $[\alpha]$ denotes the smallest integer larger than α.

Step 2 For each b_j, find $\hat{F}_{b_j,n}$, where $\hat{F}_{b,n}$ is the subsampling based distribution estimator for a given subsampling width b.

Step 3 Let $\rho(F, G) = \sup_x |F(x) - G(x)|$, and set

$$j_0 = \text{argmin}_{j=1,\cdots,J-1} \; \rho\left(\hat{F}_{b_j,n}, \hat{F}_{b_{j+1},n}\right). \tag{1.19}$$

Then the optimal block size is b_{j_0}. If the difference is minimized for a few values of b_j, then pick the largest among them.

For the transformed Gaussian processes, we use sample sizes $n \in \{400, 1000\}$, $q \in \{1, 2, 3\}$, $H \in \{0.95, 0.75, 0.55\}$ and $\theta = 0.8$. The data driven block size selection parameter g in equation (1.18) is set to be 0.75. Table 6 provides the empirical coverage probabilities of the 90% two-sided confidence intervals for

transformed Gaussian processes, with the average lengths of the confidence intervals over $1,000$ simulation runs for each configuration reported in parenthesis. Also listed in parenthesis are the mean and the median (optimal) block sizes over $1,000$ simulation runs for each parameter combination. For linear LRD series, Table 7 corresponds to FARIMA series with a standard normal innovation with sample sizes $n \in \{400, 1000\}$, $q \in \{1, 2, 3\}$, Filter$1, 2, 3$, $H \in \{0.95, 0.75, 0.55\}$ and $\theta = 0.8$.

We summarize several findings from the simulation results in Tables 6 and 7: 1. For both transformed Gaussian processes and linear processes with a normal innovation, MSW produces narrower confidence intervals and less empirical coverage (except at $n = 1000$, linear process with normal innovation and filter 1) than SW does; 2. In the case of transformed Gaussian processes, when data are strongly dependent, i.e., $\alpha = 0.1$, the coverage probabilities for SW is closer to the nominal coverage 90% than MSW when $q = 1, 2$. However, with moderate or low degree of long range dependence ($\alpha = 0.5$ or $\alpha = 0.9$), MSW in general achieves the empirical coverage probabilities that are closer to the 90% nominal level. In the case of the linear LRD processes with normal innovation, SW outperforms MSW in terms of coverage probabilities for the second filter. Except for that, the performance for SW and MSW are comparable in terms of the coverage; 3. When $n = 1000$, the average mean and median optimal block sizes corresponding to MSW are in general shorter than that for SW. We also tried $g = 0.6$ and qualitatively similar results have been found.

1.4 Conclusions

In summary, we propose a modified subsampling method to construct a confidence interval for the mean of a LRD time series. Compared to the existing subsampling window method of Hall *et al.* (1998), our method eliminates the need to consistently estimate the variance of the sample mean, which involves one rather arbitrary tuning parameter θ. Through simulations we find that our method is comparable to subsampling window method in terms of coverage accuracies. Hall *et al.* (1998)'s subsampling window method tends to have higher empirical coverage probabilities than ours, but often our modified subsampling method yields the coverage probabilities that are closer to the nominal level. In addition, our modified subsampling method delivers shorter confidence intervals.

1.4.1 Simulation results

Table 1.1: Empirical coverage probabilities for transformed Gaussian processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method. Sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.1, 0.5, 0.9$, block sizes $b = Cn^{1/2}$ with constants $C = 0.5, 1$. Entry in the parenthesis () represents the average length of the intervals. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			$b = 0.5n^{1/2}$			$b = n^{1/2}$		
n	q	method	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
400	1	SW	78.8	91.4	92.8	79.5	91.5	93.6
			(3)	(1.4)	(0.4)	(3.3)	(1.5)	(0.5)
		MSW	69.8	86.2	89.9	65	84.3	88
			(2.1)	(0.9)	(0.3)	(1.9)	(0.9)	(0.3)
	2	SW	88.3	95.7	92.5	84.7	95.1	93.9
			(5.6)	(1.1)	(0.7)	(5.5)	(1)	(0.6)
		MSW	80.5	92.2	87.7	74.8	89.7	86.1
			(3.9)	(0.7)	(0.4)	(3.4)	(0.6)	(0.3)
	3	SW	91.5	96.2	91.7	89.3	95.1	91.7
			(8.8)	(1.1)	(0.8)	(8.4)	(1.1)	(0.8)
		MSW	88.9	92	83.5	82.5	88.3	81.3
			(6)	(0.7)	(0.5)	(5.1)	(0.6)	(0.5)
1000	1	SW	86	94.8	96	85.1	94.3	95
			(4.1)	(1.5)	(0.4)	(3.8)	(1.5)	(0.4)
		MSW	74.1	88.2	90.8	69.8	86.6	88.5
			(2.2)	(0.8)	(0.2)	(1.9)	(0.7)	(0.2)
	2	SW	90.8	97.2	97.3	88.2	96.1	96.6
			(7.1)	(0.9)	(0.5)	(6.2)	(0.8)	(0.4)
		MSW	79.4	93.9	92.2	72.5	91.5	90.1
			(3.7)	(0.5)	(0.3)	(3.1)	(0.4)	(0.2)
	3	SW	94.7	97.5	96.1	92.4	96.7	94.9
			(10)	(0.9)	(0.7)	(8.8)	(0.8)	(0.6)
		MSW	89.7	93.8	88.3	84	89.6	84.6
			(5)	(0.5)	(0.3)	(4.2)	(0.4)	(0.3)

Table 1.2: Empirical coverage probabilities for transformed Gaussian processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method. Sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.1, 0.5, 0.9$, block sizes $b = Cn^{1/2}$ with $C = 3, 6$. Entry in the parenthesis () represents the average length of the intervals. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			$b = 3n^{1/2}$			$b = 6n^{1/2}$		
n	q	method	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
400	1	SW	68	84.3	87.6	54.6	71	78.8
			(2.5)	(1.3)	(0.4)	(1.8)	(1)	(0.3)
		MSW	56.5	75.6	83.6	45.4	66.1	74.2
			(1.6)	(0.8)	(0.3)	(1.1)	(0.6)	(0.2)
	2	SW	72.1	88.3	86.5	59.6	76.1	77
			(4)	(0.8)	(0.4)	(2.6)	(0.6)	(0.4)
		MSW	63.2	81.7	79.7	51	69.7	70.9
			(2.6)	(0.5)	(0.3)	(1.8)	(0.4)	(0.2)
	3	SW	78.4	89.8	83.5	66.7	77.7	73.7
			(5.9)	(0.9)	(0.7)	(3.9)	(0.7)	(0.6)
		MSW	70.6	82.1	74.7	66.7	71.9	67
			(3.8)	(0.6)	(0.5)	(3.9)	(0.5)	(0.4)
1000	1	SW	71	88	89.5	59.6	80.5	82.8
			(2.6)	(1.1)	(0.3)	(2)	(0.9)	(0.2)
		MSW	59.4	79.6	85.8	50.5	73.3	78.4
			(1.6)	(0.7)	(0.2)	(1.3)	(0.6)	(0.2)
	2	SW	77.6	91.3	91.9	64.6	85.3	84.6
			(4.2)	(0.5)	(0.3)	(2.9)	(0.4)	(0.2)
		MSW	64.8	87.1	87.4	56.3	81.7	81.1
			(2.5)	(0.3)	(0.2)	(1.9)	(0.3)	(0.2)
	3	SW	82.3	91.5	87.9	72.9	85.1	81.8
			(5.9)	(0.6)	(0.5)	(4)	(0.5)	(0.4)
		MSW	74.6	83.8	81.1	66.8	79.1	77
			(3.4)	(0.4)	(0.3)	(2.6)	(0.3)	(0.3)

Table 1.3: Empirical coverage probabilities for linear processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method, with standard normal, $\chi^2 - 1$ and t_3 innovations, filters: $\varphi = 0.7, \vartheta = -0.3$ (Filter1);$\varphi = -0.7, \vartheta = 0.3$ (Filter2);$\varphi = \vartheta = 0$ (Filter3); sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.1$, block sizes $b = Cn^{1/2}$ with constants $C = 0.5, 1, 2$. Entry in the parenthesis () represents the average length of the intervals. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			Standard Normal Innovations			Chi-Square Innovations			t Innovations		
n	Filter	Method	$C = 0.5$	$C = 1$	$C = 2$	$C = 0.5$	$C = 1$	$C = 2$	$C = 0.5$	$C = 1$	$C = 2$
$\alpha = 0.1$											
400	1	SW	94.9	92.1	87.4	94.1	92.1	87.1	93.6	90.5	84.4
			(19)	(17)	(13.5)	(31.3)	(27.5)	(21)	(35.3)	(29.8)	(23.3)
		MSW	93.1	86.8	78.4	93.3	86.9	79.6	92.9	87.2	77.9
			(14)	(10.7)	(8.4)	(22.5)	(16.9)	(12.9)	(27)	(19.8)	(14.9)
	2	SW	85.6	87.1	83.6	87.5	88.1	84.2	86.3	87.2	83
			(3.6)	(4.2)	(4)	(6.1)	(6.4)	(5.6)	(6.6)	(7.5)	(6.8)
		MSW	80.6	80.4	76	81.9	79.9	74.5	81.3	79.5	74.3
			(2.6)	(2.7)	(2.5)	(4.3)	(4)	(3.5)	(4.8)	(4.8)	(4.4)
	3	SW	88.5	89.3	84	90.4	89.5	84.5	88.5	88	83.5
			(5.3)	(5.9)	(5.1)	(9.3)	(9.5)	(7.8)	(10)	(10.6)	(9.2)
		MSW	81.8	79.5	75.3	84.5	81.1	75.4	83.5	79.4	74
			(3.7)	(3.6)	(3.3)	(6.3)	(5.6)	(4.8)	(6.9)	(6.4)	(5.7)
1000	1	SW	97	93.5	90.1	97.4	93.8	90.2	96.7	93.1	88.1
			(25.4)	(17.6)	(14.5)	(37.9)	(25.4)	(20.4)	(47.3)	(32.2)	(25.8)
		MSW	94.7	88.9	81.5	93.7	89.6	83.7	93.3	88.6	82.8
			(13.5)	(10.6)	(8.9)	(19.5)	(14.8)	(12)	(24.8)	(19)	(15.6)
	2	SW	93.1	91	87.6	93.2	90.9	87.4	93.9	90.2	85.9
			(5.1)	(4.5)	(4.2)	(8.3)	(6.7)	(5.9)	(10.1)	(8.5)	(7.6)
		MSW	83.3	81.8	78.6	86.8	85	80.8	85.4	82.5	79.3
			(2.6)	(2.7)	(2.5)	(4.3)	(4)	(3.6)	(5.2)	(5)	(4.6)
	3	SW	95	92.1	89.1	95.2	91.8	88.7	94.5	90.5	87.8
			(7.9)	(6.3)	(5.6)	(13.1)	(9.7)	(8.4)	(14.4)	(11.1)	(9.8)
		MSW	87.4	85	81.1	87.6	85.2	81.1	88.3	84.5	80.7
			(3.9)	(3.7)	(3.4)	(6.2)	(5.5)	(5)	(7)	(6.5)	(5.8)

Table 1.4: Empirical coverage probabilities for linear processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method, with standard normal, $\chi^2 - 1$ and t_3 innovations, filters: $\varphi = 0.7, \vartheta = -0.3$ (Filter1);$\varphi = -0.7, \vartheta = 0.3$ (Filter2);$\varphi = \vartheta = 0$ (Filter3); sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.5$, block sizes $b = Cn^{1/2}$ with constants $C = 0.5, 1, 2$. Entry in the parenthesis () represents the average length of the intervals. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			Standard Normal Innovations			Chi-Square Innovations			t Innovations		
n	Filter	Method	$C = 0.5$	$C = 1$	$C = 2$	$C = 0.5$	$C = 1$	$C = 2$	$C = 0.5$	$C = 1$	$C = 2$
$\alpha = 0.5$											
400	1	SW	96.3	95.7	92	95.9	93.7	90.5	95.4	93.9	90.8
			(5.2)	(5)	(4)	(8.8)	(7.7)	(6)	(9.2)	(8.6)	(7)
		MSW	96.1	91.9	84.8	94.7	89.4	83.2	95.1	90.1	83.5
			(3.7)	(3)	(2.4)	(6.1)	(4.6)	(3.5)	(6.4)	(5)	(4.2)
	2	SW	86.9	87.7	86.9	90.8	91.6	89.3	86.5	89.4	86.1
			(0.9)	(1)	(1)	(1.7)	(1.7)	(1.6)	(1.6)	(1.8)	(1.8)
		MSW	81.7	81.8	79.9	83.4	83.2	80.4	77	77.4	76.6
			(0.6)	(0.7)	(0.7)	(1.1)	(1.1)	(1)	(1.1)	(1.1)	(1.1)
	3	SW	88.4	90.2	86.5	91.3	92.1	88	91.1	91.2	88.4
			(1.3)	(1.5)	(1.3)	(2.7)	(2.6)	(2.2)	(2.4)	(2.6)	(2.4)
		MSW	85.4	83.4	79.2	85	83.5	79.1	85.5	82.4	79.8
			(0.9)	(0.9)	(0.8)	(1.8)	(1.5)	(1.4)	(1.6)	(1.6)	(1.5)
1000	1	SW	97.9	94.2	91	97.9	94.2	91.7	97.5	94.7	91.6
			(5.3)	(3.8)	(3.3)	(8.5)	(5.7)	(4.7)	(9.3)	(6.6)	(5.7)
		MSW	94.6	91.1	84.8	94.8	91.3	86.3	95.6	90.8	85.8
			(2.7)	(2.3)	(2)	(4.2)	(3.3)	(2.7)	(4.6)	(3.7)	(3.2)
	2	SW	92	90.9	89	95.6	92.2	91	93.3	90.1	88.6
			(1)	(0.9)	(0.9)	(1.8)	(1.5)	(1.4)	(1.8)	(1.5)	(1.5)
		MSW	83.8	84.1	82.5	86.7	84.9	82.5	82.2	83	80.9
			(0.5)	(0.5)	(0.5)	(0.9)	(0.9)	(0.8)	(0.9)	(0.9)	(0.9)
	3	SW	95.2	92.5	90.5	95	92.2	90.4	94.8	91.7	89
			(1.5)	(1.3)	(1.2)	(2.7)	(2.1)	(1.8)	(2.8)	(2.3)	(2.1)
		MSW	86.2	85.1	83.6	88.7	87.7	84.8	88	85.9	84.3
			(0.7)	(0.7)	(0.7)	(1.3)	(1.2)	(1.1)	(1.4)	(1.3)	(1.3)

Table 1.5: Empirical coverage probabilities for linear processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method, with standard normal, $\chi^2 - 1$ and t_3 innovations, filters: $\varphi = 0.7, \vartheta = -0.3$ (Filter1);$\varphi = -0.7, \vartheta = 0.3$ (Filter2);$\varphi = \vartheta = 0$ (Filter3); sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.9$, block sizes $b = Cn^{1/2}$ with constants $C = 0.5, 1, 2$. Entry in the parenthesis () represents the average length of the intervals. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			Standard Normal Innovations			Chi-Square Innovations			t Innovations		
n	Filter	Method	$C = 0.5$	$C = 1$	$C = 2$	$C = 0.5$	$C = 1$	$C = 2$	$C = 0.5$	$C = 1$	$C = 2$
$\alpha = 0.9$											
400	1	SW	97.9 (1.5)	97.2 (1.5)	93.7 (1.3)	96.3 (2.9)	95.7 (2.6)	92.7 (2)	96.3 (2.5)	95.3 (2.5)	93 (2.1)
		MSW	97.7 (1.1)	93.6 (0.9)	89.9 (0.8)	96.3 (1.9)	93.1 (1.5)	87.6 (1.2)	94.8 (1.7)	91.3 (1.4)	84.8 (1.2)
	2	SW	91.4 (0.3)	91.8 (0.3)	89.7 (0.3)	90.8 (0.5)	91.6 (0.5)	89.6 (0.4)	87.3 (0.4)	89 (0.5)	87.1 (0.5)
		MSW	84.1 (0.2)	84.8 (0.2)	82.7 (0.2)	85.4 (0.3)	86.4 (0.3)	85.1 (0.3)	81.4 (0.3)	82 (0.3)	81.4 (0.3)
	3	SW	91.2 (0.4)	91.4 (0.4)	90.5 (0.4)	93.6 (0.9)	93.2 (0.8)	91.3 (0.7)	91 (0.7)	91.4 (0.8)	89.1 (0.7)
		MSW	89.1 (0.3)	86.7 (0.3)	85.2 (0.3)	90.2 (0.6)	89.3 (0.5)	86.7 (0.4)	85.8 (0.4)	84.2 (0.4)	81.6 (0.4)
1000	1	SW	98.1 (1.4)	96.1 (1)	93.4 (0.9)	98.9 (2.2)	97.1 (1.4)	95 (1.2)	99 (2.4)	96.5 (1.7)	94.3 (1.5)
		MSW	97 (0.7)	93.5 (0.6)	89.7 (0.5)	96.2 (1.1)	92.5 (0.8)	88.4 (0.7)	95.8 (1.1)	92 (0.9)	86.3 (0.8)
	2	SW	94.6 (0.2)	91.7 (0.2)	90.7 (0.2)	93.9 (0.4)	90.6 (0.3)	90.1 (0.3)	93.1 (0.4)	91.5 (0.3)	89.9 (0.3)
		MSW	87.2 (0.1)	88 (0.1)	87.5 (0.1)	83.6 (0.2)	84.2 (0.2)	83.1 (0.2)	84.7 (0.2)	85.1 (0.2)	84.4 (0.2)
	3	SW	95.1 (0.3)	91.8 (0.3)	90 (0.3)	94.8 (0.6)	92.7 (0.5)	91.1 (0.4)	96.1 (0.6)	93.5 (0.5)	91.9 (0.5)
		MSW	90.2 (0.2)	89 (0.2)	87 (0.2)	87.2 (0.3)	86.5 (0.3)	85.5 (0.3)	88.7 (0.3)	88.4 (0.3)	87.4 (0.3)

Table 1.6: Empirical coverage probabilities, average interval widths for transformed Gaussian processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method, with sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.1, 0.5, 0.9$. Data driven optimal bandwidth selection with $g = 0.75$ is employed. Entry in the parenthesis () represents the average length of the intervals. Entries in the square bracket [,] list the average mean optimal block size and average median optimal block size, respectively. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			$n = 400$			$n = 1000$		
n	transform	method	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
	1	MSW	72.1 (2.3) [6.8,6]	88.3 (1.0) [7.0,6]	91 (0.3) [8.0,6]	75.9 (2.2) [7.3, 6]	89.3 (0.8) [7.5,6]	91.2 (0.2) [9.3,6]
		SW	83.7 (3.8) [8.9,4]	93.9 (1.8) [8.6,6]	96.5 (0.6) [8.4,6]	85.5 (3.9) [16.3,14]	94.5 (1.4) [16.2,14]	95.7 (0.4) [15.8,14]
	2	MSW	82.8 (4.6) [7.3,6]	91.8 (0.8) [9.4,6]	86.6 (0.5) [11.2,8]	83.7 (4.2) [8.2,6]	94.1 (0.5) [13.9,11]	89.7 (0.3) [17.8,14]
		SW	94 (8.4) [7.0,4]	97.5 (2.4) [8.3,6]	96.4 (1.4) [9.9,6]	89.1 (6.8) [16.7,14]	96.9 (1.1) [17.3,18]	96 (0.6) [18.4,18]
	3	MSW	92.4 (7.0) [7,6]	93.8 (0.8) [8.4,6]	86.5 (0.5) [8.9,8]	92.5 (5.8) [8.1,6]	93.4 (0.5) [9.9,6]	90.5 (0.3) [10.2,8]
		SW	95.5 (13.6) [6.7,4]	97.9 (1.9) [8.2,6]	96.1 (1.3) [9.9,6]	94.7 (9.9) [16.3,14]	96.9 (0.9) [16.2,14]	97 (0.7) [15.8,14]

Table 1.7: Empirical coverage probabilities and average interval widths for linear processes using the subsampling window (SW) method with $\theta = 0.8$ and the modified subsampling window (MSW) method, with a standard normal innovation, filters: $\varphi = 0.7, \vartheta = -0.3$ (Filter1);$\varphi = -0.7, \vartheta = 0.3$ (Filter2);$\varphi = \vartheta = 0$ (Filter3); with sample sizes $n = 400, 1000$, long range dependence parameters $\alpha = 0.1, 0.5, 0.9$. Data driven optimal bandwidth selection with $g = 0.75$ is employed. Entry in the parenthesis () represents the average length of the intervals. Entries in the square bracket [,] list the average mean optimal block size and average median optimal block size, respectively. The results are based on $M = 1000$ replications. The nominal coverage is 90%.

			$n = 400$			$n = 1000$		
n	Filter	method	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
	1	MSW	96.1 (16.5) [6.8,6]	96.9 (4.2) [6.5,6]	98.8 (1.2) [6.5,6]	96.3 (16.3) [7.6,6]	97.2 (3.4) [6.7,6]	98.8 (0.8) [7,6]
		SW	97.3 (27.6) [5.6,4]	98.1 (7.5) [7.3,4]	99.2 (2.5) [7.6,4]	95.9 (22.7) [17.6,18]	97 (5.1) [17.6,18]	98.1 (1.4) [16.8,18]
	2	MSW	79.4 (2.4) [7.3,6]	81.3 (0.6) [7.8,6]	83.8 (0.2) [9,8]	83 (2.5) [9.8,6]	81.3 (0.5) [9.1,6]	85.8 (0.1) [10.1,8]
		SW	90.3 (5.7) [10.1,6]	91.5 (1.3) [8.3,6]	91.8 (0.3) [8.6,6]	91.9 (5) [15.3,14]	94.2 (1.1) [14.1,14]	94 (0.3) [15.2,14]
	3	MSW	84.4 (3.8) [6.7,6]	87 (1) [7,6]	89.4 (0.3) [7.9,6]	97.2 (3.4) [6.9,6]	88.2 (0.8) [7.4,6]	90.1 (0.2) [9.2,6]
		SW	91.2 (6.8) [10.7,6]	93.4 (1.9) [9.9,6]	95.1 (0.6) [8.1,6]	97 (5.1) [15.5,14]	95.7 (1.4) [15.6,14]	96.8 (0.4) [15.7,14]

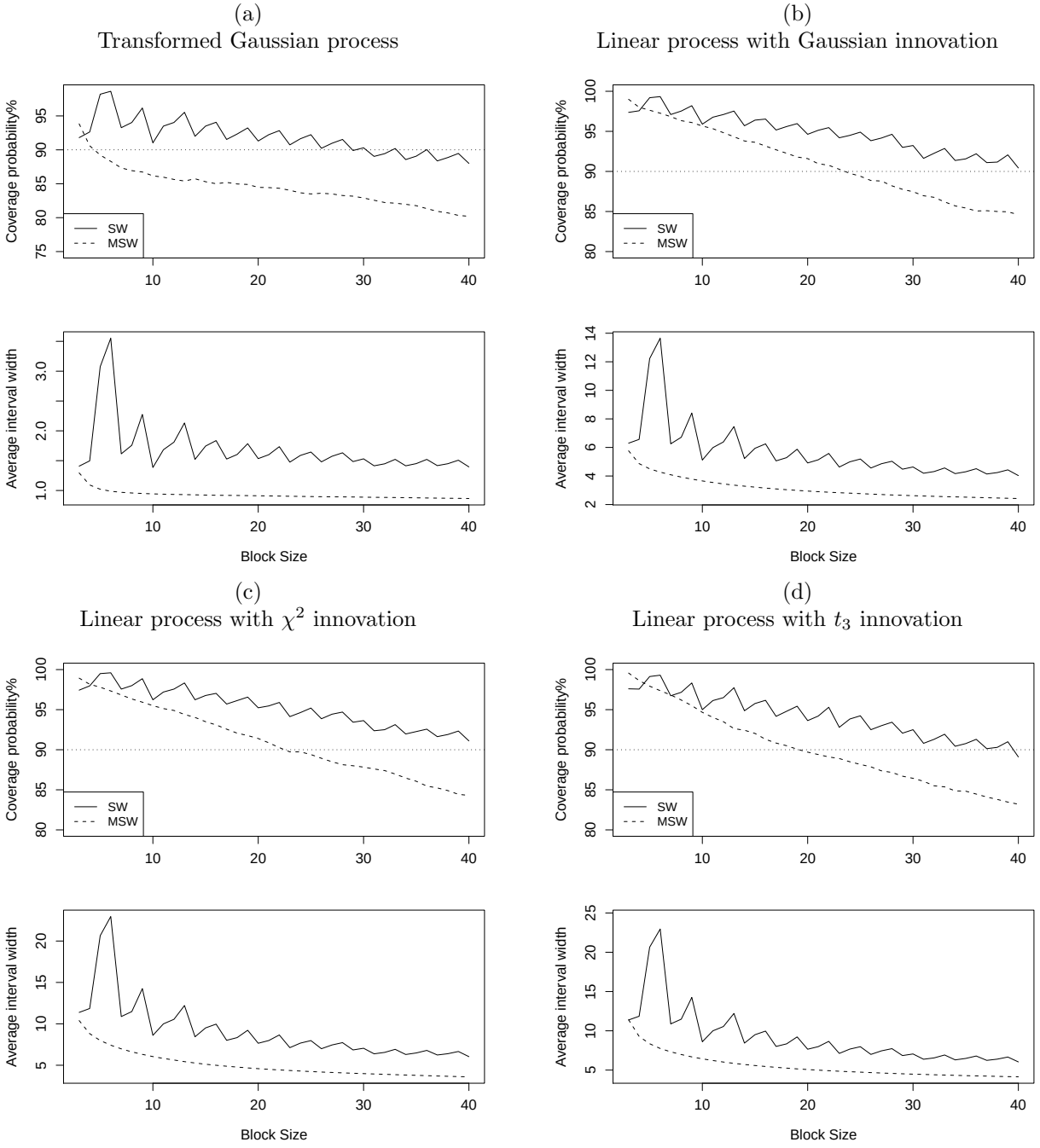

Figure 1.1: Coverage probabilities and average widths of the confidence intervals at the 90% nominal level. Sample size = 400, $\theta = 0.8$, block sizes vary from 3 to 40 and Hurst parameter=0.75. For linear processes, the filters are (0.7,-0.3). The results are based on $M = 3000$ replications. (a) Gaussian process with Hermite rank =1. (b) Linear process with Gaussian innovation. (c) Linear process with $\chi^2_1 - 1$ innovation. (d) Linear process with t_3 innovation.

Chapter 2

Clustering Analysis of Data Arising from Psychometric Latent Variable Models

2.1 Introduction

2.1.1 Latent Variable Models

We consider several models that are used in educational measurement, and how hierarchical agglomerative cluster analysis (HACA) and minimum diameter partitioning (MDP) behave when applied to data arising from these models. In particular, our analysis focuses on models from item response theory (IRT), linear factor analysis (LFA), and structured latent class models for cognitive diagnosis (CDM). When these models are used in educational testing, one typically assumes to know which underlying psychological constructs are responsible for the dependence in the data, and latent variables are used to represent them. Though HACA is a widely used technique in exploratory psychometric problems, it is typically not used in the measurement or classification phase of a testing endeavor. Latent variable models and HACA are seen as distinct approaches, and little has been done to explore their relationship. Our aim is to determine when HACA and MDP can provide consistent groupings of the subject variables in latent variable models, and see how much information about the underlying structure the distance measure must be supplied in order to expect useful partitions.

In general, suppose that responses to J variables are obtained on N subjects, and let Y_{nj} be the response of the nth subject on the jth variable. Latent variable models most often assume conditional independence of the J responses, given the vector-valued or scalar-valued latent variable $\boldsymbol{\theta}_n$. Responses of distinct subjects are assumed to be statistically independent. The conditional probability density function for the nth case is given by

$$f(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n; \boldsymbol{\beta}) = \prod_{j=1}^{J} f_j(y_{nj} \mid \boldsymbol{\theta}_n; \boldsymbol{\beta}_j),$$

where f_j is the density function for the jth variable, and depends on $\boldsymbol{\theta}_n$ and item-specific parameter $\boldsymbol{\beta}_j$. The value of a latent variable model is to reduce the complexity of the J-dimensional random vector $\boldsymbol{Y}$ by explaining all of the dependence in its components by a latent variable $\boldsymbol{\theta}$ that has a much smaller dimension.

The marginal distribution is obtained by integrating over the distribution of $\boldsymbol{\theta}$,

$$f(\boldsymbol{y}_n \mid \boldsymbol{\beta}) = \int f(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n; \boldsymbol{\beta}) dG(\boldsymbol{\theta}),$$

where the distribution of $\boldsymbol{\theta}$ may be either Lebesgue-dominated or dominated by counting measure. The particular latent variable models we study are reviewed chapter-by-chapter, along with theory and simulations for each case.

2.1.2 Hierarchical Agglomerative Cluster Analysis and Minimum Distance Partitioning

Hierarchical agglomerative clustering is a fast and popular method for arranging objects into homogeneous groups. HACA begins by defining a matrix of distances for all pairs of distinct observations, say $d_{nn'} = \sqrt{\sum_{j=1}^{J}(y_{nj} - y_{n'j})^2}$, in the case of Euclidean distance. Each object begins as its own cluster. A hierarchy of clusters is then formed by sequentially combining the two clusters at each stage that have the minimum distance between them. How one defines the distance between two clusters is what distinguishes different methods of HACA. In this research our aim is to investigate the tightness of clusters, and we use complete linkage HACA, which strives for compact clusters. Consider clusters C_m and $C_{m'}$, where $m \neq m'$. Complete linkage defines the distance between cluster C_m and $C_{m'}$, $d^*_{mm'}$, as the maximum distance between two points, one taken from cluster C_m and the other from $C_{m'}$,

$$d^*_{mm'} = \max_{n \in C_m, n' \in C_{m'}} d_{nn'}.$$

This definition of inter-cluster distance implies that every data point in the combined cluster would not be farther than $d^*_{mm'}$ away from every other data point in the cluster. Complete linkage clustering tends to produce clusters for which the largest diameter of all clusters remains as small as possible, when proceeding in such a sequential hierarchical manner. Complete linkage HACA provides a simple clustering approach that strives to keep all clusters compact, and tends to keep the diameter of the largest cluster in check as it proceeds to combine all of the initial N clusters into a single cluster. To arrive at a useful solution, one must decide where to end the process and cut the tree-like hierarchical structure of clusters and take the M clusters that exist at that stage.

Minimum diameter partitioning is a more direct way of controlling the maximum diameter across all clusters. Suppose we intend to partition the N objects into M clusters, and let $\pi_M = \{C_1, C_2, ..., C_M\}$ be a feasible partition. A feasible partition must include clusters that are nonempty, disjoint, and whose union

includes all N objects. Let Π_M denote the set of all feasible partitions. Define the diameter of the mth cluster by $d_m = \max_{n \in C_m n' \in C_m} d_{nn'}$. The largest of them will be denoted by $d_{max} = \max_{1 \leq m \leq M} d_m$. Let $\pi_M^* \in \Pi_M$ be a minimum diameter partition, a feasible partition for which $d_{max} = d_{max}^*$ is minimized over all feasible partitions.

There are close relationships between MDP and complete link HACA. The complete link HACA solution combines clusters to result in the minimum increase in the maximum diameter at each step. However, it does not necessarily reach a global optimum (Brusco and Stahl, 2005). Hansen and Delattre (1978), show that the complete link HACA solution can be quite far from an MDP. However, complete link HACA groups objects in the same spirit, and is quite easy to conduct, whereas solving for the MDP is an exponential computing problem. In our analysis, theory for the MDP is derived, and is derived for complete link HACA in some cases, though HACA is used for all simulations.

2.1.3 Cluster Analysis and Latent Variable Models

The quality of a partition of the subjects, when a latent variable model underlies their responses, is determined by how similar the values of $\boldsymbol{\theta}$ are within clusters. When $\boldsymbol{\theta}$ takes real values many clusters might be needed to achieve within-cluster homogeneity. In the case of latent class models, when $\boldsymbol{\theta}$ can take only finitely many values, there is a correct number of clusters that corresponds to the number of values $\boldsymbol{\theta}$ can take with positive probability. In either case, distance measures on the parameter space of $\boldsymbol{\theta}$ can be defined that correspond to the distance measures in the space of the data, that were used in the formation of the clusters. Suppose that $\boldsymbol{\theta}$ is a K-dimensional vector, and define the distance between the latent trait values of the nth and n'th subjects by,

$$\omega_{nn'} = \sqrt{\sum_{k=1}^{K} (\theta_{nk} - \theta_{n'k})^2}, \tag{2.1}$$

which corresponds to $d_{nn'}$. Similarly, the distance between two clusters in the latent variable space is $\omega_{mm'}^* = \max_{n \in C_m, n' \in C_{m'}} \omega_{nn'}$, which corresponds to $d_{mm'}^*$. Also, for any partition we let $\omega_m = \max_{n \in C_m n' \in C_m} \omega_{nn'}$, be the diameter of the mth cluster, and ω_{max} be the maximum of these. Denote ω_{max} by ω_{max}^* for a partition which is a MDP. Note that the MDP is determined on the distance measures formed from the data $\boldsymbol{Y}$ rather than the latent variables $\boldsymbol{\theta}$. Nevertheless, for any partition formed from applying HACA or MDP to the data, we can consider the tightness of the clusters measured by distances between $\boldsymbol{\theta}$ values for the subjects in the clusters.

The theory and simulations of the following sections concern the tightness of clusters in the space of $\boldsymbol{\theta}$. When cluster analysis is applied to data that can be effectively modeled with latent variables, one must

consider what the clusters are saying about the homogeneity of the latent traits of subjects within the same cluster. The ability of a clustering procedure to achieve tight groups may depend very critically on how one forms the distance measure $d_{nn'}$. There may be cases when Euclidean measure between the item response vectors $\boldsymbol{y}_n$ and $\boldsymbol{y}_{n'}$ may be effective and cases when it leads to poor results. We consider alternative versions of $d_{nn'}$ that are found by taking Euclidean distance between summaries of the data that use some, but not complete, knowledge of the underlying model. For example, in the next section in which unidimensional item response models are analyzed, we study the behavior of clusters obtained by forming $d_{nn'}$ from the difference in proportion correct values on the J items. This recognizes the unidimensionality of the latent trait to arrive at an efficient summary score, but does not use any further information about the underlying model. In all the models we consider, comparisons between simple Euclidean distance and distance measures that utilize partial knowledge of the latent variable model are compared.

2.2 Cluster Analysis and Item Response Models

Educational tests are often divided into subtests for which a linear ordering of abilities is desired. In such cases, latent variable models in which θ is scalar-valued are useful, and are appropriate if the dependence in the item responses can be explained by conditioning on a single latent trait. In particular, in the common case in which items are scores as right or wrong ($Y_{nj} \in \{0, 1\}$), IRT has become the dominant methodology. In item response models, the likelihood of a subject's vector of scores can be written as

$$P[\boldsymbol{y}_n \mid \theta_n] = \prod_{j=1}^{J} P_j(\theta_n)^{y_{nj}} (1 - P_j(\theta_n))^{1-y_{nj}},$$

where P_j is a function called the item characteristic curve that assigns the probability of a correct response to the jth item as a function of the latent ability θ. In the most general models, P_j is only assumed to be a monotone increasing function with a range in some subinterval of (0,1). A popular special case is the two-parameter logistic model in which $P_j(\theta) = 1/[1 + \exp(-\beta_{1j}(\theta - \beta_{0j}))]$. Parameter β_{1j} is the log odds-ratio for the latent covariate θ, and β_{0j} is a difficulty parameter that indicates the value of θ at which the item is most informative.

2.2.1 Consistency Results

When clustering data that arise from a unidimensional item response model, one would expect distances between total scores or proportion correct scores to work well. When a unidimensional IRT models holds, the sum-score $S_n = \sum_{j=1}^{J} Y_{nj}$ has a strictly increasing expected value as a function of θ_n, provided all of the

item characteristic curves are strictly increasing. In the important special case of the Rasch model (Rasch, 1960), S_n is a sufficient statistic for θ_n, but can be quite useful even when the Rasch model, which is much like the two-parameter logistic model described above but with a constant discrimination parameter, does not hold. When a unidimensional IRT model is suspected, the distance between proportion correct scores is

$$d^U_{n,n'} = \frac{1}{J} \mid S_n - S_{n'} \mid = \frac{1}{J} \mid \sum_{j=1}^{J} (Y_{nj} - Y_{n'j}) \mid. \tag{2.2}$$

It will be shown that the unidimensional distance measure in (2.2) yields a consistent clustering result, for a particular definition of consistent clustering given below. This distance measure is studied and compared with Euclidean distance between item response vectors $\boldsymbol{Y}_n$ and $\boldsymbol{Y}_{n'}$, for the purpose of clustering. Whether performing HACA with complete linkage, or studying the consistency results of MDP, Euclidean distance with binary data yields equivalent results to Hamming Distance. The Hamming distance, scaled by the reciprocal of test length, between the response vectors of two subjects can be expressed as

$$d^H_{n,n'} = \frac{1}{J} \sum_{j=1}^{J} |Y_{nj} - Y_{n'j}|. \tag{2.3}$$

Later we will show that clustering IRT data with Hamming distance is not likely to give consistent results.

In order to define consistent clustering with data arising from continuous latent variable models, we need to recall how the diameter of a cluster is defined in the space of θ. In the previous section, we defined the diameter of cluster C_m by $\omega_m = \max_{n \in C_m n' \in C_m} \omega_{nn'}$, where $\omega_{nn'}$ is defined in (2.1). In the unidimensional case $\omega_{nn'}$ reduces to $|\theta_n - \theta_{n'}|$. The maximum of these across all M clusters is ω_{max}, with value ω^*_{max} corresponding to an MDP solution.

Given this, we can consider how diameters shrink as N, J, and M all increase, to arrive at a definition of consistent clustering. Note that for continuous latent variables, the number of clusters M should approach infinity as N does, otherwise these diameters can never approach 0 with probability 1, under any possible method. To avoid trivial solutions, we require that the ratio of subjects to clusters, N/M, must also go to infinity. Thus consistent clustering means that ω_{max} has 0 as its limit, even though the clusters are growing in membership. Implicit in this definition is that we have a sequence of datasets that are increasing in N and J as we add subjects and items, and we think of M growing as a function of N and J.

Definition 2.2.1 (Consistent clustering). *Let N, J, and M be the number of subjects, items and clusters, respectively. Suppose π_M is a sequence of feasible partitions indexed by M, which is a function of J and N. If $\omega_{max} \to 0$ in probability when $N \to \infty$, $J \to \infty$, $M \to \infty$, and $\frac{N}{M} \to \infty$, then we call π_M a consistent*

sequence of partitions.

Next we state a theorem and conditions for consistent sequences of partitions under MDP.

Theorem 2.2.1. *Assume that responses arise from a unidimensional IRT model in which responses are conditionally independent given θ. Suppose that $N, J, M \to \infty$, $\frac{J}{M} \to \infty$, $\frac{N}{M} \to \infty$ and $N^2 \exp\left[-\frac{J}{M^\alpha}\right] \to 0$, for some $\alpha \in (0,1)$. Further assume that the support of θ is bounded, and that the slopes of all item characteristic curves are uniformly bounded above 0 in a compact set that contains the support of θ. Then when using distance measure d^U, any sequence of MDP solutions results in consistent clustering.*

The proof is given in the appendix. The assumption that θ has a bounded support is used to avoid tedious work with the tails where observations become sparse. However, it is common in IRT to assume θ has a standard normal distribution, and we could imagine truncating it at 10,000 or more, beyond which we would never observe a value with any likelihood in our finitely populated world. The assumption of a uniform lower bound on the slopes of item characteristic curves makes sense over a distribution with a support contained in a closed and bounded interval of real numbers. For example, it would follow in a two-parameter logistic model when slope parameters are bounded above 0 and difficulty paramors can be contained in a bounded interval. Concerning rates N, J and M, many possibilities exist. For example, we could let $J = \log(N)^3$ and $M = \log(N)$. Then we'd have convergence at a rate of at least $1/M^{0.5}$, corresponding to $\alpha = 1/2$.

The theory has shown that recognizing the unidimensionality underlying the data can lead to consistent clustering. Ignoring this and using Hamming distance, or equivalently Euclidean distance, breaks down when data arise from the unidimensional IRT model. To demonstrate this, it suffices to show that for some θ_n, $\theta_{n'}$ and θ_{n*},

$$\mathbf{E}[d^H_{nn'} \mid \theta_n = \theta_{n'}] > \mathbf{E}[d^H_{nn*} \mid \theta_n \neq \theta_{n*}],$$

where $\mathbf{E}[d^H_{n,n'}] = \sum_{j=1}^J P_j(\theta_n)[1 - P_j(\theta_{n'})] + \sum_{j=1}^J P_j(\theta_{n'})[1 - P_j(\theta_n)]$. Assume that for all j, $P_j(\theta_n) = P_j(\theta_{n'}) = 0.9$ and $P_j(\theta_{n*}) = 1$. So

$$\mathbf{E}[d^H_{n,n'} \mid \theta_n = \theta_{n'}] = 0.18,$$

and

$$\mathbf{E}[d^H_{n,n*} \mid \theta_n = \theta_{n*}] = 0.1.$$

This means that for Hamming distance, there can be a tendency to cluster data with differing θ's more than with similar θ's, which leads to inconsistent clustering.

2.2.2 Simulation

The simulation design for the performance of complete linkage cluster analysis utilized 100 independent replications in a design that varied test length and the number of clusters. For all simulations we set the sample size at $N = 1000$, and studied test lengths of $J = 20 and 100$ and grouped into $M = 10, 40$, and 80 clusters. The latent variable θ was drawn from a standard normal distribution, and items parameters of a two-parameter logistic model were drawn from distributions, $\beta_0 \sim N(0, 1)$ and $\beta_1 \sim U[1, 2.5]$.

Results on the distributions of cluster diameters are given in Table 3.1, and Table 3.2 summarizes the root-mean square distance between $\boldsymbol{\theta}$ values within the same cluster. By taking advantage of the unidimensionality of the model through forming the sum-scores S_n, we see that cluster diameters are uniformly much smaller than when this information is ignored and Euclidean distance is used. The same is true for root mean square distance of points within a common cluster as seen in Table 3.2. This indicates that knowledge of the underlying dimensionality, even without knowing specific item parameters, can be quite useful to incorporate when forming tight clusters. Note that improvements are generally seen when increasing the test length from 20 to 100, though increasing the number of clusters from 40 to 80 did not have a substantial impact.

2.3 Cluster Analysis and Latent Class Models

Cognitive diagnosis models (CDMs) are latent class models with constraints that represent a theory for how exam items are answered. They utilize assumptions made by experts concerning the attributes or skills that are required for each item and how these are combined to generate responses. The appeal of CDMs is that they promise to pinpoint the precise skills or abilities a subject has or has not mastered, which cannot be achieved by a single score. The relationship between the latent skills or attributes to the probability of a correct response is often dictated by whether the skills or attributes operate in a compensatory, disjunctive, or conjunctive fashion.

Specialized latent class models for cognitive diagnosis are derived under assumptions on which attributes are needed for which items, and how the attributes are utilized to construct a response. Let $\boldsymbol{\theta}$ be a K-dimensional vector for which the k^{th} entry θ_k, indicates whether or not a subject possesses the k^{th} attribute or skill, for $k = 1, 2, \cdots, K$. An attribute might refer to a clearly defined skill in some applications, or a more abstract psychological construct in another. All CDMs that we consider require a $J \times K$ matrix $\boldsymbol{Q}$, referred to as a Q-matrix (Tatsuoka, 1985), with (j, k) entry q_{jk} denoting whether or not the j^{th} item requires the k^{th} attribute. The vector $\boldsymbol{\theta}$ can take 2^K distinct values. These values index the 2^K latent classes in such models.

An example of a conjunctive model is the DINA (Deterministic Input, Noisy Output "AND" gate) model (Junker and Sijtsma, 2001). The DINA model extends the work of Macready and Dayton (1977), which considers a two-class version of it for assessing mastery of a skill. The item response function of the DINA model is,

$$P(Y_{nj} = 1|\boldsymbol{\theta}_n) = \beta_{0j}^{(1-\eta_{nj})}(1 - \beta_{1j})^{\eta_{nj}}, \tag{2.4}$$

where $\beta_{0j} = P(Y_{nj} = 1|\eta_{nj} = 0)$, $\beta_{1j} = P(Y_{ij} = 0|\eta_{nj} = 1)$, and η_{nj} is the ideal response which connects the attribute pattern possessed by a subject and the elements of $\boldsymbol{Q}$ in the following way,

$$\eta_{nj} = \prod_{k=1}^{K} \theta_{nk}^{q_{jk}}. \tag{2.5}$$

The variable η_{nj} indicates whether the subject possesses all the attributes needed for answering the particular item. Parameters β_{0j} and β_{1j} allow for stochastic deviations from the ideal responses, but should be somewhat close to 0. Otherwise, the validity of the conjunctive assumption would be drawn into question. The DINA model is characterized by its strong conjunctive feature that the probability of answering an item correctly will severely drop if any required attribute is missing. Estimation can be done with the EM algorithm (Haertel, 1989), or by use of Markov chain Monte Carlo (de la Torre and Douglas, 2004; Tatsuoka, 2002).

The NIDA (Noisy Input, Deterministic Output "And" gate) model, introduced in Maris (1999), and named in Junker and Sijtsma (2001), considers item responses as arising from a sequence of subtasks, and departures from ideal response patterns happen if a single misstep is taken along the path of solving a problem. Let η_{njk} indicate whether the nth subject correctly applied the kth attribute in completing the jth item. This leads to the parameters $\beta_{0k} = P(\eta_{njk} = 1 \mid \theta_{nk} = 0, q_{jk} = 1)$, and $\beta_{1k} = P(\eta_{njk} = 0 \mid \theta_{nk} = 1, q_{jk} = 1)$. An item response Y_{nj} is 1 if all η_{njk}'s are equal to 1. By assuming the η_{njk}'s are independent conditional on $\boldsymbol{\theta}_n$, the item response function is

$$P(Y_{nj} = 1 \mid \boldsymbol{\theta}_n) = \prod_{k=1}^{K} P(\eta_{njk} = 1 \mid \theta_{nk}) = \prod_{k=1}^{K} \left[\beta_{0k}^{(1-\theta_{nk})q_{jk}}(1 - \beta_{1k})^{\theta_{nk}q_{jk}}\right].$$

The NIDA model is quite restrictive because parameters at the subtask level are constant across all of the items. A generalization of this that loosens that restriction is a reduced version of the Reparameterized Unified Model (Hartz et al., 2005).

Whereas conjunctive models require the intersection of a set of attributes or successful implementations of these attributes, disjunctive models essentially replace "and" with "or". As an example, Templin and

Henson (2006), introduced the DINO (Deterministic Input, Noisy Output "Or" gate) model. The item response function of the DINO model is expressed as

$$P(Y_{nj} = 1|\boldsymbol{\theta}_n) = \beta_{0j}^{(1-\eta_{nj})}(1 - \beta_{1j})^{\eta_{nj}},$$

where $\eta_{nj} = 1 - \prod_{k=1}^{K}(1 - \theta_{nk})^{q_{jk}}$ and indicates whether at least one of the attributes corresponding to the item is possessed.

All of the models discussed above and many more can be represented in a log-linear model framework developed by Henson et al. (2009). Our analysis considers whether simple distance measures on the data can be used to obtain clusters that correspond to the 2^K values that $\boldsymbol{\theta}$ can take in these models, or if specialized distance measures that require more knowledge of the underlying structure are needed. In particular, we are interested in comparing clustering results obtained from Euclidean distance between $\boldsymbol{y}_n$ and $\boldsymbol{y}_{n'}$ and cluster solutions based on a summary score vector that assumes knowledge of the Q-matrix.

Let $\boldsymbol{W}_n = (W_{n1}, W_{n2}, ..., W_{nK})'$ be a vector of summed scores for items that measure each of the K attributes. The kth component is $W_{nk} = \sum_{j=1}^{J} Y_{nj}q_{jk}$. Because each item may require more than one attribute, an item may contribute to more than one component of $\boldsymbol{W}_n$. An alternative to Euclidean distance is to first summarize the data into these summed scores, and then compute $d_{nn'}$ as the Euclidean distance between $\boldsymbol{W}_n$ and $\boldsymbol{W}_{n'}$. This is similar to the summed-score approach in the previous section on IRT, but uses K different summed scores, one for each attribute. Chiu et al. (2009), studied the properties of HACA with distances obtained from $\boldsymbol{W}$, and the theoretical results of that work are summarized below, along with results indicating that consistent clustering cannot be obtained from Euclidean distance. However, simulation results show little difference.

2.3.1 Consistency Results

CDMs are restricted latent class models, and consistency results for clustering must utilize the same number of clusters as there are distinct latent classes, which is 2^K for the models we consider. In order to use data to distinguish between latent classes, some conditions on the Q-matrix must be satisfied. Call $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1, \eta_2, \cdots, \eta_J)'$ an ideal vector for the J items where $\eta_j = \prod_{k=1}^{K}\theta_k^{q_{jk}}$, and let $\boldsymbol{e}_k$ be a $K \times 1$ vector with the k^{th} entry being 1 and all other entries 0. A matrix $\boldsymbol{Q}$ is **complete** if it can identify all possible attribute patterns; that is, $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\eta}(\boldsymbol{\theta}^*)$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

Completeness refers to the ability of an exam to determine attribute patterns from one another, in a purely algebraic sense. Next we consider two lemmas that are needed for the main consistency theorem to follow. The first concerns necessary and sufficient conditions for $\boldsymbol{Q}$ be complete. Completeness is generally

needed for the identifiability of a model, which must be satisfied for correct classification. The first lemma implies that an exam must include some items that measure each attribute alone, among its many items, if identifiability of attribute patterns is to be achieved. Proofs of the lemmas and theorem are in Chiu et al. (2009).

Lemma 2.3.1. *A $J \times K$ matrix $\boldsymbol{Q}$ is complete if and only if it includes rows $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_K$, among its J rows.*

A complete $\boldsymbol{Q}$-matrix is required for the expected value of the sum-score vector $\boldsymbol{W}$ to distinguish between attribute patterns. Let $\boldsymbol{T}(\boldsymbol{\theta}) = \mathbf{E}[\boldsymbol{W} \mid \boldsymbol{\theta}]$. Then we see by the next lemma that a complete $\boldsymbol{Q}$-matrix is all that is required for $\boldsymbol{T}(\boldsymbol{\theta})$ to discriminate between latent classes in the DINA model.

Lemma 2.3.2. *Assume that all item responses are generated according to the DINA model and $0 \leq \beta_{0j} < 1 - \beta_{1j} \leq 1$ for $j = 1, 2, ..., J$. Also, assume that $\boldsymbol{Q}$ is complete. For attribute patterns $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$, if $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, then $\mathbf{T}(\boldsymbol{\theta}) \neq \mathbf{T}(\boldsymbol{\theta}^*)$.*

The following lemma is the basis for the proof of the consistency theorem, and essentially says that consistent clustering will take place if the data from a mixture model are replaced with their expected values. This means that observed distances are near their expected values, which is what takes place when the number of items J becomes large.

Lemma 2.3.3. *Let $\mathbf{V}$ be a random vector in K-dimensional space, with a mixture probability density function $f(\boldsymbol{v}) = \sum_{m=1}^{M} f_m(\boldsymbol{v})\zeta_m$, where ζ_m denotes the population proportion for the m^{th} latent class, and f_m is a probability density function in K-dimensional Euclidean space with expected value $\boldsymbol{\mu}_m$. Assume that for some positive number δ, $\min_{m \neq m'} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| > \delta$. Consider data $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N$, and let $\boldsymbol{v}_n^{(m)}$ denote that the n^{th} observation arose from the m^{th} component of the mixture density f. There exists a small enough $\varepsilon > 0$ such that if $\max_{i=1,2,\cdots,N} \|\boldsymbol{v}_i^{(m)} - \boldsymbol{\mu}_m\| < \varepsilon$ for all $i = 1, \cdots, N$, then the hierarchical agglomerative cluster analysis solution with complete linkage will place objects in clusters corresponding exactly with their latent class membership when the algorithm is cut at M clusters.*

The proof of the following theorem essentially amounts to showing that if J is sufficiently large, then $\boldsymbol{W}_n/J$ will be sufficiently close to its expected value for all n so the the result of the previous lemma will directly apply. Note that clustering based on $\boldsymbol{W}$ and $\boldsymbol{W}/J$ yield identical results when doing either HACA with complete link or MDP. Before stating the theorem, we define an **exact** cluster solution for data arising from a mixture model as one for which clusters correspond precisely with the components of the mixture.

Theorem 2.3.1. *Assume that responses arise from a cognitive diagnosis model in which responses are conditionally independent given a K-dimensional vector with binary components $\boldsymbol{\theta}$, and each of the 2^K values*

of $\boldsymbol{\theta}$ *are sampled with a probability greater than 0. Also, define* $\mathbf{E}[\boldsymbol{W}/J|\boldsymbol{\theta}^{(m)}] = \boldsymbol{\mu}_m$ *for* $m = 1, 2, \cdots, 2^K$, *and assume that for some positive number* δ, $\min_{m \neq m'} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| > \delta$. *Provided* $Ne^{-J} \to 0$ *as* $J \to \infty$, *the hierarchical agglomerative cluster analysis solution with complete linkage using Euclidean distances between* $\boldsymbol{W}/J$ *as input will be exact with probability converging to 1. Further, if* Ne^{-J} *is summable, the solution is inexact only finitely often with probability 1.*

Chiu et al. (2009), showed that this theorem holds for the DINA model. Though the focus was on HACA, the theorem also holds when HACA is replaced by MDP. In fact, it is more directly proven in that case. Though $\boldsymbol{W}$ leads to a consistent theory for the DINA model, it may not be a good statistic on which to base clusters when data arise from the NIDA model, for example. Other statistics need to be explored. Even in the case of the DINA model, in which the theory for consistent clustering has been developed, $\boldsymbol{W}$ appears to produce no better clusters than Euclidean distance between $\boldsymbol{Y}$ vectors, which amounts to the square root of Hamming Distance. This is somewhat surprising because $\boldsymbol{W}$ uses information in the $\boldsymbol{Q}$-matrix, and Euclidean distances between response vectors make no use of the underlying model structure.

Despite its strong performance in simulations to be displayed below, it can be shown that a theory for consistency using Hamming distance is unlikely to be proven, because it lacks a fundamental property required for consistent clustering. It is critical that $\mathbf{E}[d_{nn'}]$ is minimized when $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'}$, but this is not always the case under the DINA model. To show that clustering with Hamming distance breaks down when data arise from the DINA model, consider the following example. To prove that Hamming distance leads to inconsistent clustering, we must show that for some $\boldsymbol{\theta}_n$, $\boldsymbol{\theta}_{n'}$, and $\boldsymbol{\theta}_{n^*}$,

$$\mathbf{E}[d^H_{nn'}|\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'}] > \mathbf{E}[d^H_{nn^*}|\boldsymbol{\theta}_n \neq \boldsymbol{\theta}_{n^*}],$$

where d refers to either Euclidean distance or Hamming distance, which give identical results. Assume that the number of attributes is $K = 2$, so we have 4 latent classes corresponding to the attribute patters, $\boldsymbol{\theta}_1 = (0, 0)$, $\boldsymbol{\theta}_2 = (1, 0)$, $\boldsymbol{\theta}_3 = (0, 1)$, $\boldsymbol{\theta}_4 = (1, 1)$. Let the Q matrix include all possible entries in equal

proportions, that require at least one attribute,

$$\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ \vdots & \end{pmatrix}.$$

Let $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n'} = (0,1)$ and $\boldsymbol{\theta}_{n^*} = (0,0)$. Also assume that $\beta_{0j} = \beta_0$ and $\beta_{1j} = \beta_1$, for $1 \leq j \leq J$. Then

$$\mathbf{E}[d^H_{n,n'}] = \frac{4}{3}\beta_0(1-\beta_0) + \frac{2}{3}\beta_1(1-\beta_1). \tag{2.6}$$

$$\mathbf{E}[d^H_{n,n^*}] = \frac{4}{3}\beta_0(1-\beta_0) + \frac{1}{3}\beta_1\beta_0 + \frac{1}{3}(1-\beta_1)(1-\beta_0). \tag{2.7}$$

When $\beta_1 > 0.5$ and $1-\beta_1 > \beta_0$, (2.6) > (2.7). This means there would be a tendency to cluster data arising from different latent classes, rather than clustering subjects that have identical attribute patterns. Though this example suggests that Hamming distance may not lead to consistent clustering for some cognitive diagnosis models, it happens to perform well in simulation with realistic parameter values.

2.3.2 Simulations

In all simulations, the number of attributes was $\mathbf{K} = 3$, and the 8 possible attribute patterns were drawn with equal probability. Test lengths were $J = 20$ and $J = 100$. The Q matrix when $J = 100$ was 5 replications

of $\boldsymbol{Q}$ for $J = 20$, which was,

$$\boldsymbol{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}. \tag{2.8}$$

The item responses were generated generated from the DINA model in (2.4) with β_0 parameters drawn from a uniform distribution $U[0, 0.3]$, and β_1 parameters drawn from a uniform distribution $U[0, 0.15]$. Hierarchical clustering with complete linkage was performed using both $\boldsymbol{W}$ and the raw data as summaries. In the tables these are referred to as Q-matrix and Euclidean distance, respectively. The number of clusters was fixed to be 8.

Results on the distributions of cluster diameters are given in Table 3.3, and Table 3.4 summarizes the root-mean square distance between $\boldsymbol{\theta}$ values within the same cluster. We see that when test length is small, i.e., when $J = 20$, there are mixed results. The analysis of cluster diameters summarized in Table 3.3 appears to slightly favor distances based on $\boldsymbol{W}$, though root mean squared distances within all pairs within clusters slightly favors Euclidean distance. However, when the test length is 100, Euclidean distance, or equivalently Hamming distance, appears to result in tighter clusters. These results indicate that while no consistency

theorem for Euclidean distance is likely to be proven, it may be more practical in many situations. Also, it is quite possible that there are more efficient ways to summarize the data than with $\boldsymbol{W}$, which surely depend on what the underlying cognitive diagnosis model is. Further research could yield distance measures that outperform Euclidean distance, which does not take into account any knowledge of the model.

2.4 Cluster Analysis and the Linear Factor Analysis Model

The large set of statistical methods which encompass factor analysis, are among the most frequently used techniques in psychometrics. The level of a psychological construct may be thought of as a latent variable over the population that is being sampled. In factor analysis these latent variables are usually called factors, because they are used in factor analysis models as the key elements in explaining the relationships among a set of observable variables. In an influential paper by Spearman (1904), factor analysis was first used to model children's examination scores in an effort to explore a theory that there might be one general ability that accounts for performance on several exams of different academic subjects.

Spearman's model easily generalizes to the linear factor analysis model with K factors. In this model we can express Y_{nj} by

$$Y_{nj} = \beta_{j1}\theta_{n1} + \beta_{j2}\theta_{n2} + \cdots \beta_{jK}\theta_{nk} + e_{nj} = \boldsymbol{\beta}_j'\boldsymbol{\theta}_n + e_{nj}. \tag{2.9}$$

By forming the matrix of factor loadings $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_j$ in its jth row, the entire random vector of responses is

$$\boldsymbol{Y}_n = \boldsymbol{\beta}\boldsymbol{\theta}_n + \boldsymbol{e}_n.$$

In the orthogonal linear factor model, it is assumed that the covariance of $\boldsymbol{\theta}$ is the $K \times K$ identity matrix, and $\boldsymbol{\theta}$ is uncorrelated with $\boldsymbol{e}$. The residual vector $\boldsymbol{e}$ is usually assumed to have a diagonal covariance matrix $\boldsymbol{\Psi}$, with diagonal entries that represent the proportion of the variation of each variable not explained by the K factors. This results in the covariance matrix,

$$\text{var}(\boldsymbol{Y}) = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Psi}.$$

Equivalent models can be represented by taking orthogonal or oblique rotations of $\boldsymbol{\theta}$, which result in the same covariance matrix for $\boldsymbol{Y}$, but different factor loadings and possibly intercorrelated components of the factor $\boldsymbol{\theta}$.

In this section, we study the behavior of cluster analysis when applied to data arising from the linear factor analysis model. Three different distance measures are compared. The first is simply the Euclidean

distance between $\boldsymbol{Y}_n$ and $\boldsymbol{Y}_{n'}$. The second is to replace $\boldsymbol{Y}$ with its first K principal components, and form distances based on Euclidean distances between these K components. This utilizes some knowledge of the underlying model, including the number of factors K. Also, if the error variance is small, the coefficients for these components will be very similar to $\boldsymbol{\beta}$. Finally, distances obtained from $\boldsymbol{W}$ are used based on replacing the data with K sum scores, like in the cognitive diagnosis section. When all loadings are nonnegative, and if we had knowledge of which of them are nonzero, it makes sense to consider the vector $\boldsymbol{W}_n = (W_{n1}, W_{n2}, ..., W_{nK})'$ in which $W_{nk} = \sum_{j=1}^{J} Y_{nj} I[\beta_{jk} > 0]$. The first distance measure makes no assumptions about the underlying model, whereas the final two use a considerable amount of knowledge about the model.

2.4.1 Consistency Results

A theorem is given below for MDP based on Euclidean distances between values of the sum-scores $\boldsymbol{W}$. This is the same statistic as was used in Theorem 2 in the case of cognitive diagnosis. However, note that the unidimensional proportion correct statistics of Theorem 1 also amounts to the special case where $K = 1$, and the method of proof for Theorem 3 below is quite similar to the of Theorem 1. The differences are that it takes place in higher dimensions and with continuous unbounded random variables. Using $\boldsymbol{W}$ for clustering essentially amounts to knowing the basic structure of the factor pattern, but without knowing the exact values of the factor loadings.

Unlike we have seen in previous chapters, it appears that some consistency result is possible for clustering based on Euclidean distance between the raw data vectors. In the linear factor analysis model, the expected distances are minimized when latent variable values are the same, which is the basic requirement for consistency. However, a rigorous proof has so far been elusive. Next we present the theorem for consistency of clusters based on $\boldsymbol{W}$, and the proof may be found in the appendix.

Theorem 2.4.1. *Assume that responses arise from a linear factor analysis model in which responses are conditionally independent given $\boldsymbol{\theta}$. We assume that $\boldsymbol{\theta}$ is multivariate normal with mean $\mathbf{0}$ and covariance equal to the $K \times K$ identity matrix and that the jth variable Y_j is scaled to have mean 0 and variance 1 for $j = 1, 2, ..., J$. Assume that factor loadings are nonnegative, and all factors are asymptotically represented according to $\min_k \sum_{j=1}^{J} \beta_{jk} > CJ$ for some $C > 0$. Then any sequence of MDP solutions using clusters based on Euclidean distances between sum-scores $\boldsymbol{W}$ will be consistent, provided that $\boldsymbol{Q'\beta}$ is full rank, and N, J, and M satisfy $N \to \infty$, $J \to \infty$, $M \to \infty$, and for some $\alpha \in (0, 1)$, $\frac{\sqrt{\log N}}{M^{\alpha}} \to 0$, $\frac{\sqrt{J \log N}}{M^{\alpha}} \to \infty$ and $N^2 \frac{M^{\alpha}}{\sqrt{J \log N}} \exp\left[-\frac{J \log N}{M^{2\alpha}}\right] \to 0$.*

2.4.2 Simulations

In all simulations the number of factors was $K = 3$, and the number of variables was varied between $J = 20$ and 100. The sample size was fixed at $N = 1000$, and we recorded results for $M = 10, 40$ and 80 clusters. In the simulation model $\boldsymbol{\theta} \sim N(\mathbf{0}, I_{3\times 3})$, all errors terms $e_{n,j}$'s were independent, and $\psi_j = \text{var}(\text{e}_\text{j})$ were drawn from a uniform distribution $U[0, 0.3]$. Nonnegative factor loadings $\boldsymbol{\beta}$ were generated so that the overall variance of Y_j would be 1, which implies $\sum_{k=1}^{3} \beta_{jk}^2 = 1 - \psi_j$. The nonzero loadings were determined according to the same $\boldsymbol{Q}$- matrix defined in (2.8), replicating this matrix as needed to adjust for different test lengths. The nonzero loadings were then generated by drawing random uniform variates, and placing the squared loadings in the same proportions as them, but scaled to satisfy the constraint that they must sum to $1 - \psi_j$.

Table 3.5 listed the distribution of cluster diameters under four different distance measures and Table 3.6 summarized the root-mean square distances between $\boldsymbol{\theta}$ values within the same cluster. As expected clusters formed by randomly assigning points into one cluster yielded the worst clustering results among all the distance measures However, documenting these results serves as a baseline to measure the efficiency of the other techniques. In general, improvements were observed when either the test length increased or when the number of clusters increased. Clustering based on distance between summed scores on each dimension of the factor pattern, the first three principal components and Euclidean distance performed similarly for the same test length and number of clusters, though Euclidean distances and distances between principal components appeared to slight outperform the Q-matrix based sum-score approach, despite the consistency theorem for this approach.

2.5 Discussion

The aim of this paper has been to explore the behavior of HACA and MDP when applied to several models that are used in educational testing and educational measurement. In each case, some distances measures were used that took advantage of some knowledge of the underlying models, and Euclidean distance between response vectors was also used. In the case of IRT, it was clearly advantageous to make use of a unidimensionality assumption, and the theory supported this as well as simulations. For latent class models for cognitive diagnosis, a theory was developed for a particular sum-score vector, upon which distances may be based to yield consistent solutions. However, simulations appeared to prefer Euclidean distance, even though a counterexample was given indicating that consistency may not be possible. Several distance measures were used to cluster data arising from the linear factor analysis model, though Euclidean distance and distance

between principal components using the same number of components as factors, appeared to work best. A consistency theorem for MDP based on Euclidean distance seems likely, and remains to be proven. Consistency was shown for a sum-score approach, similar to the one considered for cognitive diagnosis, though the corresponding procedure performed poorly in simulations.

The motivation for this research was to study two views of multivariate data analysis that are often distinct, clustering and latent variable modeling. The results for IRT and factor analysis could have been expected. Because clusters tend to be finite and continuous variables can take on any real numbered values, many clusters are obviously needed before they can be homogeneous in the latent variable space. In the case of IRT, proceeding with Hamming distance or Euclidean distance can give very poor results when examined from the space of the latent variable, and some knowledge of the model is very useful. However, when the true model is a linear factor analysis model, simulations indicate that Euclidean distance can result in very homogeneous clusters. The most natural analysis was to study the behavior of clustering when data arise from latent class models, such as those used in cognitive diagnosis. Here the question becomes whether clustering with a carefully chosen distance measure can return the correct latent classes, without actually fitting a latent class model. Results are encouraging, but more work into summary statistics that capture the critical features of these models is needed.

2.5.1 Simulation results

Table 2.1: Summary statistics of cluster diameters on the latent trait scale are given for clusters formed based on distance between proportion correct (unidimensional) and Euclidean distance with different test lengths and numbers of clusters. The results are based on 100 replications, with standard deviations across replications reported in parentheses.

		test length=20			test length=100		
clusters	distance	minimum	maximum	median	minimum	maximum	median
10	unidimensional	1.33 (0.24)	2.65 (0.33)	1.86 (0.13)	0.76 (0.07)	1.99 (0.34)	1.05 (0.07)
	Euclidean	1.36 (0.34)	3.82 (0.42)	2.31 (0.25)	0.88 (0.18)	3.24 (0.33)	1.34 (0.09)
40	unidimensional	0.00 (0.00)	2.36 (0.29)	0.55 (0.39)	0.22 (0.17)	1.30 (0.26)	0.65 (0.02)
	Euclidean	0.34 (0.21)	3.35 (0.38)	1.50 (0.13)	0.23 (0.14)	3.02 (0.32)	0.86 (0.05)
80	unidimensional	0.00 (0.00)	2.32 (0.30)	0.00 (0.0)	0.01 (0.03)	1.14 (0.37)	0.52 (0.1)
	Euclidean	0.04 (0.07)	3.08 (0.37)	1.15 (0.1)	0.04 (0.06)	2.88 (0.34)	0.66 (0.04)

Table 2.2: The root mean squared distances between pairs of θ values within a cluster are given for clusters formed from distance between proportion correct (unidimensional) and Euclidean distance with different test lengths and numbers of clusters. Standard deviations across 100 simulation runs are reported in parentheses.

clusters	distance	test length=20	test length=100
10	unidimensional	0.52 (0.03)	0.31 (0.01)
	Euclidean	0.84 (0.05)	0.68 (0.05)
40	unidimensional	0.49 (0.02)	0.23 (0.01)
	Euclidean	0.76 (0.05)	0.67 (0.04)
80	unidimensional	0.48 (0.02)	0.23 (0.01)
	Euclidean	0.72 (0.05)	0.65 (0.05)

Table 2.3: Cluster diameters on the latent variable scale are given for clusters based on distance between $\boldsymbol{W}$ (Q-matrix) and Euclidean distance for different test lengths. The number of clusters is always 8, which in the number of latent classes in the model. The results are based on 100 replications, and standard deviations of the statistics are across replications are given in parentheses.

	test length=20			test length=100		
distance	minimum	maximum	median	minimum	maximum	median
Euclidean	0.37	1.73	1.33	0.00	0.63	0.00
	(0.52)	(0.03)	(0.15)	(0.0)	(0.5)	(0.0)
Q-matrix	0.08	1.73	1.16	0.00	1.24	0.05
	(0.27)	(0.03)	(0.15)	(0.00)	(0.21)	(0.15)

Table 2.4: The square root of the average squared distance between $\boldsymbol{\theta}$ values within the same cluster are given for clusters based on distance between $\boldsymbol{W}$ (Q-matrix) and Euclidean distance for different test lengths. The number of clusters is always 8, which in the number of latent classes in the model. The results are based on 100 replications, and standard deviations of the statistics are across replications are given in parentheses.

distance	test length=20	test length=100
Euclidean	0.6	0.04
	(0.08)	(0.03)
Q-matrix	0.84	0.24
	(0.07)	(0.11)

Table 2.5: Summary statistics of cluster diameters on the latent trait scale are given for clusters formed based on $\boldsymbol{W}$ (Q-matrix), the first three principal components, and Euclidean distance with different test lengths and numbers of clusters. As a baseline, results for randomly formed clusters are also given. The results are based on 100 replications, with standard deviations across replications reported in parentheses.

		test length=20			test length=100		
clusters	measure	minimum	maximum	median	minimum	maximum	median
10	Euclidean	2.64 (0.41)	4.76 (0.20)	3.94 (0.07)	2.42 (0.38)	4.62 (0.10)	3.74 (0.06)
	Q-matrix	2.71 (0.48)	5.71 (0.21)	4.53 (0.10)	2.33 (0.00)	5.57 (0.17)	4.44 (0.06)
	PCA	2.56 (0.43)	4.77 (0.16)	3.86 (0.07)	2.39 (0.34)	4.57 (0.11)	3.76 (0.06)
	random	3.95 (1.22)	7.13 (0.47)	5.92 (0.23)	3.89 (1.20)	7.09 (0.41)	5.92 (0.23)
40	Euclidean	0.24 (0.43)	2.97 (0.18)	2.13 (0.08)	0.21 (0.41)	2.64 (0.12)	1.93 (0.07)
	Q-matrix	0.28 (0.46)	3.44 (0.20)	2.43 (0.10)	0.41 (0.43)	3.15 (0.13)	2.27 (0.08)
	PCA	0.25 (0.42)	2.88 (0.15)	2.06 (0.08)	0.15 (0.40)	2.59 (0.11)	1.91 (0.06)
	random	0.51 (0.87)	6.76 (0.37)	4.83 (0.17)	3.89 (0.74)	7.09 (0.40)	5.92 (0.14)
80	Euclidean	0.01 (0.08)	2.35 (0.17)	1.64 (0.06)	0 (0)	1.91 (0.09)	1.31 (0.04)
	Q-matrix	0.00 (0.00)	2.60 (0.17)	1.66 (0.06)	0 (0)	2.26 (0.10)	1.51 (0.04)
	PCA	0.00 (0.00)	2.23 (0.15)	1.44 (0.05)	0 (0)	1.83 (0.08)	1.26 (0.04)
	random	0.03 (0.18)	6.64 (0.48)	4.24 (0.12)	0 (0)	6.56 (0.36)	4.24 (0.14)

Table 2.6: The square root of the average squared distance between pairs of $\boldsymbol{\theta}$ values within a cluster are given for clusters formed based on $\boldsymbol{W}$ (Q-matrix), the first three principal components, and Euclidean distance with different test lengths and numbers of clusters. As a baseline, results for randomly formed clusters are also given. The results are based on 100 replications, with standard deviations across replications reported in parentheses.

clusters	measure	test length=20	test length=100
10	Euclidean	1.62	1.60
		(0.06)	(0.06)
	Q-matrix	1.77	1.75
		(0.07)	(0.07)
	PCA	1.61	1.58
		(0.06)	(0.05)
	random	2.45	2.45
		(0.03)	(0.03)
40	Euclidean	1.08	1.01
		(0.03)	(0.03)
	Q-matrix	1.19	1.13
		(0.04)	(0.04)
	PCA	1.06	1.00
		(0.03)	(0.03)
	random	2.45	2.45
		(0.03)	(0.04)
80	Euclidean	0.87	0.76
		(0.03)	(0.10)
	Q-matrix	0.93	0.86
		(0.04)	(0.10)
	PCA	0.83	0.75
		(0.03)	(0.09)
	random	2.44	2.45
		(0.04)	(0.04)

Chapter 3

Response Times in Computerized adaptive testing

3.1 Introduction

Computerized adaptive testing (CAT) is a method of administering a test that adapts to the examinee's θ level (trait level). A CAT test differs from the paper-pencil test format in that in CAT, different examinees with different ability levels are tested with different sets of items. In a paper-pencil test, all examinees receive the same set of items. The purpose of CAT is to estimate each examinee's θ precisely, and to select test items sequentially from an item pool based on the current performance of a test taker. In other words, the test is tailored to the test taker's θ level, so that the more able examinees can avoid being given too many easy items, whereas less able examinees can avoid being exposed to too many difficult items. The major advantage of CAT is that it provides more precise θ estimates with relatively fewer items than that would be required in the conventional tests. Examples of large scale CATs include the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), the National Council of State Boards Nursing (NCLEX), and the Armed Services Vocational Aptitude Battery (ASVAB). The most important aspect in CAT is the item selection procedure. Heuristically, when the test taker answers an item correctly, the next item selected for him or her should be more difficult. If the answer is incorrect, the next item should be easier. To equate scores from different sets of exams, item response theory (IRT) models are utilized. IRT is widely used to estimate both the characteristics of the test items and the ability level of the test takers. Such analyses are based on the examinees responses to the test items.

When a test is administered in a computerized mode, the capability of recording the amount of time a test taker has spent on each item provides us with additional information about the test-taking experience of individuals as well as the characteristics of items. Recently, there has been rapid development on how to utilize response times on test items as an additional source of information in estimating the abilities of the test takers when the test is delivered in a computerized fashion. We propose semiparametric models for response times, and algorithms for item selection that use response time information. Such models can assist in controlling the amount of time required for a CAT, and in the appropriate circumstances can also

be used in estimation of ability.

The rest of the chapter is organized as follows. Section 3.2 reviews the classical IRT theory and several prominent models in the psychometrical response time literature. We propose a proportional hazard model for response times in section 3.3 and follow a simulation study in Section 3.3.2. We extend the response time application in computerized adaptive testing in Section 3.4, where we propose a new item selection strategy and compare it with classical method. We also discuss an important issue in CAT, item exposure rate, and study how it may be controlled. A simulation study is conducted in Section 3.4.5. Conclusions and all the simulation results are listed in Section 3.5.

3.2 Review of IRT and some Response Time Models

3.2.1 Item Response Theory

Item response theory (IRT) has been a dominant methodology in educational testing for decades, and is also widely used in psychology, marketing research and surveys of quality of life in medicine. In the case of unidimensional IRT models, we assume that item responses are statistically independent if we condition on a 1-dimensional latent trait. This trait represents the true value of the psychological construct is being measured. Furthermore, it is usually assumed that expected scores on the items are monotonically related to this trait. There are many ways to investigate the adequacy of a single latent trait, and a standard method related to factor analysis is the examination of the tetrachoric correlation matrix. Even if small departures from unidimensionality are identified, the practical use of an assessment is often to construct a linear ordering of subjects, making the unidimensional IRT model desirable. Among its desirable properties are that items can be parameterized to describe their difficulty and the information they supply at different levels of the latent trait, and they may be used in different combinations with different populations of subjects and still be used to construct scores on a common scale.

An example of a common parametric IRT model is the two-parameter logistic model. The item response function of this model is given below. Here we show the probability that the response of the nth subject to the jth variable is correct, given θ_n, the level of the latent trait of the nth subject.

$$P_j(\theta_n) = P(Y_{nj} = 1|\theta_n) = \frac{e^{a_j(\theta_n - b_j)}}{1 + e^{a_j(\theta_n - b_j)}} \tag{3.1}$$

In this model, the a parameters are log odds-ratios, and reflect the extent that changes in θ result in changes in the odds of a correct response. The b parameters describe at what level of θ the probability of a correct

response is 0.5, so it can be thought of as a location or difficulty parameter.

3.2.2 Current Models for Response Times

We briefly review several response time models in the test theory literature. Verhelst, Verstraalen, and Jansen (1997) present a model that is based on the assumption of a generalized extreme-value distribution of a latent response variable given the time spent on the item and a gamma distribution for the time. For the probability of a response on item j and person n,

$$P_j(\theta_n, \tau_n) = [1 + \exp(\theta_n - \ln \tau_n - b_j)^{-\pi_j}], \tag{3.2}$$

where b_j is the difficulty parameter for item j, θ_n is the ability parameters for the nth person, τ_n is the speed parameter for the person n, and π_j is an item-dependent shape parameter. For $\pi = 1$, the model reduces to a Rasch (1980) type model with $\xi_j = \theta_n - \ln \tau_n$ replacing the traditional ability parameter. The model in (3.2) incorporates a speed-accuracy tradeoff. If a person decides to increase the speed τ_n, parameter ξ_n decreases. Roskam (1987) proposed a similar model,

$$P_j(\theta_n) = [1 + \exp(\theta_n + \ln t_{nj} - b_j)^{-1}], \tag{3.3}$$

The model assumes a speed-accuracy tradeoff directly between the ability of the test taker and the time spent on a test item; less time on an item results in a higher speed and lower accuracy.

Under the assumption that the more capable person tends to answer the questions faster, i.e., a higher ability θ_n implies a shorter response time on the item, Thissen (1983) introduced the following model,

$$\ln T_{nj} = \mu + \tau_n + \beta_j - \rho(a_j\theta_n - b_j) + \epsilon_{nj}, \tag{3.4}$$

where $\epsilon_{nj} \sim N(0, \sigma^2)$. Parameters τ_n and β_j can be interpreted as the speed of the examinee and the amount of time required by the item, respectively. μ is a general level parameter and a_j, θ_n, and b_j are the item discrimination, ability, and item difficulty parameters, respectively. The term $\rho(a_j\theta_n - b_j)$ represents a regression of a two-parameter response model on the logtime with ρ being the regression parameter.

Rouder, Sun, Speckman, Lu, and Zhou (2003) propose a model based on a Weibull distribution. The Weibull distribution is widely used to model the waiting time for a system failure. In this model, the reaction time

distribution for person n on item j has the density

$$f(t_{nj}) = \frac{\pi_n (t_{nj} - \psi_n)^{\pi_n - 1}}{\sigma_n^{\pi_n}} \exp\left\{ - \left[\frac{t_{nj} - \psi_n}{\sigma_n} \right]^{\pi_n} \right\}, \; t_{nj} > \psi_n, \tag{3.5}$$

where ψ_n, σ_n and π_n are the shift, scale and shape parameters, respectively. The model in (3.5) does not assume the relationship between the ability of the examinee and the characteristics of the items.

Van der Linden (2007) proposes a hierarchical framework of modeling with two different levels, one for the individual test taker and one for the population of test takers. Each level includes two components: one to model speed and the other to model accuracy. At the level of the individual test taker, the framework models the test taker's responses to items (correct or incorrect) and the time he or she spent on each item. Both component models have separate parameters for the item and person effects. At the second level, the framework has a model for the population of test takers that explains how the speed and accuracy of the test takers tend to be related. Van der Linden's (2007) model is based on the following assumptions. The first assumption is that a test taker operates at a fixed level of speed, therefore the examinee operates at a fixed level of accuracy. This stationarity assumption is is a standard assumption in IRT. The next assumption is that responses and response times are conditionally independent given the levels of ability and speed at which the test taker operates, which is analogous to the local independence assumption in IRT. The items are indexed by $j = 1, \cdots, J$, and the examinees by $n = 1, \cdots, N$. For the nth test taker, his or her responses and response times are denoted by $\mathbf{Y}_n = (Y_{1n}, \cdots, Y_{Jn})'$, and $\mathbf{T}_n = (T_{1n}, \cdots, T_{Jn})'$. At the first level, two models for the responses and times are specified separately. Each response variable is assumed to follow a three-parameter logistic (3PL) model:

$$P_j(\theta_n) = c_j + (1 - c_j) \frac{\exp[a_j(\theta_n - b_j)]}{1 + \exp[a_j(\theta_n - b_j)]}. \tag{3.6}$$

A lognormal model is chosen for the response times:

$$T_{nj} \sim f(t_{nj}; \tau_n, \alpha_j, \beta_j) \equiv \frac{\alpha_j}{t_{nj}\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left[\alpha_j (\ln t_{nj} - (\beta_j - \tau_n)) \right]^2 \right\}, \tag{3.7}$$

where τ_n, β_j and α_j are the speed parameter for examinee n, the time intensity and discriminating power of item j, respectively.

At the second level, denote by $\xi_n = (\theta_n, \tau_n)$, $\psi_j = (a_j, b_j, c_j, \alpha_j, \beta_j)$ the person parameter for the nth person and item parameter for the jth item, respectively. ξ_n is assumed to be randomly drawn from a multivariate

normal distributor; that is,

$$\xi_n \sim f(\xi_n; \mu_p, \mathbf{\Sigma}_p) \equiv \frac{|\mathbf{\Sigma}_p^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\xi_n - \mu_p)^T \mathbf{\Sigma}_p^{-1}(\xi_n - \mu_p)\right], \tag{3.8}$$

with mean vector

$$\mu_p = (\mu_\theta, \mu_\tau),$$

and covariance matrix

$$\mathbf{\Sigma}_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{0\tau} \\ \sigma_{0\tau} & \sigma_\tau^2 \end{pmatrix}.$$

Analogous to (3.30), the parameter vector ψ_j is assumed to also follow a multivariate normal distribution

$$\psi_j \sim f(\psi_j; \mu_J, \mathbf{\Sigma}_J) \equiv \frac{|\mathbf{\Sigma}_J^{-1}|^{5/2}}{2\pi} \exp\left[-\frac{1}{2}(\psi_j - \mu_J)^T \mathbf{\Sigma}_J^{-1}(\psi_j - \mu_J)\right], \tag{3.9}$$

with mean vector

$$\mu_J = (\mu_a, \mu_b, \mu_c, \mu_\alpha, \mu_\beta),$$

and covariance matrix

$$\mathbf{\Sigma}_J = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{ac} & \sigma_{a\alpha} & \sigma_{a\beta} \\ \sigma_{ba} & \sigma_b^2 & \sigma_{bc} & \sigma_{b\alpha} & \sigma_{b\beta} \\ \sigma_{ca} & \sigma_{cb} & \sigma_c^2 & \sigma_{c\alpha} & \sigma_{c\beta} \\ \sigma_{a\alpha} & \sigma_{\alpha b} & \sigma_{\alpha c} & \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\beta a} & \sigma_{\beta b} & \sigma_{\beta c} & \sigma_{\beta\alpha} & \sigma_\beta^2 \end{pmatrix}.$$

Combining (3.28), (3.29), (3.30) and (3.9), the author arrives at the final model:

$$f(\mu, \mathbf{t}, \xi, \psi) = \prod_{n=1}^{N} \prod_{j=1}^{J} f(\mu_n, \mathbf{t}_n, \xi_n, \psi_j) f(\xi_n, \mu_p, \mathbf{\Sigma}_p) f(\psi_j, \mu_J, \mathbf{\Sigma}_J). \tag{3.10}$$

3.3 Proportional Hazards Model for Response Times

Survival analysis is a branch of statistics which concerns the analysis time-to-event data. In educational testing, the event is often the time required for an examinee to answer a specific question. All of the parametric models mentioned above can be generalized to regression models in which some parameter of the distribution depends on θ usually through some linear function, that is then nonlinearly linked to the parameter of interest. The standard approach in biostatistics to make such models more flexible is to combine

the parametric regression term with a non-parametrically defined baseline hazard function, and express the hazard function as

$$h(t \mid \theta) = h_0(t)e^{\gamma\theta}. \tag{3.11}$$

The hazard function is the instantaneous rate at which response times occur, and is comprised of a non-parametrically specified nonnegative baseline hazard function h_0 and a term that depends on the covariate, which is the latent trait θ in the IRT case. This model, known as the Cox proportional hazards model (Cox, 1972), is widely used for inference about the parameter γ, and sometimes estimating the entire conditional survival curves is of interest. Another commonly used measure in survival analysis is the survival function $S(t|\theta)$, which captures the probability that an event will survive beyond a specified time. $S(t|\theta)$ relates to $h(t \mid \theta)$ in (3.11) through:

$$S(t|\theta) = \exp[-H(t \mid \theta)],$$

where $H(t \mid \theta) = \int_{s=0}^{t} h(s|\theta)ds$ is the cumulative hazard function. Given $S(t|\theta)$, the probability density function $f(t \mid \theta)$ is readily available as $f(t \mid \theta) = -\frac{dS(t|\theta)}{dt}$. By (3.11), for the jth item, we model the hazard function of t_{nj}, the response time required for test taker n to answer the jth item by

$$h_j(t_{nj}|\theta_n) = h_{0j}(t_{nj})e^{\gamma_j\theta_n}. \tag{3.12}$$

We need to stress that, the above models and the model for responses that we shall introduce later all rely on the assumption of the **local independence (LI)**. Specifically, we postulate the following assumptions hold:

- Independence between responses given θ . Mathematically, this assumption is defined as

$$f(y_{n1}, \cdots, y_{nG}|\theta_n) = \prod_{g=1}^{G} f(y_{ng}|\theta_n), \tag{3.13}$$

  for $\theta_n \in \mathbb{R}$, $1 \leq n \leq N$ and each possible subset of items of size $G \leq J$, where $f(y_{n1}, \cdots, y_{nG}|\theta_n)$ and $f(y_{ng}|\theta_n)$ denote the probability functions of the responses on the subset of items and the individual items in it, respectively.

- Independence between response times given θ . This assumption is defined as

$$f(t_{n1}, \cdots, t_{nG}|\theta_n) = \prod_{g=1}^{G} f(t_{ng}|\theta_n), \tag{3.14}$$

where $f(t_{n1}, \cdots, t_{nG}|\theta_n)$ and $f(t_{ng}|\theta_n)$ are the densities of the response times on the subset of items and the individual items in it, respectively.

- Independence between responses and response times given θ. That is

$$f(y_{n1}, \cdots, y_{nG_1}; t_{n1}, \cdots, t_{nG_2}|\theta_n) = \prod_{g_1=1}^{G_1} f(y_{ng_1} \mid \theta_n) \prod_{g_1=1}^{G_2} f(t_{ng_2} \mid \theta_n), \tag{3.15}$$

for $\theta \in \mathbb{R}$, $1 \leq n \leq N$ and each possible subset of items of size $G_1, G_2 \leq J$.

Partial Likelihood

The goal of our investigation is to accurately estimate θ as well as the parametric and semiparametric terms of the response time model of (3.12). Note that in CAT, every test taker is given different items, based on his or her adaptively estimated θ level. So the random sampling of θ from a common distribution can not be assumed. Consequently, the usual marginal likelihood approaches used in latent variable modeling must be modified. Furthermore, because of the infinite-dimensional parameter h_0, the ordinary likelihood function is untractable, and we must use the so-called **partial likelihood** proposed by Cox (1975), to estimate γ and to supplement the IRT likelihood for θ.

For the jth item, suppose that there are no ties between the event times. Let $t_{(1j)} < t_{(2j)} < \cdots < t_{(Nj)}$ denote the ordered event times and $\theta_{(i)}$ be the latent trait associated with the individual whose response time is $t_{(ij)}$. Define the risk set $R(t_{(pj)})$ at time $t_{(pj)}$, $1 \leq p \leq N$ as the set of all individuals who have not answered the question yet, i.e., $R(t_{(pj)}) = \{t_{((p+1)j)}, \cdots, t_{(Nj)}\}$. Based on the hazard function as specified in (3.12), the partial likelihood function for the jth item given θ is specified as:

$$\begin{aligned} L(\gamma_j|\theta) &= \prod_{n=1}^{N} \frac{\exp[\gamma_j \theta_{(n)}]}{\sum_{t_{(pj)} \in R(t_{(pj)})} \exp[\gamma_j \theta_{(p)}]} \\ &= \prod_{n=1}^{N} \frac{\exp[\gamma_j \theta_{(n)}]}{\sum_{p \geq j}^{N} \exp[\gamma_j \theta_{(p)}]} \end{aligned} \tag{3.16}$$

The partial likelihood for the vector $\gamma = (\gamma_1, \cdots, \gamma_J)'$ is then defined as

$$L(\gamma|\theta) = \prod_{j=1}^{J} L(\gamma_j|\theta). \tag{3.17}$$

The log of the partial likelihood is $LL(\gamma_j|\theta) = \ln[L(\gamma_j|\theta)]$, and write $LL(\gamma_j|\theta)$ as

$$LL(\gamma_j|\theta) = \sum_{n=1}^{N} \gamma_j \theta_{(n)} - \ln[\sum_{p\geq j}^{N} \exp(\gamma_j \theta_{(p)})]. \tag{3.18}$$

Taking derivatives with respect to γ we find the score to be $U(\gamma_j|\theta) = \partial LL(\gamma_j|\theta)/\partial\gamma_j$. Then

$$U(\gamma_j|\theta) = \sum_{n=1}^{N} \theta_{(n)} - \frac{\sum_{p\geq j}^{N} \theta_{(p)} \exp[\gamma_j \theta_{(p)}]}{\sum_{p\geq j}^{N} \exp[\gamma_j \theta_{(p)}]}. \tag{3.19}$$

Taking derivatives again and changing sign we find the observed information to be

$$I(\gamma_j|\theta) = \frac{\sum_{p\geq j}^{N} \theta_p^2 \exp[\gamma_j \theta_{(p)}]}{\sum_{p\geq j}^{N} \exp[\gamma_j \theta_{(p)}]} - \left[\frac{\sum_{p\geq j}^{N} \theta_{(p)} \exp[\gamma_j \theta_{(p)}]}{\sum_{p\geq j}^{N} \exp[\gamma_j \theta_{(p)}]}\right]^2. \tag{3.20}$$

3.3.1 Parameter Estimation

Suppose the items are indexed by $j = 1, \cdots, J$, and the examinees by $n = 1, \cdots, N$. For the nth test taker, his or her responses and response times are denoted by $\boldsymbol{Y_n} = (Y_{1n}, \cdots, Y_{Jn})'$, and $\boldsymbol{T_n} = (T_{1n}, \cdots, T_{Jn})'$, respectively. We model the jth item's hazard function by (3.12) and specify the partial likelihood function by (3.16). We assume a two-parameter IRT (2PL) model for the response variable Y_n:

$$P_j(\theta_n) = \frac{\exp[a_j(\theta_n - b_j)]}{1 + \exp[a_j(\theta_n - b_j)]}. \tag{3.21}$$

Then the likelihood function for the nth subject can be specified as:

$$\text{IRT}(\theta_n) = \prod_{j=1}^{J} P_j(\theta_n)^{y_{nj}} (1 - P_j(\theta_n))^{1-y_{nj}}. \tag{3.22}$$

To estimate the parameters $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_J)'$, note that in CAT it would generally be the case that different examinees take different items, and the items they take are closely associated with their ability level θ. That means, some on-the-shelf method for marginal likelihood estimation or a frailty model procedure will not work. So in our investigation, MCMC is used to take advantage of the previously calibrated items parameters as well as an additional source of information about the latent covariate in the model, which is theta. The algorithm below, immediately generalizes to the CAT situation, in which different people take different items.

Markov Chain Monte Carlo

To explore the posterior distribution of the parameters in the models (3.18) and (3.22), we use a Bayesian MCMC method (Metropolis-Hastings). Here we assume a CAT setting in which item parameters are previously calibrated and are taken as known. Our concern is utilizing the response time information to estimate paramters of the response time distributions and also obtain more information for the estimation of θ. We assume the response time variables are independent (Local independence) given the person parameter θ.

Metropolis-Hastings

We describe the Metropolis-Hastings algorithm in the following steps:

Step 1: Denote the initial values for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\gamma}}_0 \equiv (\hat{\gamma}_{01}, \cdots, \hat{\gamma}_{0J})'$ and $\hat{\boldsymbol{\theta}}_0 \equiv (\hat{\theta}_{01}, \cdots, \hat{\theta}_{0N})'$, respectively. $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimator (MLE) by maximizing the likelihood function in (3.22). Denote by σ_θ^2 the sample variance of $\hat{\boldsymbol{\theta}}_0$, and σ_γ^2 the sample variance of $\hat{\boldsymbol{\gamma}}_0$. Conditioning on $\hat{\boldsymbol{\theta}}_0$, $\hat{\boldsymbol{\gamma}}_0$ is obtained by maximizing the partial likelihood function defined in (3.18). Set the iteration counter iter = 1.

Step 2: At rth step, Denote the previous positions $\boldsymbol{\theta}^{(r-1)} \equiv (\theta_1^{(r-1)}, \cdots, \theta_N^{(r-1)})'$ and $\boldsymbol{\gamma}^{(r-1)} \equiv (\gamma_1^{(r-1)}, \cdots, \gamma_J^{(r-1)})'$. We first sample the person parameters $\boldsymbol{\theta}$. For the nth person, draw θ_n^* from a normal distribution $N(\theta_n^{r-1}, \sigma_\theta^2)$ and denote $\boldsymbol{\theta}^* \equiv (\theta_1^{(r-1)}, \cdots, \theta_n^*, \cdots, \theta_N^{(r-1)})'$. The acceptance probability for θ_n^* is defined as:

$$\alpha(\theta_n^{(r-1)}, \theta_n^*) \equiv \min\left\{1, \frac{\mathrm{IRT}(\theta_n^*) L(\boldsymbol{\gamma}^{(r-1)}|\boldsymbol{\theta}^*)\pi(\theta_n^*)}{\mathrm{IRT}(\theta_j^{(r-1)}) L(\boldsymbol{\gamma}^{(r-1)}|\boldsymbol{\theta}^{(r-1)})\pi(\theta_n^{(r-1)})}\right\},\ 1 \le n \le N, \tag{3.23}$$

where $\pi(\theta_n^*) = \frac{1}{\sqrt{2\pi}} \exp\frac{(\theta_n^*)^2}{2}$, and $\pi(\theta_n^{(r-1)}) = \frac{1}{\sqrt{2\pi}} \exp\frac{(\theta_n^{(r-1)})^2}{2}$. $\mathrm{IRT}(\cdot)$ and $L(\cdot)$ are defined in (3.22) and (3.17) respectively. Generate a uniform $[0,1]$ quantity u. If $u \le \alpha(\theta_n^{(r-1)}, \theta_n^*)$, the move is accepted and $\theta_n^r = \theta_n^*$. If $u > \alpha(\theta_n^{(r-1)}, \theta_n^*)$, the move is not accepted, and $\theta_n^r = \theta_n^{(r-1)}$.

After obtaining the person parameters $\boldsymbol{\theta}^r$, we then draw item parameters $\boldsymbol{\gamma}^r$. Analogous to the previous steps, for jth item, draw γ_j^* from a normal distribution $N(\gamma_j^{r-1}, \sigma_\gamma^2)$. The acceptance probability for γ_j^* is defined as:

$$\alpha(\gamma_j^{(r-1)}, \gamma_j^*) \equiv \min\left\{1, \frac{L(\gamma_j^*|\boldsymbol{\theta}^r)}{L(\gamma_j^{(r-1)}|\boldsymbol{\theta}^r)}\right\},\ 1 \le j \le J, \tag{3.24}$$

where $L(\cdot)$ is defined in (3.16). Generate a uniform $(0,1)$ quantity u. If $u \le \alpha(\gamma_j^{(r-1)}, \gamma_j^*)$, the move is accepted and $\gamma_j^r = \gamma_j^*$. If $u > \alpha(\gamma_j^{(r-1)}, \gamma_j^*)$, the move is not accepted, and $\gamma_j^r = \gamma_j^{(r-1)}$.

step 3: Change the iteration counter from r to $r+1$ and return to step 1 until iter $= M$, where M is a pre-specified number.

A burn-in period of the initial K iterations is often required to allow the chain to reach equilibrium. Finally, we obtain two Markov chains for the person and the item parameters, which are denoted by $\boldsymbol{\theta}_{\text{mc}} \equiv (\boldsymbol{\theta}^{(K+1)}, \cdots, \boldsymbol{\theta}^{M})'$ and $\boldsymbol{\gamma}_{\text{mc}} \equiv (\boldsymbol{\gamma}^{(K+1)}, \cdots, \boldsymbol{\gamma}^{M})'$, respectively. Define the posterior sample means for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ as $\bar{\boldsymbol{\gamma}} \equiv \bar{\boldsymbol{\gamma}}_{\text{mc}} = \frac{1}{M-K}\sum_{t=K+1}^{M}\boldsymbol{\gamma}^{t}$ and $\bar{\boldsymbol{\theta}} \equiv \bar{\boldsymbol{\theta}}_{\text{mc}} = \frac{1}{M-K}\sum_{t=K+1}^{M}\boldsymbol{\theta}^{t}$.

Nonparametric Estimation of the Baseline Hazard

Having obtained the regression coefficients $\hat{\boldsymbol{\gamma}}$, we now consider the estimation of the baseline hazard $\{h_{0j},\ 1 \leq j \leq J\}$ defined in (3.12), or equivalently the cumulative baseline hazard rate $H_{0j}(t) = \int h_{0j}(u)du$. For the jth item, in order to estimate h_{0j}, we express the complete profile likelihood as

$$\begin{aligned} L(\gamma_j, h_{0j}(t)) &= \prod_{n=1}^{N} f(t_{nj} \mid \theta_n) = \prod_{n=1}^{N} -\frac{dS(t_{nj} \mid \theta_n)}{dt_{nj}} \\ &= \prod_{n=1}^{N} -\frac{d\exp(-H(t_{nj} \mid \theta_n))}{dt_{nj}} = \prod_{n=1}^{N} h(t_{nj} \mid \theta_n)S(t_{nj} \mid \theta_n) \\ &= \prod_{n=1}^{N} h_{0j}(t_{nj})\exp(\gamma_j\theta_n)\exp[-H_{0j}(t_{nj})\exp(\gamma_j\theta_n)]. \end{aligned}$$

Fix γ_j by its estimator $\hat{\gamma}_j$ and consider maximizing the above likelihood as a function of $h_{0j}(t)$ only. The function to be maximized can be written as

$$L(h_{0j}(t) \mid \hat{\gamma}_j) = \left[\prod_{n=1}^{N} h_{0j}(t_{nj})\exp(\hat{\gamma}_j\theta_{(n)})\right]\exp\left[-\prod_{n=1}^{N} H_{0j}(T_j)\exp(\hat{\gamma}_j\theta_n)\right]. \tag{3.25}$$

Note that (3.25) is maximized when $h_0(t) = 0$ except for times at which the events occurs. So $H_0(T_j) = \sum_{t_i \leq T_j} h_{0j}(t_i)$, and (3.25) is maximized when

$$\hat{h}_{0j}(t_i) = \frac{1}{\sum_{p \geq j}^{N}\exp[\hat{\gamma}_j\theta_{(p)}]}.$$

So

$$\hat{H}_{0j}(t) = \sum_{t_i \leq t}\frac{1}{\sum_{p \geq j}^{N}\exp[\hat{\gamma}_j\theta_{(p)}]}, \tag{3.26}$$

or

$$\hat{S}_{0j}(t) = \exp[-\hat{H}_{0j}(t)].$$

This is the Breslow estimator of the baseline cumulative hazard function suggested by Breslow (1972). In the survival analysis literature, there are other types of nonparametric estimators of the survival function

for example the Kaplan-Meier estimator.

3.3.2 Simulation Studies

The simulation design for the performance of estimation of latent traits utilized 25 independent replications in a design that varied test length J and the number of test takers N. For all simulations we set the sample size at $J = 100, 400$, and studied test lengths of $J = 20, 100$. The latent variable θ was drawn from $N(0, 1)$, the standard normal distribution; items parameters of a two-parameter logistic model were drawn from distributions, $\beta_1 \sim N(0, 1)$ and $\beta_0 \sim U[1, 2.5]$. An exponential baseline hazard function $h(\cdot) = \lambda$ with various item parameters λ was chosen to generate response times. λs were drawn from a uniform distribution $\lambda \sim U[0.25, 1.5]$. The log hazard ratios γs in (3.12) were drawn from $U[0.5, 1.5]$. To implement the Bayesian MCMC algorithm, chains of length 4000 with an initial burn-in period = 1000 were chosen.

3.4 A New CAT Item Selection Strategy: Maximum Item Information per Time Unit

3.4.1 Computerized Adaptive Testing

Computerized adaptive testing (CAT) is a desirable format for testing because it can tailor items to the ability of the examinee to more quickly obtain an accurate estimate of the examinee's ability. A predominant estimator of ability is the maximum likelihood estimator $\hat{\theta}^{mle}$, which is the value of θ that maximizes the conditional likelihood function of the responses. Under smoothness conditions on the item response functions, $\hat{\theta}^{mle}$ is asymptotically $N(\theta_0, I^{-1}(\theta_0))$, where θ_0 is the true value of θ and $I(\theta_0)$ is the Fisher information at θ_0. The function $I(\theta)$ is a useful measure of the precision with which the J items can serve to measure θ, as a function of θ. Denote $s(\theta)$ the score function, the first derivative of the log-likelihood, then

$$I(\theta) = E[s^2(\theta)] = -E[\frac{ds(\theta)}{d\theta}] = \sum_{j=1}^{J} I_J(\theta),$$

where $I_j(\theta)$ is known as the **item information function**.

The general form of an item information function is given by

$$I_j(\theta) = \frac{[P_j'(\theta)]^2}{P_j(\theta)Q_j(\theta)}$$

where P_j denotes the item response function of the jth item and $Q_j = 1 - P_j$. In the case of a 3-PL model,

$$I_j(\theta) = a_j^2 \frac{Q_j}{P_j} \left[\frac{P_j - c_j}{1 - c_j} \right]^2 .$$

Because the variance of $\hat{\theta}^{mle}$ is inversely related to the Fisher information, it motivates the procedure of selecting items to maximize Fisher information at the current ability estimate. Specifically, index the bank of all possible items by $j = 1, 2, ..., J$, and suppose that m items have been administered, $Y_{j_1}, Y_{j_2}, ..., Y_{j_m}$. Let $S_m = \{j_1, j_2, ..., j_m\}$ denote the indices for these items, and let $R_m = \{1, 2, ..., N\} \cap \bar{S}_m$ denote the remaining items. We wish to find the "best" item to give next. Let $\hat{\theta}_m^{mle}$ denote the current ability estimate. The **maximum information criterion** (MIC) selects the item which has highest information at $\hat{\theta}_m^{mle}$. In other words,

$$j_{m+1} = \max_l \{ I_l(\hat{\theta}_m^{mle}) : l \in R_m \}. \tag{3.27}$$

Chang and Ying (1996) provide an alternative to the MIC by defining information in a different way. Note that test information $I(\theta)$ is a measure of local information around the true value θ_0. The usefulness of maximizing $I(\hat{\theta}_m^{mle})$ is questionable early in the sequence when $\hat{\theta}_m^{mle}$ may be quite far from θ_0 for small m. To address this they propose a more global definition of information. Let $Y_1, Y_2, ..., Y_N$ be responses to the items of a test, and let $L(\theta)$ denote the likelihood function. If our aim is to distinguish a value θ_0 from another value θ^*, the likelihood ratio $L(\theta_0)/L(\theta^*)$ should provide useful information. In fact, Neyman-Pearson theory tells us that the likelihood ratio method is optimal for testing $\theta = \theta_0$ versus $\theta = \theta^*$. With this notion in mind, we define **Kullback-Leibler item information** as follows

$$K_j(\theta \mid \theta_0) = E_{\theta_0} \left[log \frac{L_j(\theta_0)}{L_j(\theta)} \right]$$

where $L_j(\theta) = P_j(\theta)^{y_j} Q_j(\theta)^{1-y_j}$ is the factor that the jth item response contributes to the conditional likelihood function.

This leads to the calculation

$$K_j(\theta \mid \theta_0) = P_j(\theta_0) \left[log \frac{P_j(\theta_0)}{P_j(\theta)} \right] + Q_j(\theta_0) \left[log \frac{Q_j(\theta_0)}{Q_j(\theta)} \right].$$

For the entire test the **Kullback-Leibler test information** is defined by

$$K(\theta \mid \theta_0) = E_{\theta_0} \left[log \frac{L(\theta_0)}{L(\theta)} \right].$$

By conditional independence,

$$K(\theta \mid \theta_0) = \sum_{j=1}^{N} K_j(\theta \mid \theta_0).$$

The function $K(\cdot)$ is global as opposed to $I(\cdot)$ in that for a fixed θ_0, K is a function of θ and I is a fixed number. Chang and Ying proposed incorporating $K(\cdot)$ into CAT as follows. Using the notation above and a current estimate $\hat{\theta}_m$, usually $\hat{\theta}_m^{mle}$ when it exists, select item $m+1$ according to **global information criterion** (GIC)

$$j_{m+1} = \max_l \left\{ \int_{\hat{\theta}_m - \delta_m}^{\hat{\theta}_m + \delta_m} K_l(\theta \mid \hat{\theta}_m) d\theta : l \in R_m \right\}.$$

The sequence δ_m should go to 0, and is set equal to $c/\sqrt{m}$. Here c should be large enough so that the interval $(\hat{\theta}_m - \delta_m, \hat{\theta}_m + \delta_m)$ has a high probability of containing δ_0. If δ_m is very small, the GIC is essentially equivalent to the MIC. However, if δ_m is defined as above, the GIC is global in that it is largely influenced by the tails of $K_l(\theta \mid \hat{\theta}_m)$.

Simulation studies revealed that this method outperformed MIC early in the sequence, indicating it would be a better choice for short adaptive tests. Also, because it would not always select the item with the highest discrimination parameter at every difficulty level, it takes some steps towards addressing item exposure. However, neither the MIC nor the GIC adequately address balancing item exposure, and have no features for satisfying test constraints. Next we consider a parametric model for response times, and propose a modification of the MIC that can be used to more effectively accrue information.

3.4.2 A Lognormal Model for Response Times

The items are indexed by $j = 1, \cdots, J$, and the examinees by $n = 1, \cdots, N$. For the nth test taker, his or her responses and response times are denoted by $\mathbf{Y}_n = (Y_{1n}, \cdots, Y_{Jn})'$, and $\mathbf{T}_n = (T_{1n}, \cdots, T_{Jn})'$. Assume the item responses follow a 3-PL IRT model. The probability that the response of the nth subject to the jth variable is correct, given $\theta_n \in (-\infty, \infty)$, the level of the latent trait of the nth subject can be expressed as

$$P_j(\theta_n) = P(Y_{nj} = 1 | \theta_n) = c_j + (1 - c_j) \frac{e^{a_j(\theta_n - b_j)}}{1 + e^{a_j(\theta_n - b_j)}}, \tag{3.28}$$

where c_j is the guessing parameter for the jth item. A lognormal model proposed by van der Linden is used to model the response time t_{nj} of the nth person on the jth item:

$$T_{nj} \sim f(t_{nj}; \tau_n, \alpha_j, \beta_j) \equiv \frac{\alpha_j}{t_{nj}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_j(\ln t_{nj} - (\beta_j - \tau_n))\right]^2\right\}, \tag{3.29}$$

where τ_n, β_j and α_j are the speed parameter for examinee n, the time intensity and discriminating power of item j, respectively.

At the second level, denote by $\xi_n = (\theta_n, \tau_n)$, $\psi_j = (a_j, b_j, c_j, \alpha_j, \beta_j)$ the person parameter for the nth person and item parameter for the jth item, respectively. ξ_n is assumed to be randomly drawn from a multivariate normal distribution,

$$\xi_n \sim f(\xi_n; \mu_p, \mathbf{\Sigma}_p) \equiv \frac{|\mathbf{\Sigma}_p^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\xi_n - \mu_p)^T \mathbf{\Sigma}_p^{-1}(\xi_n - \mu_p)\right], \tag{3.30}$$

with mean vector

$$\mu_p = (\mu_\theta, \mu_\tau),$$

and covariance matrix

$$\mathbf{\Sigma}_p = \begin{pmatrix} \sigma_\theta^2 & \sigma_{0\tau} \\ \sigma_{0\tau} & \sigma_\tau^2 \end{pmatrix}.$$

3.4.3 Item Information per Time Unit

In Section 3.4.1, we reviewed a widely used item selection strategy known as the maximum information criterion (MIC) to select an item that has the highest information $I(\hat{\theta}^{mle})$. However, MIC does not take into consideration the time required to answer an item. Such information is useful in that often a highly informative item can be quite time consuming, so it has less practical value compared to an equally or somewhat less informative item that requires less time to complete. Specifically, instead of maximizing raw item information $I_l(\hat{\theta}_m^{mle})$ in (3.27), we propose to maximize "item information per time unit"(**MICT**).

Specifically, we choose the next item based on

$$j_{m+1} = \max_l \left\{ \frac{I_l(\hat{\theta}_m^{mle})}{\mathbf{E}[T_l|\hat{\tau}_m^{mle}]} : l \in R_m \right\}, \tag{3.31}$$

where T_l is the time required for the l item and $\hat{\tau}_m^{mle}$ is the maximum likelihood estimator of current the speed parameter τ. Under the lognormal model in (3.29), the expected time to answer the lth item can be

expressed as

$$\mathbf{E}[T_l|\hat{\tau}_m^{mle}] = \exp\left(\beta_l - \hat{\tau}_m^{mle} + \frac{1}{2\alpha_l^2}\right),\ l \in R_m$$

and for the nth person,

$$\hat{\tau}_n = \frac{\sum_{j\in R_m}\alpha_j(\beta_j - \log t_{nj})}{\sum_{j\in R_m}\alpha_j}.$$

3.4.4 Control of Item Exposure Rate

One of the drawbacks to the item information based CAT is that only a small fraction of items that possess high item information are selected. So the MIC and MICT could lead to extremely skewed item exposure rates, i.e., some items could be constantly selected in a CAT whereas others may never be used. Having too few items being used is not cost-effective and can lead to serious test security problems when the relatively small set of items being used becomes compromised to the public. To remedy this, Chang, *et. all.,* (2001) proposed an a-stratified with b blocking method (ASB) for item selection in CAT. The ASB method can be described in the following steps:

1. Arrange the item back according to b values, and then divide the item bank into M equal-length blocks. So the first blocks contains items with the smallest b values, and the Mth block with the highest b values.

2. Further partition each of the M blocks into K strata according to their a values. Thus, in the mth block, the first stratum contains items with the smallest a values, and the Kth stratum with the highest a values.

3. For $k = 1, \cdots, K$, recombine the kth stratum items across M blocks into a single stratum. So there are now K strata.

4. Divide the test into K stages.

5. In the kth stage, select items using Difficulty Matching (DM) procedure, i.e., items are chosen from the kth stratum based on the closeness of b values to the current estimate of θ for an examinee.

6. Repeat Step 5 for $k = 1, 2, \cdots, K$.

The method ASB_DM can be adjusted for time, which we refer to as a-stratification with b blocking and time weighting (ASB_TWDM), by a simple adjustment to the fifth step in the algorithm above. Instead of matching θ to the nearest b, we minimize the product of the absolute difference between θ and b and the expected time to answer the item, given the person's current estimate of the speed parameter.

3.4.5 Simulation Studies

Comparison between MICT and MIC

We wish to simulate a $J = 1000$-item bank with the item parameters a, b, c, α and β defined in (3.28) and (3.29). Assume $a \sim$ uniform$[1, 2.5]$, $c \sim$ beta$(2, 10)$, and $\alpha \sim$ uniform$[2, 4]$. The item difficulty parameter b and time intensity parameter β are assumed to follow a bivariate normal distribution:

$$(b_j, \beta_j) \equiv \psi_j \sim \frac{|\mathbf{\Sigma}_1^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\psi_j - \mu_1)^T \mathbf{\Sigma}_1^{-1}(\psi_j - \mu_1)\right], \tag{3.32}$$

with mean vector $\mu_1 = (0, 0)$, and covariance matrix

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 0.25 \end{pmatrix},$$

or

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

$N = 1000$ person parameters θ_n and τ_n are generated from a bivariate normal distribution:

$$(\theta_n, \tau_n) \equiv \xi_n \sim \frac{|\mathbf{\Sigma}_2^{-1}|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\xi_n - \mu_2)^T \mathbf{\Sigma}_2^{-1}(\xi_n - \mu_2)\right], \tag{3.33}$$

with mean vector $\mu_2 = (0, 0)$, and covariance matrix

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

or

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Each examinee is administered 5 randomly chosen items from the item bank to acquire the initial ability and speed estimates. Then a CAT with MIC and MICT are implemented. Two scenarios are considered. The first scenario is a fixed length exam with 50 items (55 items in total) chosen either by MIC or MICT. For each method, mean squared errors (MSEs) for θ and τ, the average time to complete the 55 items and item exposure rates are calculated. The other scenario is concerned with a fixed Fisher information exam and for each examinee, the exam stops until he/she achieves Fisher information= 40.

Item Exposure Control Study

A fixed test length of 55 items was used. The lognormal model in (3.29) was used to simulate the response times for the examinees. We simulated a 500-item bank with item parameters $a \sim$ uniform$[1, 2.5]$, $c \sim$ beta$(2, 10)$, and $\alpha \sim$ uniform$[2, 4]$. The item difficulty parameter b and time intensity parameter β are assumed to follow a bivariate normal distribution defined in (3.32). A total of 1000 θ and τ values are generated from (3.33). The 3-PL model was used to simulate item responses for each examinee. The lognormal model in (3.29) was used to simulate the response times for the examinees.

We compared the maximum likelihood estimates for θ and τ for both CAT with ASB item exposure control and CAT without item exposure control. We also compared a χ^2 statistic proposed by Chang and Ying (1999), to measure the skewness of the exposure rate distribution,

$$\chi^2 = \sum_{i=1}^{n} \frac{(er_i - L/n)^2}{L/n}, \tag{3.34}$$

where

er_i is the observed exposure rate for the jth item,

$L = 55$ is the test length, and $n = 500$ is the number of the items in the item bank.

For the ABS item exposure control, the item bank is stratified into 5 strata according to the method described in Section (3.4.4). The initial estimates of θ and τ are acquired by randomly selecting 5 items each from 1 different stratum. The remaining items were selected according to Step 5 above, i.e., within the $k$tratum, an item that has the closest b value to the current estimated θ is selected:

$$j_{m+1} = \max_{l} \left\{ \frac{1}{|\theta_m^{\text{mle}} - b_l|} : l \in R_m \right\},$$

where R_m is the set of the remaining items in the kth strata that haven't been administered. Similarly, for the ABS_TWDM method, the next item is select to maximize

$$j_{m+1} = \max_{l} \left\{ \frac{1}{\mathbf{E}[T_l|\hat{\tau}_m^{mle}]|\theta_m^{\text{mle}} - b_l|} : l \in R_m \right\}.$$

3.5 Conclusions

Tables (3.1) and (3.4) listed the means and standard deviations for the estimation of $\boldsymbol{\gamma}$ with $J = 20, N = 100$ and $J = 40, N = 400$, respectively. Both sample sizes resulted in highly accurate estimates. As expected, the standard deviations of the MIMC chains reduced by a half with $J = 40$ and $N = 400$, compared to when $J = 20$ and $N = 100$. Figures (3.1) and (3.4) plot the latent trait estimates $\hat{\boldsymbol{\theta}}$ v.s. the true $\boldsymbol{\theta}$. It appears that both IRT and IRT+proportional hazard functions estimated $\boldsymbol{\theta}$ very well in that, in both cases the scatter plots hug the 45 degree lines very well. From Tables (3.2) and (3.5), the estimation errors (RMSE) for θ are reduced when both the proportional hazard functions and the IRT models are incorporated into the likelihood function. Table (3.3) tabulated the integrated absolute differences between the baseline hazard functions and their Breslow estimators with a sample size of $J = 20, N = 100$. Figure (3.2) shows the baseline hazard function estimate for a particular item. The Breslow estimator appears to reconstruct the baseline cumulative hazard functions well except at the right boundaries. The Markov chains in Figure (3.3) and (3.5) appear to have reached equilibrium, and have small autocorrelations beyond the first couple lags.

Table 3.6 compared between MIC and MICT with N=100 and the length of the item bank $J = 500$. Several covariance structures for the person parameters θ and τ and the item parameters b and β are considered. Two CAT exam scenarios are simulated: fixed length test and viable length test. Mean Squared Errors (MSE) for θ and τ estimations as well as the average time to complete the test are listed. Both MIC and MICT estimated θ and τ very well with small MSEs in both fixed length and viable length tests. As expected, MIC's estimation is sightly better than MICT's due to the design of the maximization paradigm. MICT substantially reduced the average time examinees take to complete a test. It is very desirable because it shows that with the information of response times, we can better control the duration of the exams.
Table (3.7) listed the MSE for θ and β, the average time to complete a test, and χ^2 statistic when a stratification with b blocking (ASB) exposure control is incorporated into the CAT item selection procedure. Without the exposure control, both MIC and MICT yielded very skewed exposure rate distribution with very high χ^2 values. When ASB was in place, the values for χ^2 are reduced considerably for both difficulty matching (DM) and time weighted difficulty matching (TWDM) item selection procedures. DM can be viewed as the counter part of MIC whereas TWDM can be viewed as the counter part of MICT. Furthermore, as expected, when response time information was taken into account, the average time for examinees to complete a test is reduced. it is observed for both TWDM and MICT. Figure (3.5) showed that TWDM_ASB and DM_ASB have more evenly distributed item exposure rates than that of MIC and MICT.

3.5.1 Simulation Results

Table 3.1: Means and standard deviations for γ with $J = 20$, $N = 100$. The results are based on 25 replications, each replication has a MCMC chain of length 4000.

γ	0.81	0.76	1.05	0.56	0.97	0.98	1.31	0.87	1.05	0.67
$\hat{\gamma}$	0.83	0.78	1.09	0.61	0.83	1.01	1.34	0.87	1.09	0.72
sd	0.14	0.14	0.16	0.13	0.15	0.16	0.18	0.15	0.16	0.14
γ	1.12	1.38	0.78	0.9	1.26	1.17	0.7	0.86	0.86	1.19
$\hat{\gamma}$	1.22	1.3	0.84	0.92	1.29	1.21	0.75	0.78	0.75	1.2
sd	0.17	0.18	0.14	0.15	0.18	0.17	0.14	0.14	0.14	0.17

Table 3.2: The rooted mean square errors (RMSE) for the θ estimates from IRT and IRT+response times, with $J = 20$, $N = 100$. The results are based on 25 replications.

	IRT	IRT+Response
RMSE	0.38	0.23

Table 3.3: Integrated absolute difference between the baseline cumulative hazard function $H_0(t)$ and the Breslow estimator: $\int |H_0(t) - \hat{H}_0(t)|dt$, with $J = 20$, $N = 100$. The results are based on 25 replications.

Item	1	2	3	4	5	6	7	8	9	10
Diff	1.39	1.2	1.46	1.15	1.54	1.55	1.41	1.86	1.27	0.89
Item	11	12	13	14	15	16	17	18	19	20
Diff	1.31	2.19	1.76	1.46	1.3	1.6	1.85	2.27	1.95	1.46

Table 3.4: Means and standard deviations for γ with $J = 50$, $N = 400$. The results are based on 25 replications, each replication has a MCMC chain of length 4000.

γ	0.31	0.26	0.55	0.06	0.47	0.48	0.81	0.37	0.55	0.17
$\hat{\gamma}$	0.31	0.26	0.57	0.05	0.48	0.5	0.82	0.38	0.55	0.19
sd	0.05	0.05	0.06	0.05	0.06	0.06	0.07	0.06	0.06	0.05
γ	0.62	0.88	0.28	0.4	0.76	0.67	0.2	0.36	0.36	0.69
$\hat{\gamma}$	0.62	0.89	0.29	0.41	0.81	0.68	0.2	0.33	0.37	0.7
sd	0.06	0.07	0.06	0.06	0.06	0.06	0.05	0.05	0.06	0.06
γ	0.54	0.71	0.54	0.75	0.42	0.17	0.77	0.88	0.55	0.28
$\hat{\gamma}$	0.54	0.71	0.54	0.77	0.43	0.19	0.77	0.89	0.58	0.29
sd	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.07	0.06	0.05
γ	0.49	0.93	0.35	0.95	0.7	0.89	0.18	0.63	0.99	0.13
$\hat{\gamma}$	0.49	0.95	0.35	0.96	0.67	0.9	0.18	0.64	0.99	0.14
sd	0.06	0.07	0.06	0.07	0.06	0.07	0.05	0.06	0.07	0.05
γ	0.33	0.87	0.78	0.83	0.6	0.49	0.78	0.88	0.21	0.31
$\hat{\gamma}$	0.33	0.91	0.78	0.82	0.61	0.5	0.8	0.89	0.21	0.29
sd	0.06	0.07	0.06	0.07	0.06	0.06	0.07	0.07	0.05	0.05

Table 3.5: The rooted mean square errors (RMSE) for the θ estimates from IRT and IRT+response times, with $J = 50$, $N = 500$. The results are based on 25 replications.

	IRT	IRT+Response
RMSE	0.25	0.17

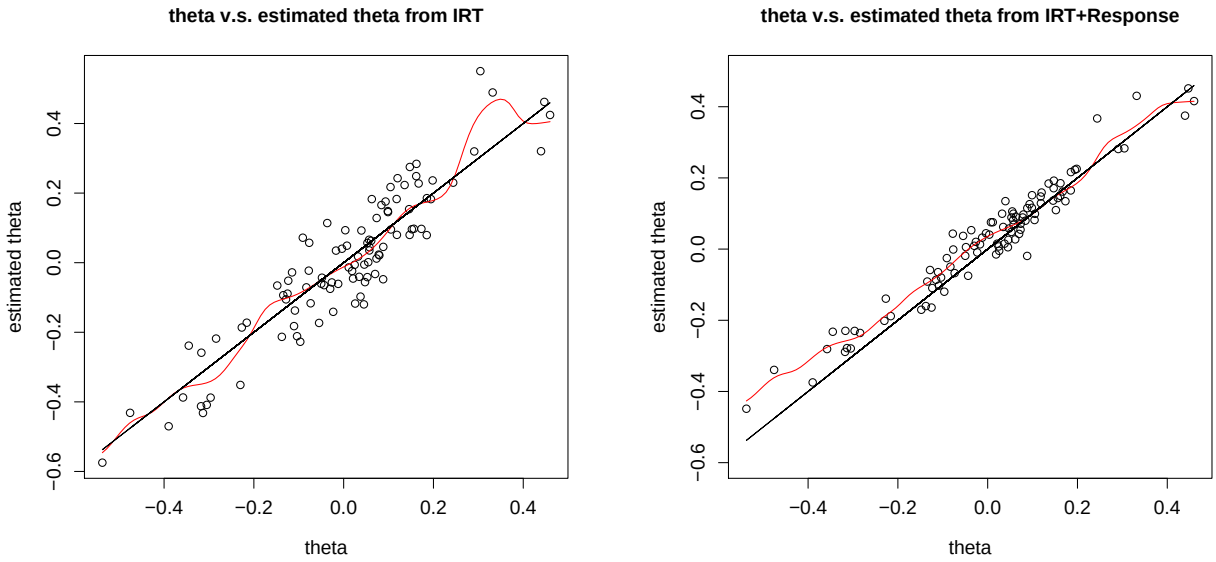

Figure 3.1: θ v.s. estimated θ from both IRT and IRT+response times, with $J = 20$, $N = 100$. The results are based on 25 replications.

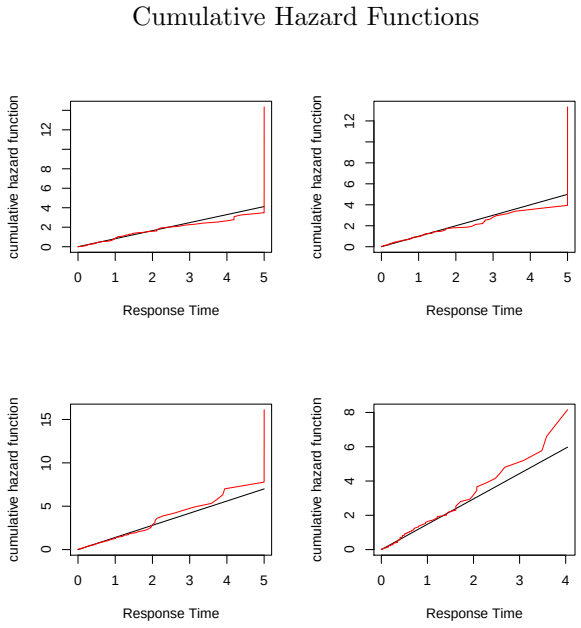

Figure 3.2: Baseline Hazard function v.s. estimated baseline hazard function for selected items, with $J = 20$, $N = 100$.

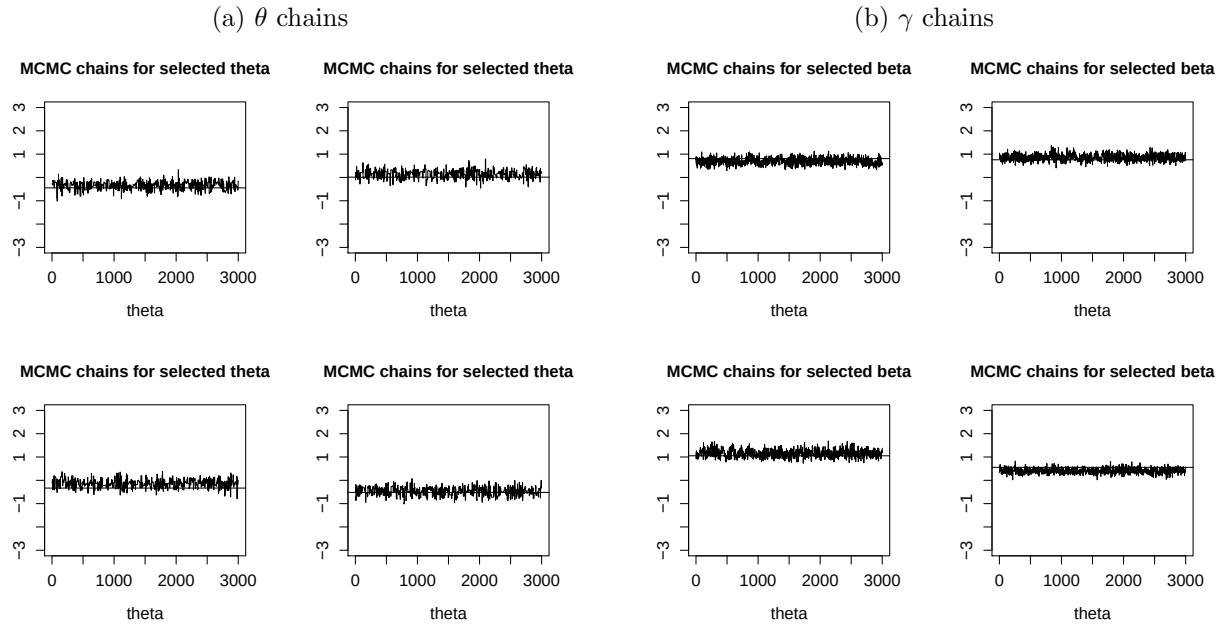

Figure 3.3: Selected MCMC chains for θ and γ $J = 20$, $N = 100$. Each chain has length = 4000 with an burn-in period =1000.

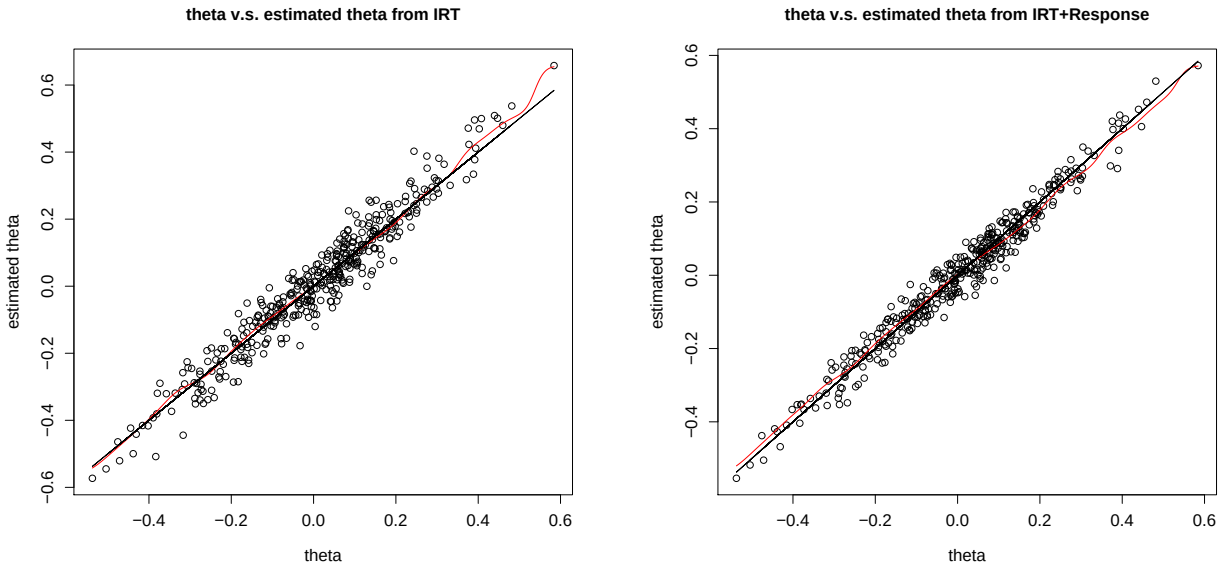

Figure 3.4: θ v.s. estimated θ from both IRT and IRT+response times, with $J = 50$, $N = 400$. The results are based on 25 replications.

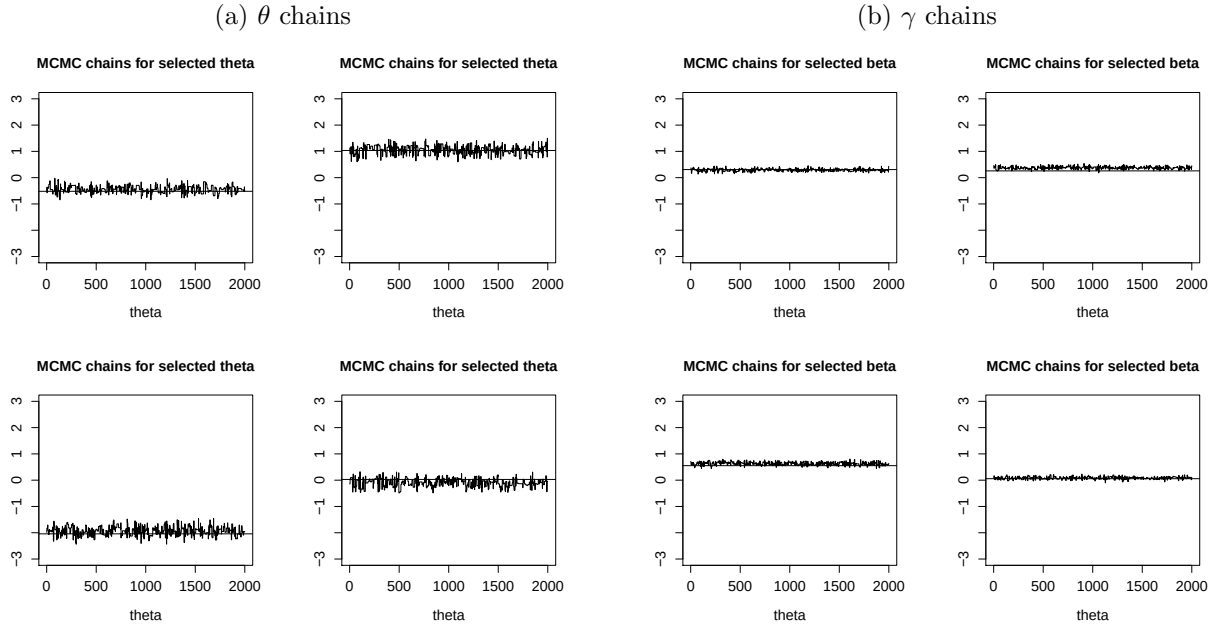

Figure 3.5: Selected MCMC chains for θ and γ with $J = 50$, $N = 400$. Each chain has length = 4000 with an burn-in period =1000.

Table 3.6: Comparison between MIC and MICT. Simulations were based on a 500-item bank, 1000 examinees.

			$\text{cov}(\theta, \tau) = 0$		$\text{cov}(\theta, \tau) = 0.5$	
		Method	Fixed Length	Fixed Fisher Info	Fixed Length	Fixed Fisher Info
$\text{cov}(\text{b}, \beta) = 0$	MSE(θ)	MICT	0.044	0.034	0.052	0.047
		MIC	0.025	0.044	0.026	0.036
	MSE (τ)	MICT	0.002	0.002	0.002	0.002
		MIC	0.002	0.003	0.002	0.003
	Time to complete	MICT	71.636	83.867	75.728	83.127
		MIC	114.965	102.432	114.443	98.671
$\text{cov}(\text{b}, \beta)$=0.5	MSE (θ)	MICT	0.040	0.034	0.028	0.033
		MIC	0.024	0.034	0.025	0.035
	MSE (τ)	MICT	0.002	0.002	0.002	0.002
		MIC	0.002	0.003	0.002	0.003
	Time to complete	MICT	71.292	88.350	69.167	93.682
		MIC	111.752	103.580	100.850	104.112

Table 3.7: Exposure control study. We implemented a-stratification with b blocking (ASB) with Difficulty Matching (DM) and Time Weighted Difficulty Matching (TWDM). Simulation was based on a 500-item bank, 1000 examinees.

	Statistic	TWDM_ASB	DM_ASB	MICT	MIC
$\text{cov}(\theta, \tau) = 0$, $\text{cov}(\text{b}, \beta) = 0$	MSE(θ)	0.036	0.045	0.028	0.024
	MSE(τ)	0.002	0.002	0.002	0.002
	Time to Complete	100.156	115.226	74.885	116.346
	χ^2	9.722	3.881	128.395	93.526
$\text{cov}(\theta, \tau) = 0$, $\text{cov}(\text{b}, \beta) = 0.5$	MSE(θ)	0.036	0.036	0.028	0.025
	MSE(τ)	0.002	0.002	0.002	0.002
	Time to Complete	102.563	117.228	77.716	116.817
	χ^2	9.006	3.533	125.455	93.346
$\text{cov}(\theta, \tau) = 0.5$, $\text{cov}(\text{b}, \beta) = 0$	MSE(θ)	0.038	0.036	0.028	0.024
	MSE(τ)	0.002	0.002	0.002	0.002
	Time to Complete	89.294	103.67	68.181	103.332
	χ^2	9.065	2.252	126.771	93.751
$\text{cov}(\theta, \tau) = 0.5$, $\text{cov}(\text{b}, \beta) = 0.5$	MSE(θ)	0.042	0.040	0.033	0.026
	MSE(τ)	0.002	0.002	0.002	0.002
	Time to Complete	86.725	97.713	66.879	99.945
	χ^2	9.373	4.526	123.571	95.827

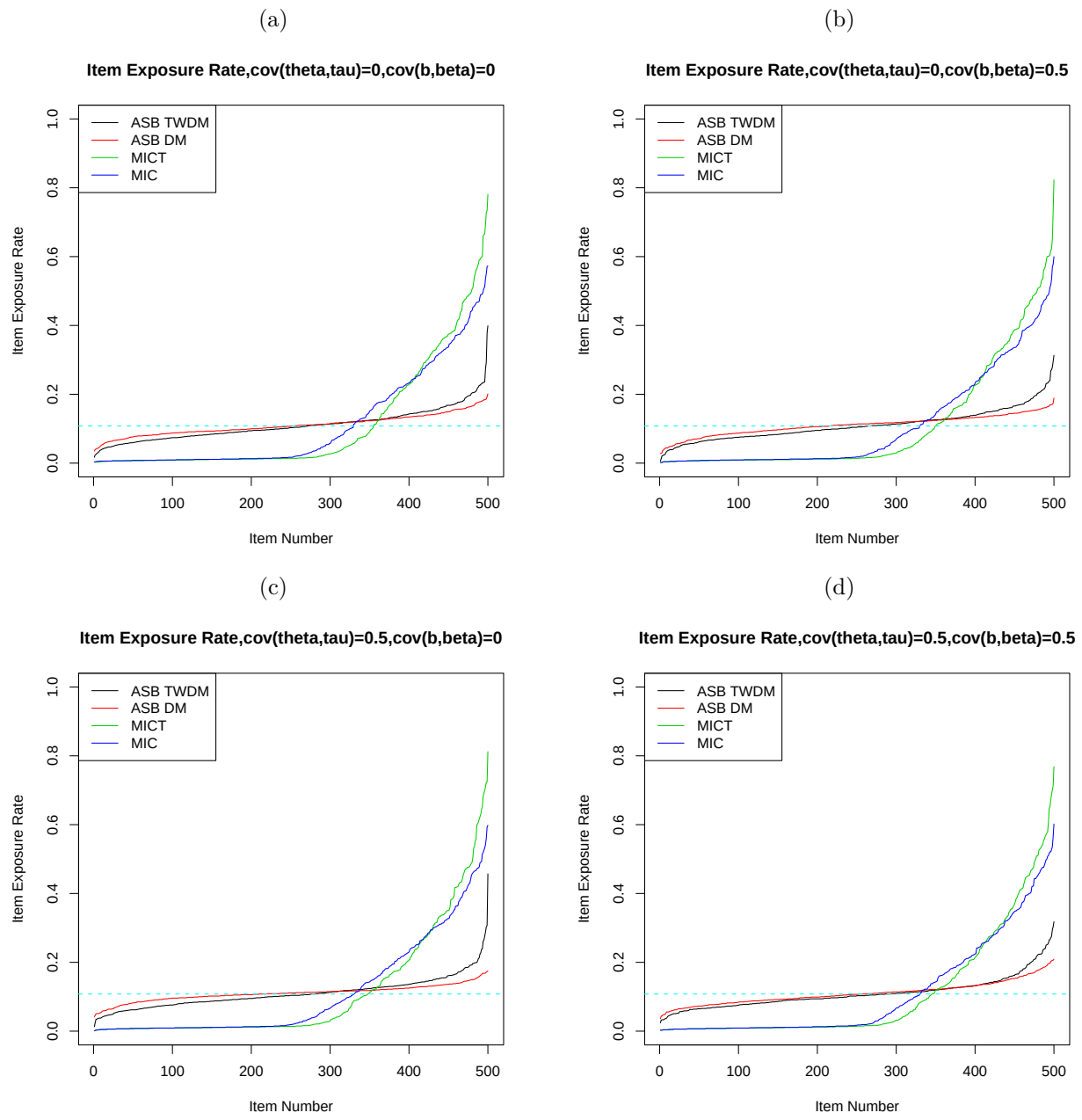

Figure 3.6: Exposure Rate Comparison with a simulated item bank $J = 500$ and $N = 1000$ simulated examinees. fixed length tests with length $L = 55$ are conducted for both MICT and MIC. (a) $\text{cov}(\theta, \tau) = 0, \text{cov}(\text{b}, \beta) = 0$. (b) $\text{cov}(\theta, \tau) = 0, \text{cov}(\text{b}, \beta) = 0.5$. (c) $\text{cov}(\theta, \tau) = 0.5, \text{cov}(\text{b}, \beta) = 0$. (d) $\text{cov}(\theta, \tau) = 0.5, \text{cov}(\text{b}, \beta) = 0.5$.

Appendix A

Proof of Theorems in Cluster Analysis

A.1 Proof of Theorem 2.2.1

The proof proceeds by constructing a sequence of partitions such that the maximum diameter of the solution with M clusters in the space of the latent trait, $\omega_{max,M}$, converges to 0 in probability as M goes to infinity, and then we show that this implies $\omega^*_{max,M}$, the maximum of the cluster diameters under MDP with M clusters, must do the same.

Construct the intervals $[0, 1/M), [1/M, 2/M), ..., [\frac{M-2}{M}, \frac{M-1}{M}), [\frac{M-1}{M}, 1]$. For a partition Π_M, let the cluster C_m include all subjects such that the proportion correct $\frac{1}{J}\sum_{j=1}^{J} Y_{nj}$ falls in the mth of these intervals. If any of these clusters are empty, we redefine π_M by subdividing nonempty clusters to arrive at a total of M nonempty clusters. Now suppose that for some number $\alpha \in (0, 1)$, we have

$$| \theta_n - \theta_{n'} | > \frac{1}{M^\alpha}.$$

Then we show that the probability that they, or any other pair satisfying this inequality, fall in the same cluster goes to 0, indicating that $\omega_{max,M}$ is less than $1/M^\alpha$, which converges to 0 as M goes to infinity..

Without loss of generality, assume $\theta_n > \theta_{n'}$. Then define the expected value of $d^U_{n,n'}$ by

$$\begin{aligned} | P_n - P_{n'} | &= \frac{1}{J} | \sum P_j(\theta_n) - \sum P_j(\theta_{n'}) | \\ &= \frac{1}{J} \sum P_j(\theta_n) - \frac{1}{J} \sum P_j(\theta_{n'}) \\ &\geq C|\theta_n - \theta'_n| \geq \frac{C}{M^\alpha}. \end{aligned} \tag{A.1}$$

The inequality in (A.1) is due to the Mean Value Theorem and the assumption that there is a uniform lower bound on slopes of item characteristic curves over the support of θ, denoted by C.

Denote the observed proportion correct for the nth subject as $\hat{P}_n \equiv \frac{1}{J}\sum_j Y_{nj}$. Then the probability that

subjects n and n' are placed into the same cluster can be bounded above by

$$\begin{aligned} & P\left[\mid \hat{P}_n - \hat{P}_{n'} \mid < \frac{1}{M}\right] = P\left[-\frac{1}{M} < \hat{P}_n - \hat{P}_{n'} < \frac{1}{M}\right] \\ \leq \; & P\left[0 < \hat{P}_n - \hat{P}_{n'} < \frac{1}{M}\right] + P\left[0 < \hat{P}_{n'} - \hat{P}_n < \frac{1}{M}\right] \\ \leq \; & P\left[0 < \hat{P}_n - \hat{P}_{n'} < \frac{1}{M}\right] + P\left[\hat{P}_{n'} - \hat{P}_n > 0\right] = I_1 + I_2. \end{aligned}$$

Then applying Hoeffding's Inequality and (A.1),

$$\begin{aligned} I_2 &= P\left[(\hat{P}_{n'} - \hat{P}_n) - (P_{n'} - P_n) > (P_n - P_{n'})\right] \\ &\leq \exp\left[-\frac{\mid P_n - P_{n'} \mid^2}{2J}\right] \leq \exp\left[-\frac{2C^2 J}{M^{2\alpha}}\right]. \end{aligned}$$

Also by using Hoeffding's inequality, we can bound I_1.

$$\begin{aligned} I_1 &\leq P\left[(\hat{P}_n - P_n) + (P_n - P_{n'}) + (P_{n'} - \hat{P}_{n'}) < \frac{1}{M}\right] \\ &= P\left[(\hat{P}_n - P_n) + (P_{n'} - \hat{P}_{n'}) < \frac{1}{M} - (P_n - P_{n'})\right] \end{aligned} \tag{A.2}$$

For big enough M (A.11) is less than

$$\begin{aligned} &< P\left[(\hat{P}_n - P_n) + (P_{n'} - \hat{P}_{n'}) < -\frac{C}{2M^\alpha}\right] \\ &< P\left[(\hat{P}_n - P_n) > \frac{C}{4M^\alpha}\right] + P\left[(\hat{P}_{n'} - P_{n'}) > \frac{C}{4M^\alpha}\right] \\ &\leq 2\exp\left[-\frac{C^2 J}{16M^{2\alpha}}\right]. \end{aligned}$$

Then the probability that there is any pair with θ values farther than $1/M^\alpha$ that fall into the same cluster is bounded by

$$N(N-1)\exp\left[-\frac{DJ}{M^{2\alpha}}\right], \tag{A.3}$$

for some positive constant D. Then (A.3) goes to 0 under the assumptions of theorem 2.2.1.

The method of proof given above shows that any sequence of partitions with $d^U_{max,M} \leq 1/M$ results in $\omega_{max,M}$ converging to 0 at a

A.2 Proof of Theorem 2.4.1

The proof proceeds much like the proof of Theorem 1. It is shown that a sequence of partitions of a given maximum diameter can be constructed with a probability converging to 1. Furthermore, it is shown that any partition having such a maximum diameter will result in the maximum diameter in the latent variable space converging to 0 with probability 1. Because we have constructed one such partition, we know the MDP must also satisfy this property.

Let $\boldsymbol{W}_n = (W_{n1}, \cdots, W_{nK})'$ be a $1 \times K$ vector of summed scores for the nth examinee and define $\boldsymbol{V}_n = \frac{1}{\sqrt{J}} \boldsymbol{Q}'_{J\times K} \boldsymbol{Y}_n$ to be a rescaled version of $\boldsymbol{W}_n$. Under the model assumptions and when conditioning on $\boldsymbol{\theta}_n$,

$$\boldsymbol{V}_n \mid \boldsymbol{\theta}_n \sim N\left(\frac{1}{\sqrt{J}} \boldsymbol{Q}' \boldsymbol{\beta} \boldsymbol{\theta}_n, \boldsymbol{\Sigma}\right),$$

for some covariance matrix $\boldsymbol{\Sigma}$. Because $\boldsymbol{\theta}_n$ is a latent random variable with mean $\mathbf{0}$, the unconditional mean of $\boldsymbol{V}_n$ is $\mathbf{0}$.

In order to determine how many clusters might be needed, we first wish to find a sequence C_n, such that for each k,

$$P[\max_n |V_{nk}| > C_N] \to 0, \tag{A.4}$$

where $V_{nk} = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} q_{jk} Y_{nj}$.

$$\begin{aligned} P[\max_n |V_{nk}| > C_N] &\leq NP[|V_{nk}| > C_N] \\ &\leq NP[Z > C_N/\sqrt{2}] = N\Phi[-C_N/\sqrt{2}], \end{aligned}$$

where $\Phi(\cdot)$ is the normal cumulative distribution function. Note that for $t \geq 3$, $\Phi(-t) \sim \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2}$. So when $C_N = \sqrt{\log N}$, (A.4) is satisfied.

For each coordinate k, Construct the intervals $[-C_N, -C_N+2C_N/M), [-C_N+2C_N/M, C_N+4C_N/M), ..., [C_N-4C_N/M, C_N - 2C_N/M), [C_N - 2C_N/M, C_N]$ of width $L_n = \frac{2C_N}{M}$. For a partition Π_M, let the cluster C_m include all subjects whose $\boldsymbol{V}$ falls in the mth of the cubicles formed by taking the Cartesian product of these

intervals over the K dimensions of $\boldsymbol{V}$. If any of these clusters are empty, we redefine Π_M by subdividing nonempty clusters to arrive at a total of M nonempty clusters. Thus Π_M will include all of the data with probability going to 1. Now we must show that any partition with equal or smaller diameters will only include points that are near one another in the latent variable space, with probability converging to 1. Note that, the maximum distance of any two points within a cluster is less than the diagonal $\sqrt{K}L_n = \frac{2\sqrt{K}C_N}{M}$. Now suppose that for some number $\alpha \in (0,1)$, we have

$$\omega_{nn'} = ||\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n'}|| > \frac{2\sqrt{K}C_N}{M^\alpha}. \tag{A.5}$$

Let $\boldsymbol{U}_{nn'} = (U_{nn'1}, \cdots, U_{nn'K})' = [\boldsymbol{V}_n - \boldsymbol{V}_{n'}]$ and define the Euclidean distance between two $\boldsymbol{V}$s as

$$d^V_{nn'} = \sqrt{\sum_{k=1}^{K} U^2_{nn'k}}.$$

Because $\mathbf{E}[\boldsymbol{V}_n] = \frac{1}{\sqrt{J}}\boldsymbol{Q'\beta\theta_n}$, $\mathbf{E}[\boldsymbol{U}_{nn'}] = \frac{1}{\sqrt{J}}\boldsymbol{Q'\beta}[\boldsymbol{\theta} - \boldsymbol{\theta_{n'}}]$. Assume that $\boldsymbol{Q'\beta}$ is full rank such that $E[\boldsymbol{U}_{nn'}] = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta_{n'}}$. For consistency, we must show that when $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_{n'}$ are at least $\frac{2\sqrt{K}C_N}{M^\alpha}$ apart, it implies that there is little chance the two subjects will be clustered together. Specifically,

$$P\left[d^V_{nn'} < \frac{2\sqrt{K}C_N}{M}\right] \to 0, \tag{A.6}$$

and for a stronger version we multiply the probability above by $N(N-1)$ to show it holds for all pairs simultaneously.

Observe that

$$\sqrt{K} \mid \theta_{nk^*} - \theta_{n'k^*} \mid > \frac{2\sqrt{K\log N}}{M^\alpha},$$

where $k^* = \max_k |\theta_{nk} - \theta_{n'k}|$. Without loss of generality, we assume that $\theta_{nk^*} > \theta_{n'k^*}$ and

$$\theta_{nk^*} - \theta_{n'k^*} > \frac{2\sqrt{\log N}}{M^\alpha}. \tag{A.7}$$

Note that $U_{nn'k^*} \sim N(\mu_{k^*}, \tilde{\psi}^2_{k^*})$, where

$$\mu_{k^*} = \frac{1}{\sqrt{J}}\sum_{j=1}^{J} q_{jk^*}(\theta_{nk^*} - \theta_{n'k^*})\beta_{jk^*} > \frac{2\sqrt{\log N}}{M^\alpha\sqrt{J}}\sum_{j=1}^{J} q_{jk^*}\beta_{jk^*},$$

and

$$\tilde{\psi}_{k^*} = \frac{2}{J}\sum_{j=1}^{J} q_{jk^*}\psi_j.$$

Then

$$\begin{aligned} P\left[d\boldsymbol{V}_{nn'} < \frac{2\sqrt{K}C_N}{M}\right] &< P\left[U_{nn'k^*} < \frac{2\sqrt{K}C_N}{M}\right] \\ &< P\left[Z < \frac{\frac{2\sqrt{K}C_N}{M} - \frac{2\sqrt{\log N}}{M^\alpha\sqrt{J}}\sum_{j=1}^{J} q_{jk^*}\beta_{jk^*}}{\frac{2}{J}\sum_{j=1}^{J} q_{jk^*}\psi_j}\right] \\ &\overset{J\to\infty}{\to} P\left[Z > \frac{\frac{2\sqrt{\log N}}{M^\alpha\sqrt{J}}\sum_{j=1}^{J} q_{jk^*}\beta_{jk^*}}{\frac{2}{J}\sum_{j=1}^{J} q_{jk^*}\psi_j}\right] \equiv P[Z > t], \end{aligned} \tag{A.8}$$

where

$$t = \frac{\frac{2\sqrt{\log N}}{M^\alpha\sqrt{J}}\sum_{j=1}^{J} q_{jk^*}\beta_{jk^*}}{\frac{2}{J}\sum_{j=1}^{J} q_{jk^*}\psi_j} \sim O\left(\frac{\sqrt{J\log N}}{M^\alpha}\right).$$

Then the probability that there is any pair with $\boldsymbol{\theta}$ values farther than $\frac{2\sqrt{K}C_N}{M^\alpha}$ that fall into the same cluster is bounded by

$$N(N-1)P\left[Z > t\right]. \tag{A.9}$$

By once again bounding the tail of a normal probability, we see that the probability there is any pair of subjects with $\boldsymbol{\theta}$ values differing by at least $\frac{2\sqrt{K}C_N}{M^\alpha}$ placed in the same cluster is of order $O\left(N^2\frac{M^\alpha}{\sqrt{J\log N}}\exp\left[-\frac{J\log N}{M^{2\alpha}}\right]\right)$. By the assumptions of the Theorem, this converges to 0, implying consistency in the latent variable space for any partition with a maximum diameter less than $\frac{2\sqrt{K}\log N}{M}$. Furthermore, by construction we know such paritions exist with probability converging to 1, so consistency must also hold for a sequence of MDP solutions.

$$\begin{aligned} P\left[d^2_{nn'} < \frac{4KC_N^2}{M^2}\right] &= P\left[d^2_{nn'} - \mathbf{E}d^2_{nn'} < \frac{4KC_N^2}{M^2} - \mathbf{E}d^2_{nn'}\right] \\ &\leq P\left[\mathbf{E}d^2_{nn'} - d^2_{nn'} > J\lambda_{min}\frac{2KC_N^2}{M^{2\alpha}} - \frac{4KC_N^2}{M^2}\right] \\ &\leq P\left[\mid d^2_{nn'} - \mathbf{E}d^2_{nn'} \mid > J\lambda_{min}\frac{2KC_N^2}{M^{2\alpha}} - \frac{4KC_N^2}{M^2}\right] \\ &\leq \frac{\mathrm{var}[\mathrm{d}^2_{\mathrm{nn'}}]}{(J\lambda_{min}\frac{2KC_N^2}{M^{2\alpha}} - \frac{4KC_N^2}{M^2})^2} \end{aligned} \tag{A.10}$$

The inequality in (A.10) is due to Chebyshev's inequality. Then we analysis the variance of $d^2_{nn'}$, Note

that $U_{nn'k} \sim N(\mu_k, \tilde{\sigma}_k^2)$, where

$$\mu_k = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} q_{jk}(\theta_{nk} - \theta_{n'k})\beta_{jk} \sim O(\sqrt{J}),$$

and

$$\tilde{\sigma}_k^2 = \frac{2}{J} \sum_{j=1}^{J} q_{jk}^2 \sigma_j^2 \sim O(1).$$

Note that $\sigma_j^2 = \mathrm{var}[\mathrm{e_{nj}}]$, where e_{nj} is defined in (2.9). So

$$\begin{aligned} \mathrm{var}[\mathrm{d}_{\mathrm{nn'}}^2] &= \mathrm{var}[\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{U}_{\mathrm{nn'k}}^2] = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{var}[\mathrm{U}_{\mathrm{nn'k}}^2] \\ &= \sum_{k=1}^{K} 2\tilde{\sigma}_k^4 + 4\mu_k^2\tilde{\sigma}_k^2 \sim O(J). \end{aligned}$$

$$P\left[d_{nn'}^W < \frac{2\sqrt{K}C_n}{M}\right] \leq P\left[\max_k |U_{nn'k}| < \frac{2\sqrt{K}C_n}{M}\right]$$

the maximum distance of any two points is less than the diagonal

$$\begin{aligned} & P\left[\mid \hat{P}_n - \hat{P}_{n'} \mid < \frac{1}{M}\right] \\ = & \frac{n(n-1)}{2} P\left[\mid \hat{P}_1 - \hat{P}_2 \mid < \frac{1}{M}\right] \\ = & \frac{n(n-1)}{2} P\left[\mid (\hat{P}_1 - P_1) - (P_1 - P_2) + (P_2 - \hat{P}_2) \mid < \frac{1}{M}\right] \\ < & n(n-1)P\left[(\hat{P}_1 - P_1) + (P_2 - \hat{P}_2) < -\frac{\beta_1^*}{M^\alpha}\right] \\ \leq & 2n(n-1)P\left[\hat{P}_1 - P_1 > \frac{\beta_1^*}{2M^\alpha}\right] \\ \leq & 2n(n-1)e^{\frac{2\beta_1^* J}{M}} \to 0. \end{aligned} \tag{A.11}$$

The last inequality in (A.11) is obtained by Hoeffding's Inequality. Based on (**??**) and (A.11), we see that the probability of two subjects whose latent traits are far apart to be clustered into one cluster goes to zero provided $N, J, M \to \infty$, $\frac{J}{M} \to \infty$, $\frac{N}{M} \to \infty$ and $N^2 e^{-\frac{2J}{M}} \to 0$. Therefore, the unidimensional distance measure in (2.2) yields consistent clustering result.

References

[1] Adenstedt, R.K. (1974). On Large Sample Estimation for the Mean of A Stationary Random Sequence, *Annals of Statistics, 2*, 1095-1107.

[2] Beran, J. (1994). *Statistical Methods for Long Memory Processes*, London: Chapman and Hall.

[3] Bickel, P.J., and Sakov, A. (2008). On the Choice of M in the M Out of N Bootstrap and Confident Bounds for Extrema, *Statistica Sinica, 18*, 967-985.

[4] Bingham, N.H., Goldie, C.M., and Teugels, J.L. (1987). *Regular Variation*, Cambridge: Cambridge University Press.

[5] Conti, P.L., Giovanni, L.D., Stoev, S.A., and Taqqu, M.S. (2008). Confidence Intervals for the Long Memory Parameter Based on Wavelets and Resampling, *Statistica Sinica, 18*, 559-579.

[6] Davydov, Y.A. (1970). The invariance principle for stationary processes, *Theory of Probability and Its Applications, 15*, 487-498.

[7] Dobrushin, R.L., Major, P. (1979). Non-Central Limit Theorems for Non-Linear Functionals of Gaussian Fields, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 50*, 827-842.

[8] Götze, F. Racǩauskas, A. (2001). Adaptive Choice of Bootstrap Sample Sizes, *Lecture Notes-Monograph Series, 36*, 286-309, State of the Art in Probability and Statistics.

[9] Granger, C.W.J. and Joyeux, R. (1980). An Introduction to Long Memory Time Series Models and Fractional Differencing, *Journal of Time Series Analysis, 1*, 15-39.

[10] Hall, P., Jing, B.Y., and Lahiri, S.N. (1998). On the Sampling Window Method for Long-Range Dependent Data, *Statistica Sinica, 8*, 1189-1204.

[11] Hosking, J.R.M. (1981). Fractional Differencing, *Biometrika, 68*, 165-176.

[12] Ibragimov, I.A., and Rozanov, Y.A. (1978). *Gaussian Random Processes*, New York: Springer.

[13] Kiefer, N.M., and Vogelsang, T.J. (2005). A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests, *Econometric Theory, 21*, 1130-1164.

[14] Künsch, H.R. (1989). The Jackknife and Bootstrap for General Stationary Observations, *Annals of Statistics, 17*, 1217-1261.

[15] Lahiri, S.N. (1993). On the Moving Block Bootstrap Under Long Range Dependence, *Statistics and Probability Letters, 18*, 405-413.

[16] Liu, R.Y., and Singh, K. (1992). Moving Blocks Jackknife and Bootstrap Capture Weak Dependence, *In R. LePage and L. Billard (eds.), Exploring the Limits of Bootstrap*, New York: John Wiley.

[17] Lobato, I.N. (2001). Testing That a Dependent Process is Uncorrelated, *Journal of the American Statistical Association, 96*, 1066-1076.

[18] Mandelbrot, B.B., and Van Ness, J.W. (1968). Fractional Brownian Motions, Fractional Noises and Applications, *SIAM Review, 10*, 422-437.

[19] McElroy, T., and Politis, D.N. (2009). Fixed-b asymptotics for the studentized mean from time series with short, long or negative memory, *Technical Report, Department of Economics*, University of California, San Diego.

[20] Nordman, D., and Lahiri, S.N. (2005). Validity of the Sampling Window Method for Long-Range Dependent Linear Processes, *Econometric Theory, 21*, 1087-1111.

[21] Nordman, D., Sibbertsen, P., and Lahiri, S.N. (2007). Empirical Likelihood for the Mean Under Long-Range Dependence, *Journal of Time Series Analysis, 28*, 576-599.

[22] Politis, D.N., and Romano, J.P. (1994). Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, *Annals of Statistics, 22*, 2031-2050.

[23] Politis, D.N., Romano, J.P., and Wolf, M. (1999). *Subsampling*, New York: Springer-Verlag.

[24] Shao, X. (2009). Confidence Intervals for Spectral Mean and Ratio Statistics, *Biometrika, 96*, 107-117.

[25] Shao, X. (2010). A Self-Normalized Approach to Confidence Interval Construction in Time Series, To appear *Journal of the Royal Statistical Society, Series, B.*

[26] Taqqu, M.S. (1975). Weak Convergence to Fractional Brownian Motion and to the Rosenblatt Process, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 31*, 287-302.

[27] Taqqu, M.S. (1979). Convergence of Integrated Processes of Arbitrary Hermite Rank, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 50*, 53-83.

[28] Wu. W., and Shao, X. (2006). Invariance principles for fractionally integrated nonlinear processes, *IMS Lecture notes-Monographs series, Festschrift for Michael Woodroofe, 50*, 20-30.

[29] Brusco, M.J., & Stahl, S. (2005). Branch-and-Bound Applications in Combinatorial Data Analysis, Springer, New York, NY.

[30] Chiu, C.Y., Douglas, J. & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications, *Psychometrika, 74*, 633-665.

[31] de la Torre, J. & Douglas, J.A. (2004). Higher order latent trait models for cognitive diagnosis, *Psychometrika, 69*, 333-353.

[32] Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

[33] Hands, S., & Everitt, B. S. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techiniques, *Multivariate Behavioural Research, 22*, 235-243.

[34] Hansen, P., & Delattre, M. (1978). Complete-Link cluster analysis by graph coloring. *Journal of the American Statistical Association, 73*, 397-403.

[35] Hartz, S., Roussos, L., Henson, R. & Templin, J. (2005). The Fusion Model for skill diagnosis: Blending theory with practicality. Unpublished manuscript.

[36] Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210.

[37] Hoeffding, W. (1963). Probabilistic inequalities for sums of bounded random variables. *Annals of Mathematical Statistics, 58*, 13-30.

[38] Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological measurement, 25*, 258-272.

[39] Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379-416.

[40] Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

[41] Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology, 15*. 201-293.

[42] Tatsuoka, C. (2002). Data-analytic methods for latent partially orderedclassification models. *Applied Statistics* (JRSS-C), 51, 337-350.

[43] Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

[44] Templin, J.L., & Henson, R.A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.

[45] Brusco, M.J., & Stahl, S. (2005). Branch-and-Bound Applications in Combinatorial Data Analysis, Springer, New York, NY.

[46] Chiu, C.Y., Douglas, J. & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications, *Psychometrika, 74*, 633-665.

[47] de la Torre, J. & Douglas, J.A. (2004). Higher order latent trait models for cognitive diagnosis, *Psychometrika, 69*, 333-353.

[48] Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

[49] Hands, S., & Everitt, B. S. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techiniques, *Multivariate Behavioural Research, 22*, 235-243.

[50] Hansen, P., & Delattre, M. (1978). Complete-Link cluster analysis by graph coloring. *Journal of the American Statistical Association, 73*, 397-403.

[51] Hartz, S., Roussos, L., Henson, R. & Templin, J. (2005). The Fusion Model for skill diagnosis: Blending theory with practicality. Unpublished manuscript.

[52] Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210.

[53] Hoeffding, W. (1963). Probabilistic inequalities for sums of bounded random variables. *Annals of Mathematical Statistics, 58*, 13-30.

[54] Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological measurement, 25*, 258-272.

[55] Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379-416.

[56] Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

[57] Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danish Institute for Educational Research.

[58] Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology, 15*. 201-293.

[59] Tatsuoka, C. (2002). Data-analytic methods for latent partially orderedclassification models. *Applied Statistics* (JRSS-C), 51, 337-350.

[60] Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

[61] Templin, J.L., & Henson, R.A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.

[62] Chang, H.H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213–229.

[63] Cox, D.R. (1975) Partial likelihood. *Biometrika*, 62, 269-276.

[64] Klein, J.P. & Moeschberger, M.L. (1997). Survival Analysis. Springer, New York, NY.

[65] Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press.

[66] Roskam, E.E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam and R. Suck (Eds.), *Progress in mathematical psychology*, 151-171.

[67] Rouder, J.N., Sun, D., Speckman, P.L., Lu,J., & Zhou,D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68, 589-606.

[68] Verhelst, N.D., Verstraalen, H.H.F.M., & Jansen, M.G. (1997). A logistic model for time-limit tests. In W.J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*, Springer, New York, NY.

[69] van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika,72*, 287–308.

[70] van der Linden, W.J. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*, 120–139.

[71] Brusco, M.J., & Stahl, S. (2005). Branch-and-Bound Applications in Combinatorial Data Analysis, Springer, New York, NY.

[72] Chiu, C.Y., Douglas, J. & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications, *Psychometrika, 74*, 633-665.

[73] de la Torre, J. & Douglas, J.A. (2004). Higher order latent trait models for cognitive diagnosis, *Psychometrika, 69*, 333-353.

[74] Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

[75] Hands, S., & Everitt, B. S. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techiniques, *Multivariate Behavioural Research, 22*, 235-243.

[76] Hansen, P., & Delattre, M. (1978). Complete-Link cluster analysis by graph coloring. *Journal of the American Statistical Association, 73*, 397-403.

[77] Hartz, S., Roussos, L., Henson, R. & Templin, J. (2005). The Fusion Model for skill diagnosis: Blending theory with practicality. Unpublished manuscript.

[78] Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210.

[79] Hoeffding, W. (1963). Probabilistic inequalities for sums of bounded random variables. *Annals of Mathematical Statistics, 58*, 13-30.

[80] Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological measurement, 25*, 258-272.

[81] Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379-416.

[82] Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

[83] Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests, Copenhagen: Danish Institute for Educational Research.

[84] Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology, 15*. 201-293.

[85] Tatsuoka, C. (2002). Data-analytic methods for latent partially orderedclassification models. *Applied Statistics* (JRSS-C), 51, 337-350.

[86] Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

[87] Templin, J.L., & Henson, R.A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

Vita

Zhewen Fan graduated from University of Science and Technology of China in 2004 with a Bachelor of Science degree in Electrical Engineering. She completed a Master of Science degree in Statistics from the University of Illinois at Urbana-Champaign in 2007. She is expected to complete her Ph.D., in Statistics from the University of Illinois at Urbana-Champaign in 2010.