THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Multivariate Outlier Detection in Latent Variable Models

Yan Lu

Department of Statistics
London School of Economics and Political Sciences

A thesis submitted to the Department of Statistics of
the London School of Economics and Political Science
for the degree of Doctor of Philosophy

March 2022

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of 36,202 words.

Statement of conjoint work

I confirm that Chapter 3 was adapted into a paper entitled "Detection of Two-Way Outliers in Multivariate Data and Application to Cheating Detection in Educational Tests" jointly co-authored with my supervisors, Dr Yunxiao Chen and Professor Irini Moustaki, and published in Annals of Applied Statistics (Chen, Lu, & Moustaki, 2022).

Abstract

Outliers often pose serious problems for statistical models since they can distort the model fit and bias parameter estimation. Outliers are also worthy of attention in their own rights, as they are often informative of substructures of the data. This thesis aims to develop methods of detecting multivariate outliers in latent variable modelling contexts. Outliers are defined as data subsets deviating from a baseline model specified for the majority of the data. By this definition, we specify one-way outliers on the basis of atypical attributes of either individuals or variables and two-way outliers on the basis of atypical attributes of both individuals and variables.

In this thesis, we develop the Forward Search (FS) procedures for detecting outlying individuals, latent groups of individuals and DIF variables. The FS does not examine just one subset of the data but instead fits a sequence of augmented subsets in order to decide which part of the data deviates from the baseline model. Outliers are identified through monitoring the effect of the sequential addition of individuals or items on the fitted model. The performance of the FS is assessed through simulated data and cross-national survey data under latent class models, factor mixture models and multiple-group latent variable models.

To detect two-way outliers, the thesis proposes to impose a latent class model component for capturing two-way outliers upon a latent factor model component for capturing normal item response behaviour. Statistical inference is carried out under a fully Bayesian framework. The detection of two-way outliers is formulated based on the proposed Bayesian decision rules and compound decision rules that control local false discovery rate and local false non-discovery rate. The proposed method proves to be particularly useful in simultaneously detecting compromised items and test takers with item pre-knowledge in educational tests. To further improve two-way outlier detection, the two-way outlier detection model is extended in an explanatory framework by accounting for covariate effects and the relationships between latent variables.

Acknowledgements

This thesis was completed at the end of the COVID-19 pandemic. I would like to sincerely thank my supervisors, Professor Irini Moustaki and Dr Yunxiao Chen, for their patience, encouragement and unwavering support during my studies and particularly during the COVID-19 pandemic. The completion of this thesis would not have been possible without their continued academic guidance. Their critical engagement and generosity with time played a significant role in shaping my research. I am deeply grateful to them for the discussions that helped me sort out major methodological problems and technical issues.

I would like to thank Professor Irini Moustaki, Dr Yunxiao Chen, Dr James Abdey, and Professor Fiona Steele for offering me teaching opportunities. I would also like to thank Dr Anastasia Kakou for her encouragement and advice on teaching. I am thankful to Daniel Linehan at the LSE LIFE for offering me the position of Quantitative Methods Adviser in the summer months. It was an eye-opening experience, and I benefited hugely from communicating statistics and data to a broader audience.

I owe a debt of gratitude to my fellow PhD students in Columbia House 502 for their comradeship and critical thought. My thanks also go to the Department of Statistics for the financial support and Penny Montague for her constant support during the last four years. Finally, I am grateful to my parents for their unconditional love and endless encouragement. To Martin, Sydney, Xinyu, Jiarong, and Mariam, thank you for always being there and being good company.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Outliers are isolated or clustered observations which differ from the main bulk of the data (Rousseeuw & Hubert, 2018). Outliers often pose serious problems for statistical models, as they may cause departures from model assumptions, distort the model fit and often lead to biased parameter estimates. However, outliers are not always seen as problematic data that need to be excluded from model estimation procedures. Instead, they can be worthy of attention in their own rights, especially when they contain information about substructures of the data. Outliers in univariate data are clearly defined as extreme values and can be easily visualised. In contrast, multivariate outliers are not necessarily defined by extreme values along a single variable. For many multivariate datasets, especially the ones containing categorical variables, it is difficult to reveal their real structure using two-dimensional or three-dimensional visualisations.

Outliers in this thesis are defined with respect to the postulated model. This definition specifies one-way outliers as individuals or variables that deviate from a specific baseline model and specifies two-way outliers as item responses consisting of both individuals and variables deviating from a specific baseline model. In illustrating one-way and two-way outliers, we give the following examples.

1.1 Examples of One-way and Two-way Outliers

von Davier and Lee (2019) mentioned a type of one-way outliers characterised by individuals whose response patterns result from cheating, guessing, random responding and all other aberrant response behaviour in educational tests. Those who do not conform to the standard response behaviour are likely to give unexpected correct responses to particularly difficult questions and wrong answers to very basic questions. Aberrant response patterns as such threaten the validity of test scores and therefore should be identified and accounted for by the model.

One-way outliers can also be defined as atypical items, for example, items exhibiting Differential Item Functioning (DIF; Holland & Wainer, 1993; Millsap, 2012). Items exhibiting DIF do not measure the same construct across groups of individuals. Given a baseline model assuming measurement invariance (i.e. observed variables in the data measure the same construct across groups), the items showing DIF would be poorly fitted. One example of DIF variables (or DIF items) which are extensively studied in the thesis is compromised items in educational tests. In many computer-based testing programs, although new test items are developed for each administration of the test, many test items have been used in previous tests. The reused items are likely to get exposure prior to the time of administration, and examinees with prior access to them will benefit from high scores on these items. Compromised items have a harmful impact on test validity because they measure person ability differently between those who have item preknowledge and those who do not.

Since outliers are defined with respect to a hypothesised model, model misspecifications could result in outliers and affect the way outliers are identified. The heterogeneity among individuals and/or variables could lead to violation of local independence, which is one of the underlying assumptions of latent variable models. In the presence of population heterogeneity, individuals not captured by a baseline model assuming population homogeneity or fewer homogeneous subpopulations can

be seen as outliers. If one or more variables no longer measure the latent construct in the same way across different subpopulations, the assumption of measurement invariance may not hold as well. This shows that aberrations on both the individual and variable levels, if not adequately addressed, can simultaneously result in departures from the hypothesised model. Therefore, the detection of outliers is essentially a classification problem for individuals or/and items in the data not dominated by outliers.

In following the definition of one-way outliers, two-way outliers are defined as the data subsets consisting of outlying individuals' responses to atypical variables. One prominent example of two-way outliers arises from aberrant response behaviour due to item preknowledge in educational tests. Individuals with preknowledge of leaked or compromised items are expected to benefit from score inflation, but their inflated scores on the compromised items cannot fairly reflect their ability. The responses from persons with prior access to compromised items are not compatible with a baseline model specified for honest response behaviour and are thus considered as outliers. When the data that are used for parameter estimation contain the test takers who respond to the compromised items based on their true ability and preknowledge, estimates for item difficulties and person abilities will be distorted.

It is also worth mentioning that like univariate data, multivariate data in the presence of multiple outliers are also subject to masking and swamping effects. The intuitive definitions of these two effects have been given by Acuna and Rodriguez (2004); Barnett and Lewis (1984); Davies and Gather (1993); Iglewicz and Martinez (1982). The masking effect occurs when some outliers strongly affect the fitted model, so much so that other outliers remain undetected. The swamping effect occurs when some "good" observations are detected as outliers only because of deviations from the model caused by the existence of other outliers. Both issues have been extensively studied in univariate contexts, particularly in linear models (Hadi & Simonoff, 1993; Lawrance, 1995). Any methods designed to detect outliers in multivariate data are expected to address these issues as well.

1.2 The Purpose of the Thesis

This thesis aims to develop model-based approaches for detecting outliers in multivariate data. We base our study on different specifications of latent variable models which have been widely used for modelling the associations in multivariate data with latent variables. The types of latent variables can be categorical, continuous or hybrid, leading to latent class models (Clogg, 1995; McLachlan & Peel, 2004), factor models with continuous outcomes (Reise, Widaman, & Pugh, 1993; Thompson, 2004) and with categorical outcomes, also known as Item Response Theory (IRT) models (Embretson & Reise, 2013), and factor mixture models (Lubke & Muthén, 2005) which can be viewed as an extension to the classical factor models in the presence of population heterogeneity.

We propose to develop methods of detecting one-way and two-way outliers in multivariate data within the latent variable modelling framework, without relying on information as to which individuals or/and items are truly outliers. We adapt the Forward Search (FS; Atkinson, 1994) for detecting one-way outliers, including outlying response patterns, latent groups of individuals and DIF items under latent variable models through defining and monitoring appropriate context-specific diagnostic statistics. Moreover, we propose two-way outlier detection models for simultaneously detecting two-way outliers deviating from a baseline model due to latent DIF. The two-way outlier detection model based on item responses is composed of a latent factor model component as the baseline model for standard item response behaviour and a latent class model component for capturing the effect of two-way outliers due to atypical response behaviour. We further formulate the detections of outlying individuals and items under a statistical decision framework and propose Bayesian decision rules and compound decision rules that control the local false discovery rate or local false non-discovery rate. The proposed models and compound decision rules are applied to high-stake educational test data to simultaneously detect test takers with preknowledge and compromised items.

1.3 Overview of Chapters

The thesis is organised as follows. In Chapter 2, we first review the FS and its advantages over conventional backward methods based on deletion diagnostics. We then adapt the FS for detecting one-way outliers that cause deviations from a baseline latent variable model and its assumptions. Specifically, we develop the FS algorithm for detecting outlying response patterns, latent groups of individuals and DIF items in multiple-group data within the latent variable modelling framework.

The models for detecting two-way outliers on the basis of item responses only (i.e. the reduced model) and both item responses and response times (i.e. the full model) are proposed in Chapter 3. Statistical inference is carried out under a fully Bayesian framework. The detections of outlying individuals and DIF items are assessed based on the proposed Bayesian decision rules and compound decision rules.

In Chapter 4, the reduced model is extended in an explanatory framework. The resultant explanatory model incorporates covariate information into the structural model as exogenous variables while relaxing the previous model assumption about the independence between latent indicators of outliers and continuous latent variables. The reduced, full and explanatory models are applied to real data gathered from a single administration of a computer-based non-adaptive licensure assessment. The classification and detection under the two-way outlier detection models within and without an explanatory framework are compared. Simulation studies are carried out to address the impacts of sample sizes, numbers of items, and model misspecifications on the performance of two-way outlier detection.

Finally, findings and directions for future research are summarised in Chapter 5. In addition, details on derivations and algorithms not covered by the main text are presented in Appendices.

Chapter 2

The Forward Search for One-way Outlier Detection

2.1 Introduction

After defining and showing examples of outliers in multivariate data, this chapter focuses in detail on the detection of one-way outliers in multivariate data using the Forward Search (FS). Prior to the proposal of the FS, one-way outliers were detected using a backward search based on deletion diagnostics derived for assessing the effect of individual observations on parameter estimation and model fit. However, the idea of monitoring the deletion diagnostics backwards raises several issues, most notably, the masking effect. The FS, on the other hand, can overcome these issues and therefore has been widely used to detect outliers and more generally, uncover hidden structures in the data. In this chapter, we seek to build on the current literature by extending the FS to the detection of outlying individuals and variables in multivariate data within a latent variable modelling framework.

This chapter is organised as follows. In Section 2.1, we review the backward search and deletion diagnostics. This is followed by a review of the FS, which includes its advantages over the backward deletion, the use of diagnostics in the FS, and key phases in the development of the FS. The latent variable modelling framework is in-

troduced in Section 2.2. Under the modelling framework, we specify the latent class model, factor mixture model, and their multiple-group representations. Section 2.3 describes the FS procedures for detecting outlying response patterns, latent groups of individuals and variables exhibiting DIF. Diagnostics which appraise the effect of the sequential addition of individuals or variables on the fitted model are developed. Simulation examples are used to demonstrate the FS procedures for detecting one-way outliers under latent class models, factor mixture models and multiple-group latent class and factor models. This is followed by a real data example in Section 2.4, where the Round 7 European Social Survey dataset (European Social Survey, 2014) is used to demonstrate the detection of DIF items under multiple-group latent class and multiple-group IRT models. This chapter concludes in Section 2.5 by discussing findings and possible future developments of the FS in outlier detection and beyond.

2.1.1 Review of Backward Search and Deletion Diagnostics

Starting with the whole dataset, the backward procedure searches backwards through the data and alternates between measuring outlyingness and removing observations until a certain number of potential outliers are excluded from the data. The measure of outlyingness typically includes problem-specific deletion diagnostics which assess the influence of omitting individual observations on model fit and parameter estimation. An observation is considered to be influential if its removal makes a marked difference in the fit of the model and parameter estimates according to the deletion diagnostic statistic used.

Deletion diagnostics have been widely used within and beyond the context of regression analysis (Atkinson & Riani, 2000). Examples of deletion diagnostics in regression analysis include deletion residuals and Cook's Distance (Cook, 1977). Observations for which the deletion residual is larger than a critical value are detected as potential outliers. Cook's Distance provides an overall measure of changes in parameter estimates when an observation is omitted. A large value of Cook's

distance indicates the observation is influential. Outside the scope of regression analysis, residual-implied Mahalanobis distance case diagnostic has been used to detect response patterns that are inconsistent with a given model in factor analysis (Yuan, Fung, & Reise, 2004), and more generally, structural equation models (SEM; Yuan & Hayashi, 2010) for continuous data. Case-deletion diagnostics have also been used to examine categorical data under SEMs.

The deletion diagnostic statistics we have discussed are all aimed at individuals. Sawatzky et al. (2018) moved on to detect and remove variables lacking measurement invariance (also known as measurement equivalence; Meredith, 1993; Widaman & Reise, 1997) in a latent variable mixture modelling framework using a backward deletion approach. In the *exploratory stage* of their proposed approach, mixture IRT or latent factor models assuming different numbers of latent classes are compared in order to determine the number of latent subpopulations. In the case of population heterogeneity, the dataset goes on to the *detection stage*, where the variables which contribute to the heterogeneity – i.e. the variables lacking measurement invariance across the identified latent groups – are detected by item-level DIF analyses. After removing these variables from the data, the reduced dataset once again goes through the *exploratory stage*, and in the case of population heterogeneity, proceeds to the *detection stage*. Such an iterative process goes on by removing variables lacking measurement equivalence one by one until the remaining variables are well fitted by the measurement-equivalent model. The remaining variables are considered the most invariant.

The backward search is known to suffer from the masking effect. The masking effect occurs when the data contain multiple outliers which strongly affect the fitted model - so much so they may all together keep every single one of them from being identified. The backward search is, therefore, unable to detect these outliers all at once. To cope with masked outliers, one needs to first calculate a robust estimator, which is designed to be only slightly influenced by a few isolated outliers and less sensitive to clusters of outliers and then use the residuals resulting from the robustly fitted

model to find outliers. Another issue regarding the backward search is the swamping effect. Not all observations removed under this procedure are outliers. Some of them are removed only because the resultant deletion diagnostic statistics are distorted by the existence of other outliers. The actual status of these observations remains to be further investigated. Furthermore, in the case of multiple deletions, one may want to check all possible combinations of observations to be removed, which could be computationally demanding. It also remains to be seen how many observations should be removed from the data.

2.1.2 Review of Forward Search

The Forward Search (FS; Atkinson, 1994) was proposed as a data-driven approach for detecting masked outliers through monitoring the effect of sequentially adding observations on the fit of a baseline model which is assumed to be compatible with the majority of the data.

The FS starts with a small and "outlier-free" subset and gradually moves toward the entire dataset. The initial subset is selected among a sufficiently large number of random subsets of smaller size, to each of which the baseline model is fitted. The subset that fits the baseline model best is chosen to initialise the search. This subset is known as the "basic" set, and its size increases as the FS proceeds by adding the observations closest to the established "basic" set. By moving forward from an "outlier-free" subset to the whole dataset, the FS manages to overcome the masking problem faced by the backward procedure in the presence of multiple outliers.

A sequence of subsets of increasing size established throughout the FS helps to decide which part of the data is compatible with the baseline model and which part is not. The effect of sequentially adding observations outside of the "basic" set on the baseline model can be assessed by monitoring problem-specific diagnostic statistics whenever a new observation is added to the "basic" set. Abrupt changes in these statistics indicate that observations poorly fitted by the baseline model are

present in the "basic" set. Since outliers are unlikely to join the "basic" set until the very late stage of the search, we can use the "basic" set obtained just before this stage to robustly estimate the baseline model. In doing this, the amount of data used to robustly estimate the baseline model is conditional on or driven by the data, which distinguishes the FS from the backward search and other robust methods.

The FS has since been extensively developed in the context of regression analysis. The deletion diagnostics that measure the effect of omitting an observation (Section 2.1.1) were adapted to suit the FS. For example, Atkinson and Riani (2000) proposed to monitor the effect of outliers on the parameter estimates during the FS via a modified Cook's distance statistic. Moreover, the minimum deletion residual among the observations outside of the "basic" set and the maximum studentised residual among those in the "basic" set can also be monitored. Atkinson and Riani (2006) proposed to generate simulation envelopes for the minimum deletion residual whose distribution is unavailable. The progression of the minimum deletion residual against the simulation envelopes was monitored. When the trajectory of the minimum deletion residual exceeds an envelope, the newly added observation and all the subsequently added observations are detected as outliers. The FS has been applied to many other univariate contexts as well, including time series (Riani, 2004) and spatial linear models (Cerioli & Riani, 1999).

Extensions have also been made to multivariate contexts, notably clustering (Atkinson & Riani, 2007), discriminant analysis (Riani & Atkinson, 2001), multidimensional scaling (Solaro & Pagani, 2007), factor analysis for continuous data (Mavridis & Moustaki, 2008), and factor analysis for binary data (Mavridis & Moustaki, 2009). The book by Atkinson, Riani, and Cerioli (2013) summarises key phases in the development of the FS in multivariate contexts.

It is worth noting that the FS is not limited to outlier detection but can be regarded as a general diagnostic approach for uncovering the cluster structure of multivariate data. For example, Atkinson and Riani (2007) applied the FS to find clusters in multivariate normal data through monitoring minimum Mahalanobis distances com-

puted based on a sequence of subsets of increasing size. The authors also elaborated on the use of simulation envelopes for the robust Mahalanobis distance statistics.

When data are suspected to come from more than one population, implementing FS multiple times is found to be useful in identifying clusters in the data. Atkinson et al. (2013) empirically justified using multiple forward searches with random starts rather than a single search starting from a carefully selected, "outlier-free" initial subset. The rationale for using random starts is that all searches eventually converge regardless of where they start once the presence of clusters is revealed. This strategy is also applicable to the situation where the sources of population heterogeneity are unknown a priori.

This brings us to the next point about simultaneously running multiple searches with random starts to detect latent population heterogeneity in the presence of outliers. The idea consists in finding a homogeneous group of observations using a single search, removing individuals belonging to the tentatively identified latent group, and repeating the FS procedure on the remaining data until all groups are revealed. This strategy also involves reallocating the observations to the 'nearest' group. Latent class models, also known as finite mixture models (McCutcheon, 2002; McLachlan & Peel, 2004), are often used to characterise population heterogeneity and provide a model-based framework for classifying individuals. However, in the presence of a heterogeneous population due to different unknown sources as well as multiple isolated or clustered outliers, even conventional robust methods might fail to reveal latent classes and outliers. The FS, on the other hand, is helpful in this situation since it is able to identify latent groups of individuals while detecting outliers.

Although the FS has been mainly used to detect individuals or groups of individuals, it can be applied to detect variables based on a set of attributes, particularly those exhibiting DIF across groups of individuals. By the logic of the FS, a sequence of subsets of incremental item size can be established and fitted by a baseline model assuming measurement equivalence. The change in model fit or item fit is monitored

whenever an item is included in the subset.

The purpose of this chapter is therefore to extend the FS to detect outlying individuals and DIF variables within a latent variable modelling framework. More specifically, an FS Algorithm is developed to address the detection of first outlying response patterns under latent class models and mixture IRT models, second latent dimensionality in terms of the number of distinct sub-populations under factor mixture models, and third items lacking measurement equivalence under multiple-group latent class and multiple-group IRT models.

2.2 Modelling Framework

In this section, we introduce a latent variable mixture modelling framework that can be used to specify all baseline models appearing in the FS applications later. Latent variable models (LVM) such as factor analysis models (with continuous latent variables), latent class models (with categorical latent variables) and factor mixture models (with both types of latent variables) have been used to handle multivariate data. LVMs are particularly widely used in social sciences for measuring unobserved constructs of interest such as ability, attitude, happiness, and state of health through a set of observed indicators, also known as manifest variables or items. LVMs are also used for data reduction and classification of individuals into unobserved homogeneous groups.

LVMs postulate certain assumptions such as (i) conditional independence of the items given the latent variables, (ii) normality for the continuous latent variables, and (iii) measurement equivalence of items across multiple groups defined by demographic, socio-economic or cultural categories (multiple group analysis). The second and third assumptions may need to be relaxed to fit the data in the presence of population heterogeneity.

Under LVMs, the number of latent variables or latent classes that are required to adequately explain the associations among measured items, and the detection

of non-equivalence of measured items across observable groups can be determined based on full-information goodness-of-fit tests (e.g., Pearson chi-squared difference tests, likelihood ratio tests etc.), limited-information goodness-of-fit tests based on lower-order marginal contingency tables (Bartholomew & Leung, 2002; Maydeu-Olivares & Joe, 2005), or model selection criteria such as the Akaike's information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). (or information criteria are performed on a sequence of nested LVMs with an increasing number of latent factors or latent classes or increasingly restrictive across-group equality constraints on item parameters.

2.2.1 A general modelling framework for LVMs

We first define a general modelling framework for LVMs. Consider a dataset consisting of N individuals' responses to J measured items. Let η and ξ denote a single continuous (classical factor analysis/latent trait model) and categorical (latent class model) latent variable respectively for a single respondent i, for $i = 1, \ldots, N$. For simplicity of presentation, we focus on univariate latent variables, η_i and ξ_i, and omit the person subscript i in the following sections. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_J)$ denote J binary or continuous items that are regarded as measures of the latent variable, η or ξ, depending on the type of latent variable used to summarise the associations in the data. $\boldsymbol{y} = (y_1, \ldots, y_J)$ are the realised values of $\boldsymbol{Y}$. For the multiple-group representation, we assume that each respondent belongs to one of G observed groups: $g = 1, \ldots, G$. The joint distribution of $\boldsymbol{Y}$ and η within group g is given by

$$p_g(\boldsymbol{y}, \eta) = p_g(\boldsymbol{y} \mid \eta; \boldsymbol{\Lambda}^{(g)}) p_g(\eta; \boldsymbol{B}^{(g)}). \tag{2.2.1}$$

The subscript g denotes a distribution depends on a observed group $g = 1, \ldots, G$, and parameters with the superscript (g) are group-specific. $p_g(\cdot)$ denotes a conditional or marginal probability density function. Specifically, $p_g(\boldsymbol{y} \mid \eta; \boldsymbol{\Lambda}^{(g)})$ is the *measurement model* that describes how the items measure the continuous latent

variable, η, in group g. Parameters associated with the *measurement model* within group g are represented by $\boldsymbol{\Lambda}^{(g)}$. $p_g(\eta; \boldsymbol{B}^{(g)})$ is the *structural model* that specifies the distribution of the latent variable in group g and the corresponding parameters $\boldsymbol{B}^{(g)}$.

Under the assumption that responses to the J items are conditionally independent after accounting for the latent variable η, the *measurement model* $p_g(\boldsymbol{y} \mid \eta; \boldsymbol{\Lambda}^{(g)})$ can be further represented by

$$p_g(\boldsymbol{y} \mid \eta; \boldsymbol{\Lambda}^{(g)}) = \prod_{j=1}^{J} p_g(y_j \mid \eta; \boldsymbol{\Lambda}_j^{(g)}). \tag{2.2.2}$$

It remains to specify the univariate measurement model for each of the J items, $p_g(y_j \mid \eta; \boldsymbol{\Lambda}_j^{(g)})$, for $j = 1, \ldots, J$. This conditional independence assumption is held for a latent class model with a categorical latent variable ξ.

The equations above provide a multiple-group representation of an LVM in the presence of observed population heterogeneity. In single-group analysis ($G = 1$), the group subscript g is dropped (and we also drop it in the rest of the chapter depending on the application we discuss).

When a comparison is made across multiple groups of individuals, the assumption that the items on which the comparison is based maintain measurement invariance across groups of individuals is required so that the latent variable is comparable on the same measurement scale across the groups. Any dependence of the *measurement model* on the observable grouping variable after controlling for the latent factor or conditioning on latent classes suggests a lack of measurement invariance; that is, parameter values pertaining to one or more items vary across the groups (Mellenbergh, 1989; Meredith, 1993; Widaman & Reise, 1997).

For the individual items, measurement invariance can be approached by using differential item functioning (DIF; Thissen, Steinberg, & Wainer, 1993) analysis. The DIF approach can be manifest or latent, depending on whether the population heterogeneity is observed or unobserved. The studies in this chapter involve manifest

DIF, and latent DIF as a source of two-way outliers is the focus of the next two chapters. The effect of DIF can be either uniform or non-uniform. Uniform DIF is analogous to scalar non-invariance and non-uniform DIF is analogous to metric non-invariance in factor models. A detailed explanation is provided in Sections 2.2.2 and 2.2.3.

When the sources of population heterogeneity and hence group memberships are unobserved, the question of interest becomes whether the *measurement model* depends on latent class memberships of individuals. A Factor mixture model which integrates latent classes indicated by the categorical latent variable, ξ, with common factor models can be used to accommodate unobserved population heterogeneity. The normality of the continuous latent variable, η, is no longer necessary in this case. Instead, a multi-modal distribution can be assumed for η to reflect latent population heterogeneity.

The rest of this section focuses on three special cases of the general modelling framework given by Equation (2.2.1), namely the latent class model for binary items with single- and multiple-group analysis, the multiple-group factor analysis model for binary items, and the factor mixture model for continuous items. In the following sections, those models will be used as baseline models and their associated goodness-of-fit measures will be monitored to indicate the effect of one-way outliers during the progression of the FS.

2.2.2 Latent Class Model for Single- and Multiple-groups

Latent class analysis (LCA) aims to explain the associations among J items using K latent classes indicated by a discrete latent variable ξ, where K is much smaller than J (Clogg, 1995; McCutcheon, 1987). LCA is typically useful for clustering and classifying multivariate data.

The latent class model for the J items in group g, for $g = 1, \ldots, G$, under the

conditional independence assumption is specified as

$$p_g(\boldsymbol{y}) = \sum_{k=1}^{K} c_k^{(g)} \prod_{j=1}^{J} p_g(y_j \mid \xi = k), \tag{2.2.3}$$

where $c_k^{(g)} = P_g(\xi = k) = p_g(\xi; \boldsymbol{B}^{(g)})$, for $k = 1, \ldots, K$, denotes the class probability of belonging to Class k for group g. The *structural model* is characterised by the structural parameter set, $\boldsymbol{B}^{(g)} = (c_1^{(g)}, \ldots, c_K^{(g)})$, subject to $c_k^{(g)} \geqslant 0$, $\sum_{k=1}^{K} c_k^{(g)} = 1$.

For each binary item, we assume a Bernoulli distribution conditional on latent class membership:

$$p_g(y_j \mid \xi = k) = (\pi_{jk}^{(g)})^{y_j}(1 - \pi_{jk}^{(g)})^{1-y_j}, \quad \text{for} \quad y_j = 0, 1 \ , \tag{2.2.4}$$

where $\pi_{jk}^{(g)} = P_g(Y_j = 1 \mid \xi = k)$ $(j = 1, \ldots, J, k = 1, \ldots, K)$ denotes the probability of a positive response to item j conditional on latent class k and group g. The *measurement model* is characterised by the item parameter matrix $\boldsymbol{\Lambda}^{(g)} = \{\pi_{jk}^{(g)}\}_{J \times K}$.

For continuous items, a normal distribution given latent class and observed group memberships can be assumed: $y_j \mid \xi = k \sim \mathrm{N}\left(\mu_{jk}^{(g)}, \left(\sigma_{jk}^{(g)}\right)^2\right)$, subject to identifiability constraints.

In the single group case, g is dropped.

In the case of multiple groups, *measurement equivalence* is established when class-conditional item response probabilities, $\{\pi_{jk}^{(g)}\}_{J \times K}$, are equal across the groups so that the latent classes have the same meaning for all groups and the measurements in different groups are comparable. *Complete measurement equivalence* further requires class probabilities to be equal across groups. The question of whether the class probabilities are identical across groups can be addressed via a likelihood-ratio difference test after the equivalence of class-conditional item response probabilities has been established. The data-generating models in Sections 2.3.4 and 2.4.2 assume *partial measurement* invariance (Collins & Lanza, 2010), meaning that some but not all conditional item response probabilities to be equal across the observed groups.

The multiple-group analysis allows parameters pertaining to non-equivalent items to vary across the groups while setting parameters related to equivalent items equal.

To understand the ways in which measurement invariance could be untenable, consider one or more items functioning differently as indicative of the latent classes across different observed groups. In this case, measurement non-invariance is present due to observed population heterogeneity, which is the source of DIF for the items in question if we use the language of IRT. The difference in the expected responses to an item showing uniform DIF across the groups would be the same within every latent class, while the non-uniform DIF effect would be different for one or more of the latent classes. DIF becomes a problem when the source of DIF is also associated with the latent classes.

2.2.3 Factor Analysis Model

Factor analysis aims to explain the associations among items using continuous latent variables or latent traits. The factor model we present here assumes a single continuous latent variable denoted by η and is applied to data containing J items and G observed groups. Again, subscript $g = 1, \ldots, G$ is used to label the groups in the formulation. In the single group case, the subscript g is dropped. The model can be generalised to include multiple latent factors.

Under the conditional independence assumption, the single-factor model for group g is given by

$$p_g(\boldsymbol{y}) = \int \prod_{j=1}^{J} p_g(y_j \mid \eta^{(g)}; \boldsymbol{\Lambda}^{(g)}) p_g(\eta; \boldsymbol{B}^{(g)}) d\eta^{(g)}. \tag{2.2.5}$$

The general form of the *measurement model*, $p_g(y_j \mid \eta^{(g)}; \boldsymbol{\Lambda}^{(g)})$, can be represented by a generalised latent variable model (GLVM; Moustaki & Knott, 2000), in which a link function denoted by $h(\cdot)$ relates the conditional expectation of the responses, $\mu_j^{(g)} = \mathbb{E}\left[Y_j \mid \eta^{(g)}\right]$, to the latent factor $\eta^{(g)}$. The group-g-specific link for item j is given by

$$h\big(\mu_j^{(g)}\big) = \nu_j^{(g)} + \lambda_j^{(g)} \eta^{(g)}, \quad j = 1, \ldots, J; \quad g = 1, \ldots, G, \tag{2.2.6}$$

where group-specific item intercepts $\nu_j^{(g)}$ and item slopes (or factor loadings) $\lambda_j^{(g)}$, for $j = 1, \ldots, J$, constitute the item parameter set $\boldsymbol{\Lambda}^{(g)} = (\boldsymbol{\nu}^{(g)}, \boldsymbol{\lambda}^{(g)})$.

If item j is continuous, an identity link is used:

$$h\big(\mu_j^{(g)}\big) = \mu_j^{(g)} = \nu_j^{(g)} + \lambda_j^{(g)}\eta^{(g)}. \tag{2.2.7}$$

The *measurement model* for continuous data is hence a single-factor model. For binary items, the *measurement model* under the logit link is presented as

$$h(\mu_j^{(g)} \mid \eta^{(g)}) = \text{logit}(\mu_j^{(g)}) = \nu_j^{(g)} + \lambda_j^{(g)}\eta^{(g)}, \tag{2.2.8}$$

which is known as a (multiple-group) 2PL IRT model. $\nu_j^{(g)}$ and $\lambda_j^{(g)}$ are also known as group-specific item difficulty and item discrimination for item j, for $j = 1, \ldots, J$.

In the case of multiple groups where $g > 1$, lack of measurement invariance implies that the items may not measure the same latent construct across different groups of individuals. In IRT, lack of measurement invariance is known as DIF. Scalar non-invariance (uniform DIF items) exists when the intercept or item difficulty, $\nu_j^{(g)}$, differs across observed groups of individuals, while metric non-invariance (non-uniform DIF items) implies that factor loadings or discrimination parameters, $\lambda_j^{(g)}$'s, (and possibly item difficulty as well) vary across known groups. Simulation 4 in Section 2.3.5 and the case study in Section 2.4.3 aim to detect uniform DIF items only. An item showing uniform DIF tends to be systematically more difficult or easier for all individuals belonging to one observed group than individuals in other groups, even for those with the same level of the latent trait (e.g., ability).

A normal distribution is typically assumed for the latent factor in the absence of unobserved population heterogeneity and skewness: $\eta^{(g)} \sim \text{N}\Big(\mu_\eta^{(g)}, \sigma_\eta^{2,(g)}\Big)$. The parameter set for the *structural model* is hence comprised of distributional parameters for η, e.g., $\boldsymbol{B}^{(g)} = (\mu_\eta^{(g)}, \sigma_\eta^{2,(g)})$. For identifiability, η is assumed to follow a standard normal distribution in one of the G groups while its mean and variance in

the other groups remain to be estimated. Simulations and case studies later in this chapter assume *configural invariance*; that is, the latent factor structure remains the same across the groups. Therefore, we can drop subscript (g) when it comes to the distribution of η.

2.2.4 Factor Mixture Model

While the normality assumption for η is convenient for computational reasons, the latent trait distribution in the population might be skewed or multi-modal when the observed data come from more than one distinct population, and unlike multiple-group data, the sources of population heterogeneity are unknown a priori. For this reason, a more flexible distribution is needed for approximating the non-normal latent trait η and accounting for unobserved population heterogeneity.

An attempt in this direction is to propose a factor mixture model (Lubke & Muthén, 2005) in which a categorical latent variable indicating K latent classes, denoted as $\xi = 1, \ldots, K$, is used to characterise latent population heterogeneity. Under a (uni-dimensional) factor mixture model, the latent factor η is assumed to follow a normal mixture with K latent classes, and the associations among items in each class are explained by a class-specific normally distributed η. The *structure model* is given by

$$\eta \sim \sum_{k=1}^{K} c_k \mathrm{N}\Big(\mu_{\eta k}, \sigma^2_{\eta k}\Big), \tag{2.2.9}$$

where $\mu_{\eta k}$ and $\sigma^2_{\eta k}$, for $k = 1, \ldots, K$, are the class-wise mean and variance. The structural parameter set consists of class-specific distributional parameters for η and class probabilities and is denoted by $\boldsymbol{B} = (\mu_{\eta k}, \sigma^2_{\eta k}, c_k; k = 1, \ldots, K)$, where the probability of belonging to Class k is $c_k = P(\xi_i = k)$.

A factor mixture model for continuous data is thus defined by (2.2.9) and (2.2.6) with an identity link. Replacing the identity link with a logit link, we have a factor mixture model for binary data, also known as a mixture IRT model (Mislevy & Verhelst, 1990; Rost, 1990; Rost & von Davier, 1995). To achieve identifiability, the

overall mean and variance for the latent factor η can be set at 0 and 1, which is equivalent to $\sum_{k=1}^{K} c_k \mu_{\eta k} = 0$ and $\sum_{k=1}^{K} c_k(\mu_{\eta k}^2 + \sigma_{\eta k}^2) = 1$.

2.3 Forward Search Overview

As previously mentioned in Section 2.1.2, the FS (Atkinson, 1994; Atkinson et al., 2013) starts from an "outlier-free" subset which is well fitted by the baseline model, and proceeds by iteratively adding observations according to their closeness to the subset until all observations are included. During the search, model estimation and statistical inference are based on a sequence of data subsets of increasing size. Outliers deviating from the baseline model can be detected by monitoring the effect of sequential addition of observations on the fitted model. Atypical variables (e.g., variable showing DIF) can also be detected following a similar procedure, and relevant statistics for assessing the effect of sequentially adding variables on model fit or item fit can be monitored.

The key steps of the FS are summarised below.

Step I Choose an initial subset of size m or p from the data of size N or J depending on whether we search for outlying cases or variables. This is the "basic" set formed at the beginning of the search while the remaining $(N - m)$ cases or $(J - p)$ items constitute the "non-basic" set.

The "basic" and the "non-basic" sets are mutually exclusive throughout the search. The ideal situation is to find a way of selecting the initial subset so that it is free of outliers but this is not always feasible with large initial subsets.

Step II Proceed by including the least outlying observations or items from the "non-basic" set so that eventually the whole dataset is included in the "basic" set. Since we start with an initial subset of m cases or p items, there are at most $(N - m)$ or $(J - p)$ steps until all observations/items are included.

Step III Monitor quantities such as parameter estimates, residuals, item fit statistics and goodness-of-fit statistics, during the progress of the search (step II).

We present below four examples that illustrate the success of the FS in detecting outlying response patterns, items lacking measurement equivalence, and data dimensionality. For each case, we define the steps of the FS, in terms of the statistics we use for selecting the initial subset, progressing in the search, and monitoring the effect of the sequential addition of observations or items on the fitted model depending on the aim of the application.

2.3.1 Forward Search for Outlier Detection in Latent Class Analysis

The FS has been applied to detect outliers in continuous and (Mavridis & Moustaki, 2008) binary (Mavridis & Moustaki, 2009) data under latent variable models with continuous latent factors. We now extend the FS to detect outlying response patterns under a latent class model.

The single-group representation of the latent class model for data $\boldsymbol{Y}$ containing J binary items and N individuals is defined in Section 2.2.2, which is

$$p(\boldsymbol{Y}) = \sum_{k=1}^{K} c_k \prod_{j=1}^{J} (\pi_{jk}^{(g)})^{y_j} (1 - \pi_{jk}^{(g)})^{1-y_j}, \quad \text{for} \quad y_j = 0, 1 \; . \tag{2.3.1}$$

Model parameters include a measurement parameter set, $\boldsymbol{\Lambda} = \{\pi_{jk}\}_{J \times K}$, and a structural parameter set, $\boldsymbol{B} = (c_1, \ldots, c_K)$.

Data Generation

We generated the data containing $N = 250$ individuals and $J = 10$ items from a 2-class model (Equation 2.3.1 with $K = 2$). Table 2.3.1 presents the values of class probabilities and class-conditional positive response probabilities in the data-generating model. These values are chosen to mimic the item responses from a

test that was administered to primary school students to evaluate their arithmetic ability (Sijtsma & Molenaar, 2002). The conditional positive response probabilities are made lower in Class 1 than Class 2 for all items, meaning that individuals in Class 2, in general, are more capable than those in Class 1. Items 1-5 are made more difficult than Items 6-10 for both classes.

A total of 20 out of the $N = 250$ individuals were randomly selected and their responses were replaced by four outlying response patterns that are not expected to be generated from the 2-class model. As Table 2.3.2 shows, these 20 individuals with outlying response behaviour tend to correctly answer the 'difficult' items (Items 1-5) and get the 'easy' ones (Items 6-10) wrong.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Class 1 (40%)	0.16	0.02	0.00	0.04	0.21	0.36	0.19	0.22	0.08	0.30
Class 2 (60%)	0.51	0.37	0.24	0.28	0.43	0.62	0.55	0.65	0.60	0.70

Table 2.3.1: *Simulation 1: Class prevalences c_k's (shown in brackets) and conditional item response probabilities π_{jk}'s used in the data-generating model.*

Frequency	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
6	1	1	0	1	1	0	0	0	0	1
4	1	1	1	1	0	0	0	0	0	0
5	1	1	1	0	1	0	0	0	0	0
5	1	0	1	1	1	0	0	0	0	0

Table 2.3.2: *Simulation 1: Outlying response patterns contained by the simulated data.*

The simulation study aims to assess the performance of the FS in detecting outlying response patterns not generated by the 2-class model (i.e. the baseline model). We describe below in detail the three steps of the FS.

Step 1: Choosing the initial subset

The first step is to choose an "outlier-free" initial subset consisting of item responses from m individuals ($m < N$). For the sake of the stable estimation and convergence of latent variable models, m should not be too small. An initial subset usually contains 5% and 15% of the entire sample depending on the sample size and the complexity of the baseline model.

Let $\mathbf{h} = (h_1, \ldots, h_m)'$ be an m-dimensional vector of row indices in the data matrix $\boldsymbol{Y}$. $\mathbf{h}$ contains m indices selected from $1, \ldots, N$: $1 \leqslant h_1, \ldots, h_m \leqslant N$. $S_{\mathbf{h}}^m = (\mathbf{y}_{h_1}, \ldots, \mathbf{y}_{h_m})$ refers to one possible initial subset of size m, e.g., the first row of the subset $S_{\mathbf{h}}^m$ $\boldsymbol{y}_{h_1}$ is the h_1-th row of the data matrix $\boldsymbol{Y}$.

The total number of all possible initial subsets of size m is $\binom{N}{m}$, but it is practically infeasible to investigate all of them even for a small N. A feasible way is to randomly select a reasonable number of subsets to investigate: $S_{\mathbf{1}}^m, \ldots, S_{\mathbf{H}}^m$, for $H < \binom{N}{m}$. The one most compatible with the baseline model, for example, the one having an exact or an asymptotic p-value of a goodness-of-fit statistic greater than a certain threshold, is chosen to initialise the FS. We can also start by checking all possible initial subsets and stop once we find a subset that meets the selection criteria.

That brings us to discuss model adequacy measures that can be used for selecting the initial subset. The fit of a latent class model can be assessed using overall goodness-of-fit test statistics (e.g., chi-squared or likelihood ratio test statistic), but they are known to be sensitive to the sparseness of the contingency table associated with multivariate categorical data (e.g., Bartholomew & Leung, 2002; Reiser & VandenBerg, 1994). The sparseness is likely to distort the asymptotic p-values for these overall goodness-of-fit statistics and makes it unfit as a criterion for assessing model adequacy. Alternatives that have been proposed include resampling methods (van Kollenburg, Mulder, & Vermunt, 2015) and test statistics based on limited-information methods (Maydeu-Olivares & Joe, 2006; Reiser, 1996). Limited-information goodness-of-fit statistics based on lower order margins such as bivariate residuals (BVR) are less sensitive to sparsity (Bartholomew, Steele, & Moustaki, 2008; Jöreskog & Moustaki, 2001). The BVR is given by

$$\text{BVR}_{jj'} = \sum_{j \in \{0,1\}} \sum_{j' \in \{0,1\}} \frac{(n_{jj'} - e_{jj'})^2}{e_{jj'}}, \tag{2.3.2}$$

where $n_{jj'}$ and $e_{jj'}$ are the observed and expected frequency under the latent class model of a pair of binary variables, Y_j and $Y_{j'}$, respectively. The expected frequency

of the pair of variables can be calculated using the parameter estimates from the fitted model. By summing BVR values across all pairs of variables, we have an alternative overall fit measure, the total bivariate residuals (TBVR):

$$\text{TBVR} = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \text{BVR}_{jj'} \,. \tag{2.3.3}$$

To determine whether a realised TBVR indicates model misfit in the form of the local dependence in residuals, an empirical finite-sample reference distribution of TBVR is needed. This requires resampling techniques (van Kollenburg et al., 2015). One popular resampling technique is parametric bootstrap. However, the parametric bootstrap approach is computationally time-consuming, because it requires estimating the model hundreds of times with hundreds of replicated datasets. The fast-bootstrap resampling method (van Kollenburg, Mulder, & Vermunt, 2018) offers an alternative by directly comparing observed data with a large number of model-generated datasets without having to repeat the model estimation procedure on each of the replicated datasets. We use the fast-bootstrap p-value for the TBVR statistic as the criterion for selecting the initial subset.

Let $\text{TBVR}(S^m_{\mathbf{h}})$ be the observed TBVR computed from a potential initial subset, $S^m_{\mathbf{h}}$. The p-value is calculated as the proportion of L model-generated datasets for which the TBVR value is at least as large as the observed TBVR:

$$p(\text{TBVR}(S^m_{\mathbf{h}})) = L^{-1} \sum_{l=1}^{L} \mathbb{1}\{\text{TBVR}^{(l)} \geqslant \text{TBVR}(S^m_{\mathbf{h}})\}. \tag{2.3.4}$$

Among the potential initial subsets, $S^m_{\mathbf{1}}, \ldots, S^m_{\mathbf{H}}$, the one with the highest fast-bootstrap p-value for the TBVR statistic is selected as the initial subset, denoted as S^m_*.

Another criterion for choosing the initial subset is the (log-)likelihood contribution of individual observations. Parameter estimates from the fitted model are first obtained for each potential initial subset: $\hat{\boldsymbol{B}}_{\{S^m_{\mathbf{h}}\}}, \hat{\boldsymbol{\Lambda}}_{\{S^m_{\mathbf{h}}\}}$, where $\mathbf{h} = \mathbf{1}, \ldots, \mathbf{H}$ represent H m-

dimensional vectors of row indices of $\boldsymbol{Y}$. For each set of parameter estimates, the median of the absolute likelihood contributions for the whole sample is taken. The log-likelihood contribution of individual i (for $i = 1, \ldots, N$) is defined as

$$\ell_i\left(\hat{\boldsymbol{B}}_{\{S_{\mathbf{h}}^m\}}, \hat{\boldsymbol{\Lambda}}_{\{S_{\mathbf{h}}^m\}}; \boldsymbol{y}_i\right) = \log\Big[\sum\nolimits_{k=1}^{K} \hat{c}_{k\{S_{\mathbf{h}}^m\}} \prod_{j=1}^{J} \hat{\pi}_{jk,\{S_{\mathbf{h}}^m\}}^{y_{ij}} (1 - \hat{\pi}_{jk,\{S_{\mathbf{h}}^m\}})^{1-y_{ij}}\Big]. \quad (2.3.5)$$

Finally, the subset which has the minimum median of absolute likelihood contributions is selected as the initial subset, denoted by S_*^m.

The initial subset S_*^m is the one with the minimum median of the absolute log-likelihood contributions:

$$median[S_*^m, \boldsymbol{\ell}] = \min_{\mathbf{h}} \left\{ median \left| \ell\left(\hat{\boldsymbol{B}}_{\{S_{\mathbf{h}}^m\}}, \hat{\boldsymbol{\Lambda}}_{\{S_{\mathbf{h}}^m\}}; \boldsymbol{y}_i\right) \right|; i = 1, \ldots, N \right\}. \quad (2.3.6)$$

In the example of the latent class model, we randomly selected $H = 100$ possible initial subsets, each of size $m = 50$. The one with the highest p-value for the TBVR was chosen to be the initial subset.

Step 2: Progressing in the Forward Search

The "basic" set at the beginning of the FS is the initial subset S_*^m. Moving from S_*^m, it takes a maximum of $(N - m)$ steps until all observations are included in the "basic" set. At each step of the FS, one or more observations can be added to the "basic" set. In addition, when the FS moves from Step m to $m + 1$, the baseline model (i.e. the 2-class model) is fitted to the "basic" set S_*^m. The standard FS sorts all N observations, from the "basic" and the "non-basic" set according to their closeness to the "basic" set. Closeness is established via a criterion based on model estimates from S_*^m. This allows observation to enter and leave the "basic" set at each step of the FS. Alternatively, one can sort just the observations in the 'non-basic set by their 'closeness' to the "basic" set, which is the way used to progress in the FS in our examples.

Common progression criteria include residuals, goodness-of-fit statistics, and likelihood contributions. We used the likelihood contributions. Outliers are expected to be poorly fitted by the baseline model that generates the majority of the data and therefore have lower likelihood contributions. The contribution of individual observation $\mathbf{y}_i$ to the log-likelihood is defined as

$$\ell_i = \log\Big[\sum\nolimits_{k=1}^{K} c_k \prod_{j=1}^{J} \pi_{jk}^{y_{ij}} (1-\pi_{jk})^{1-y_{ij}}\Big]. \tag{2.3.7}$$

One issue that may arise using this criterion is that for categorical data, the likelihood contribution of a response pattern depends on its sample frequency. Therefore, response patterns that are observed only once are unlikely to join until the late stage of the search. To reduce the impact of sample frequencies, one can weight likelihood contributions by dividing them by sample frequencies of corresponding response patterns in the "basic" set, given that the sample frequency is non-zero (Mavridis & Moustaki, 2009). We used the weighted likelihood contribution as the progression criterion while illustrating this example.

Step 3: Monitoring the Forward Search

To reveal individual responses that deviate from the baseline model, we can use forward plots, in which the evolution of goodness-of-fit statistics or model-implied residuals is monitored through the progress of the search as a function of the subset size. A substantial change in the value for the statistics monitored indicates the addition of an outlier to the subset.

In this example, we monitored the fast-boostrap p-value for the total bivariate residuals (TBVR) statistic during the progression of the search. As mentioned in **Step 1**, this fast-bootstrap p-value is less sensitive to sparseness and computationally efficient. The p-value for the TBVR statistic was computed given a sequence of subsets established during the progression: $p(\text{TBVR}(S_*^t))$ for $t = m+1, \ldots, N$.

Figure 2.3.1 is a forward plot showing the evolution of the fast-bootstrap p-value for

the TBVR computed from the "basic" set at each move t of the FS: $p(\text{TBVR}(S_*^t))$, for $t = 51, \ldots, 250$. There is a dramatic decrease from sets S_*^{234} to S_*^{250}. During these steps, 17 out of the 20 constructed outliers join the subset, indicating that the FS manages to distinguish them from the rest of the data. Three other outliers join the "basic" set a bit earlier at $t = 220, 221, 222$.

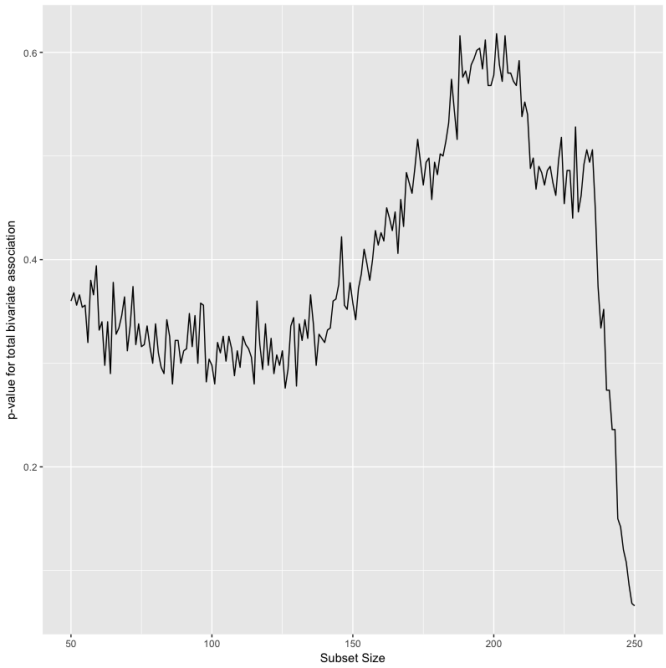


Figure 2.3.1: *Simulation 1: Forward plot of the fast-bootstrap p-value for the* TBVR *computed from the "basic" set at each step of the search. 50 observations are included by the initial subsets.*

2.3.2 Simulation Study: Detecting Latent Dimensionality

The FS has been applied to detect response patterns not generated from the baseline model. Now we are interested in the performance of the FS with the misspecification of latent dimensionality. Specifically, we would like to see how the FS works when the baseline model assumes a normal distribution for the latent trait in the presence of latent population heterogeneity.

Data Generation

In this study, we generated binary data mimicking zero-inflated data containing an excessive number of zero responses. Zero-inflation is common in clinical data and psychiatric data, where a large number of individuals are in good health and therefore do not have symptoms (denoted as zeros) in response to all items that are designed to measure a disorder (Kelley & Anderson, 2008). As a result, an all-zero response pattern often dominates the data in contrast to other not-all-zero response patterns which are very like to come from those with different levels of severity of the disorder.

Wall, Park, and Moustaki (2015) mentioned that the assumption of normality for the latent trait, which is often highly skewed due to zero inflation, leads to biased parameter estimates in modelling zero-inflated data. A mixture IRT model, which is formed by incorporating mixtures into an IRT model, is often used instead for modelling zero-inflated data.

The data-generating model is a mixture IRT model given by Equations (2.2.8) and (2.2.9), dropping the multiple-group label (g). The distribution of the latent trait η is approximated by a 3-class mixture: a mixture of two normal distributions for the heterogeneous pathological sub-population with two levels of severity (i.e. mild and severe), and a degenerate component for the healthy sub-population. The structural part of the mixture IRT model is denoted by

$$\eta \sim \sum_{k=1}^{3} c_k \mathrm{N}\left(\mu_{\eta k}, \sigma^2_{\eta k}\right), \text{ subject to } \sum_{k=1}^{3} c_k = 1,$$

where c_k for $k = 1, 2, 3$ are class proportions for the severe-pathological class, mild-pathological class and non-pathological class, respectively. The non-pathological class ($k = 3$) is fixed at an extremely negative value, because individual responses in this class contribute nothing to the estimation of the latent trait: $\mu_{\eta 3} = -100$ and $\sigma^2_{\eta 3} = 0$. To avoid the indeterminacy, the latent trait is scaled by fixing its overall mean and variance based on the two non-degenerate pathological classes at 0 and

1, without involving the non-pathological class $\sum_{k=1}^{2} c_k \mu_{\eta k} = 0$ and $\sum_{k=1}^{2} c_k(\sigma_{\eta k}^2 + \mu_{\eta k}^2) = 1$.

We consider two simulation settings under different class probabilities. As shown in Table 2.3.3, the non-pathological class ($k = 3$) accounts for 25% in Simulation 2.1 and 50% in Simulation 2.2, and the mild-pathological class ($k = 2$) takes up 50% and 25% in Simulations 2.1 and 2.2, respectively. The proportion of the severe-pathological class ($k = 2$) remains the same in both settings. The table also presents the values for structural parameters.

	N	J	c_1 "severe-"	c_2 "mild-"	c_3 "non-"	$\mu_{\eta 1}$	$\mu_{\eta 2}$	$\sigma_{\eta 1}^2$	$\sigma_{\eta 2}^2$
Simulation 2.1	1000	10	25%	50%	25%	1.00	-0.500	0.833	0.833
Simulation 2.2	1000	10	25%	25%	50%	0.800	-0.800	1.360	1.360

Table 2.3.3: *Simulations 2.1 & 2.2: Class probabilities in the data-generating model.*

Data consisting of $N = 1000$ individuals and $J = 10$ binary items were generated from the above 3-class mixture unidimensional IRT model under the two settings. Under the assumption of conditional independence, individual responses to different items are independent given η. Item parameter values in the data-generating model are specified as follows. Item slopes were fixed at 1: $\lambda_j = 1$ for $j = 1, \ldots, 10$. The values for item intercepts reflect a ladder of severity and are shown in Table 2.3.4.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
ν	-1.500	-1.167	-0.833	-0.500	-0.167	0.167	0.500	0.833	1.167	1.500

Table 2.3.4: *Simulations 2.1 & 2.2: Values for item intercepts ν_j, for $j = 1, \ldots, 10$, in the data-generating model.*

Now a 2-class model with a single pathological class and a non-pathological class is used as the misspecified baseline model. With the assumption that the pathological population is homogeneous, the latent trait for the pathological class under the 2-class model follows a normal distribution rather than a normal mixture.

2.3.2.1 Implementation of Forward Search

We apply the FS procedure for detecting individual cases. To start with, 100 subsets, each of size 200, were randomly drawn from the data. The subset with the highest fast-bootstrap p-value for the TBVR was chosen to initialise the FS, denoted as S_*^{200}. Among 200 individuals in the initial subset, S_*^{200}, 143 individuals are from the non-pathological class, 9 individuals are from the severe-pathological class, and 48 individuals belong to the mildly pathological class. This means that the initial subset was mainly established by individuals at the less severe end of the latent trait. The FS proceeded with the augmented subset using the likelihood contribution as the progression criterion. The fast-bootstrap p-value for the TBVR was computed from the two-class model based on only the subset at each step of the search "i.e.$S_*^{201}, \ldots, S_*^{1000}$). The evolution of the fast-bootstrap p-value is shown in the forward plot (Figure 2.3.2). It is clear that a substantial drop in the p-value starts when the subset size amounts to 534 (S_*^{534}). Before the p-value dives, only 3 observations from the severe-pathological class were added to the "basic" subset when its size is between 201 and 533. S_*^{533} mainly consists of individuals from the non-pathological class and the mild-pathological class. The rest of the observations from the severe-pathological class were added after the subset size amounts to 534. It seems that the drop was largely due to the big inclusion of individuals from the severe-pathological class. It is worth noting that the significant drop in the p-value halfway through the search indicates the inadequacy of the 2-class model.

The same procedure is applied to the data under the second simulation setting. The difference is that the data being analysed here are comprised of nearly 50% of all-zero patterns, twice as many as the first simulation setting, and therefore more skewed. An initial subset of size 100 was chosen based on their associated fast-bootstrap p-value for the TBVR statistic, denoted as S_*^{100}. S_*^{100} contains 5 observations come from the high-pathological class and the rest are all from the non-pathological class. The "basic" set was mainly established by individuals in the non-pathological class. Figure 2.3.3 shows how the fast-bootstrap p-value for the TBVR, an indicative of

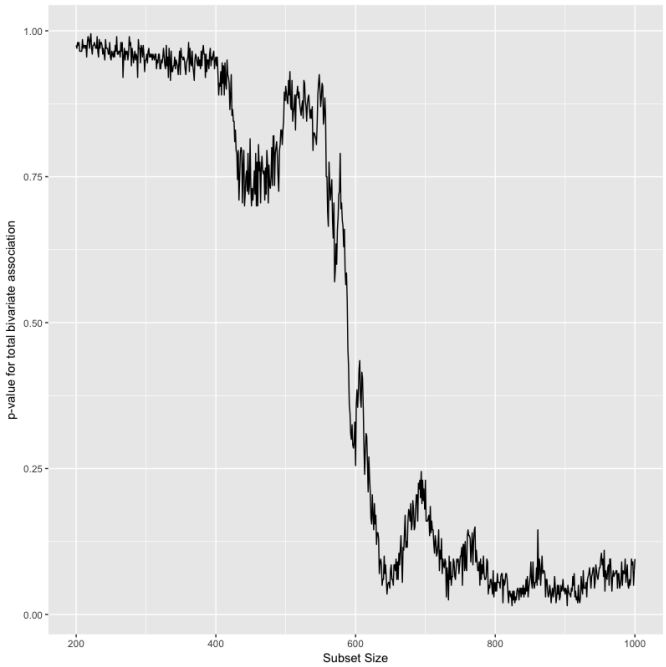


Figure 2.3.2: *Simulation 2.1: Forward plot of fast-bootstrap p-value for the* TBVR *computed from fitting a misspecified 2-class model assuming a homogeneous pathological group. The x-axis label is the size of the "basic" set during the FS.*

model misfit, evolves as more individuals are added to the "basic" set.

According to Figure 2.3.3, 282 individuals from the non-pathological class, 7 individuals from the mild-pathological class, and 1 from the high-pathological class entered the subset before subset size amounts to 390. Among the 100 observations added during subset size $t = 391 - 490$, 78 observations come from the mild-pathological class and the other 22 observations are from the non-pathological class. A sharp decline in the p-value started when individuals from the severe-pathological class began to enter the "basic" set en masse. 196 individuals from the severe-pathological class, 31 from the mild-pathological class, and 7 from the non-pathological entered the "basic" set during the course of subset size $t = 491 - 724$. The inclusion of a large number of individuals from the severe-pathological class makes the fast-bootstrap p-value slump to nearly zero. From the moment when the subset size reaches 725 until the very end of the search, among the newly added observations, there are 48 individuals from the high-pathological class, 134 individuals from the mild-pathological class, and 94 individuals from the non-pathological class.

To summarise, the p-value being monitored dramatically drops to nearly zero halfway

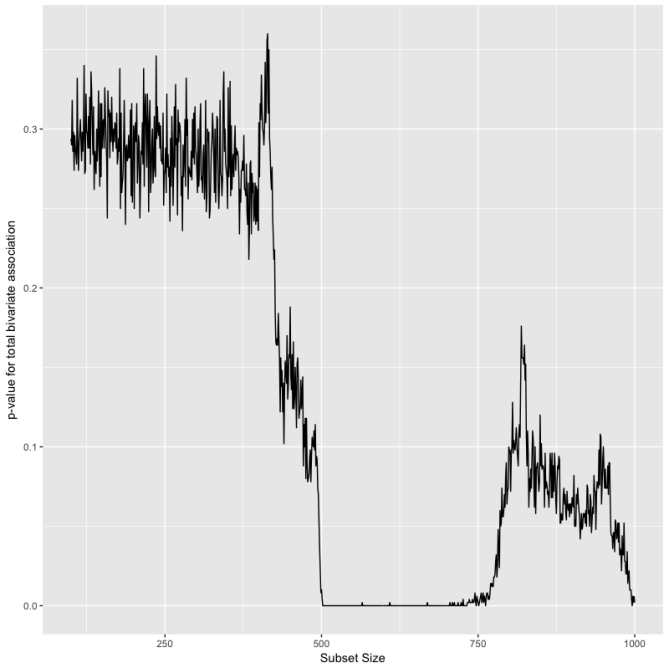


Figure 2.3.3: *Simulation 2.2: Forward plot of fast-bootstrap p-value for the* TBVR *computed from fitting a misspecified 2-class model assuming a homogeneous pathological group.*

through and maintains extremely low values for a while. This implies that the severe-pathological class is not captured by the 2-class model assuming a homogeneous unhealthy class and a non-pathological class. After the subset size reaches 725, the p-value increases a bit but finally drops to nearly zero. The increase was probably due to the inclusion of observations from the mild- or non-pathological class. It is worth noting that the significant drop in the p-value halfway through the search indicates the inadequacy of the 2-class model which assumes the pathological population is homogeneous, although the number of sub-population is remained to be detected.

The simulation study shows that an "outlier-free" initial subset is not always easily found. Initial subsets under both simulation settings were comprised mainly of response patterns well fitted by the baseline model, but there were a few exceptions. For example, in the first setting, individuals at the high-severity level tend to be poorly fitted, as the initial subset was primarily established by individuals belonging to non-pathological and mild-pathological classes.

The presence of response patterns not incorporated by the baseline model in the

initial subset might be problematic, but the problem can be fixed by repeating the FS procedure with multiple randomly chosen initial subsets (Atkinson, 1994). This leads to our next example in which multiple random-started forward searches are carried out to detect latent groups of individuals under a factor mixture model.

2.3.3 Forward Search for Detecting Latent Population Heterogeneity under Factor Mixture Model

Simulation 2 in Section 2.3.2 shows that, in the presence of population heterogeneity, when a carefully selected initial subset mainly consists of individuals from one sub-population, the observations from other sub-populations would be considered as outliers under the fitted model. It is difficult to pick out an "outlier-free" initial subset in this case. The composition of initial subsets in Simulation 2 leads to the idea of using multiple forward searches to explore the structure of multivariate data. This idea is also mentioned in Section 2.1.2. The FS was carried out multiple times to uncover clusters in continuous data through monitoring the Mahalanobis distances whenever the "basic" set size increases (Atkinson et al., 2013).

We now apply multiple forward searches to detect latent population heterogeneity that manifests as latent classes in a factor mixture model. Assume that individual membership in latent classes is indicated by a categorical latent variable $\xi = 1, \ldots, K$. There are two strategies. The first one is running K forward searches, each starting with a carefully-selected, almost "outlier-free" initial subsets solely composed of individuals from one of the K homogeneous classes. However, finding K robust initial subsets is computationally demanding and sometimes infeasible without any prior information about latent classes. The second strategy is running more than K forward searches, each starting with randomly selected initial subsets. As pointed out by Atkinson et al. (2013), forward searches eventually converge regardless of their starting points. The second strategy is also less computationally time-consuming with parallel computation and does not require any prior informa-

tion about the number of latent classes and latent class membership. Therefore, we use the second strategy in the simulation study.

Data Generation

The data consisting of $N = 500$ individuals and $J = 10$ variables were simulated from a single-factor mixture model. The measurement model is given by Equation (2.2.7), and the structural model (2.2.9) assumes the single latent factor, η, to follow a mixture of two normal distributions: $\eta \sim \sum_{k=1}^{2} c_k \mathrm{N}(\mu_{\eta k}, \sigma^2_{\eta k})$, subject to $\sum_{k=1}^{2} c_k \mu_{\eta k} = 0$ and $\sum_{k=1}^{2} c_k(\sigma^2_{\eta k} + \mu^2_{\eta k}) = 1$.

Values for structural parameters $\boldsymbol{B} = (\mu_{\eta k}, \sigma^2_{\eta k}, c_k; k = 1, 2)$ in the data-generating model are listed as follows. $c_1 = 0.400$, $c_2 = 0.600$, $\mu_{\eta 1} = -1.125$, $\mu_{\eta 2} = 0.750$, $\sigma^2_{\eta 1} = 0.150$, and $\sigma^2_{\eta 2} = 0.160$. Values for item parameters $\boldsymbol{\Lambda} = (\boldsymbol{\nu}, \boldsymbol{\lambda})$ in the measurement part of the data-generating model are shown in Table 2.3.5.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
ν	0.941	0.369	-0.911	-1.042	-1.374	0.474	0.651	-0.990	0.942	0.160
λ	1.128	0.804	1.096	1.026	0.888	1.138	0.937	0.820	0.662	0.932

Table 2.3.5: *Simulation 3: Values for item parameters in the data-generating model.*

Now suppose we don't know if the data were generated from more than one distinct population. First, we need to determine whether the data is homogeneous with respect to the baseline model, which is a single-class factor model assuming homogeneity ($K = 1$). The FS procedure for detecting outlying individuals is carried out many times until the deviations from the baseline model show evidence of population homogeneity or heterogeneity. If population heterogeneity is detected, the data are partitioned into subsets consisting of individuals from identified latent classes. To further confirm individuals' latent class memberships, the FS is performed on each partitioned data subset. This two-step procedure for detecting latent population heterogeneity is described as follows.

2.3.3.1 Phase I: Detecting Homogeneous Subgroups

An adequate number of forward searches are performed on the simulated data based on the outlier detection procedure described in Section 2.3.1. Each search starts from a randomly chosen initial subsets consisting of individuals accounting for 5% to 15% of the data depending on sample size. In this simulation study, we ran 100 forward searches, each starting from a random initial subset of size $m = 50$. Each search proceeded by adding individuals to the subset according to their likelihood contributions. No observations are allowed to leave the subset once they are included. The baseline single-factor model assuming homogeneity (with $K = 1$) was fitted to a sequence of data subsets established by each search. To monitor the deviations from the baseline model, the standardised root mean squared residual (SRMR; Hu & Bentler, 1999) which assesses the discrepancy between the sample and model-implied covariance matrices were computed at each step of each FS. The SRMR under a common factor model for continuous outcomes is defined as

$$\text{SRMR} = \sqrt{\sum_{j} \sum_{j' \leqslant j} \frac{r_{jj'}^2}{J(J+1)/2}}, \tag{2.3.8}$$

where $r_{jj'} = \dfrac{s_{jj'}}{\sqrt{s_{jj} s_{j'j'}}} - \dfrac{\hat{\sigma}_{jj'}}{\sqrt{\hat{\sigma}_{jj} \hat{\sigma}_{j'j'}}}$. $s_{jj'}$ and $\hat{\sigma}_{jj'}$ denote the sample and the model-implied covariances between the observed outcome y_j and y'_j $(j' \neq j)$, respectively. s_{jj} and $s_{j'j'}$ denote the sample variances for y_j and y'_j. SRMR $\in [0, 1]$.

The forward plot of the SRMR is shown in the upper-left panel of Figure 2.3.4. The trajectory of the SRMR for each FS is represented by a solid curve. Hu and Bentler (1999) suggested that a value of 0.08 or less is indicative of an acceptable model fit while some others suggested 0.06 as the cutting-off point. We, therefore, obtained the empirical distribution of the SRMR from 100 replications in each step of the FS, while the baseline model assuming homogeneity was fitted to the subsets established during the FS. The 1%, 50% and 99% simulation envelopes are the values of the 99%, 50% and 1% points of the empirical distribution of the SRMR, respectively.

They are represented by dashed curves.

The forward plot shows two peaks, where some trajectory curves of the SRMR exceed the simulation envelops. The peaks are present because the initial subsets from these 88 searches were mainly established by one homogeneous group of individuals. Therefore, the addition of those belonging to the other latent group distorted the fit of the baseline model assuming homogeneity. The presence of two peaks suggests that it is reasonable to assume two sub-populations or two latent classes for the continuous latent variable η.

41 searches pass through the first peak when the "basic" set size is 208, and another 47 searches pass through the second peak when the "basic" set contains 291 individuals. Therefore, one latent class could be formed by the 208 individuals in the "basic" set just before or at the first peak, and the other latent class could be established by the 291 individuals in the "basic" set just before or at the second peak. The sizes and memberships of the two latent classes need further confirmation, which leads to **Phase II**.

The remaining 12 trajectories in grey do not cover either peak, because individuals in both classes are present in initial subsets. All black and grey trajectories eventually settle down and converge regardless of their starting points, which justifies the use of random initial subsets. The end of converged trajectories is still within the simulation envelopes, meaning that the 500 individuals in the dataset belong to either of the two tentatively identified classes.

2.3.3.2 Phase II: Partitioning the data into homogeneous subgroups

To confirm the sizes of two latent classes and the classification of individuals, an adequate number of forward searches are performed solely on each homogeneous class tentatively identified in **Phase I**. The upper-right and the lower-left panel panels in Figure 2.3.4 show the evolution of the SRMR during 100 random-start forward searches from the two tentative latent classes respectively. In **Phase II**,

209 individuals from the "basic" set around the first peak in the upper-left panel were allocated to the first latent class, while 292 individuals from the "basic" set around the second peak in the upper-left panel were classified into the second latent class. One individual is left out and can be allocated to either class. Again, the forward plots do not show the presence of outliers as all trajectories are within their simulation envelopes in the end. This result is consistent with our simulation design, which involves generating the data from a 2-class single-factor model without the presence of isolated or clustered outliers.

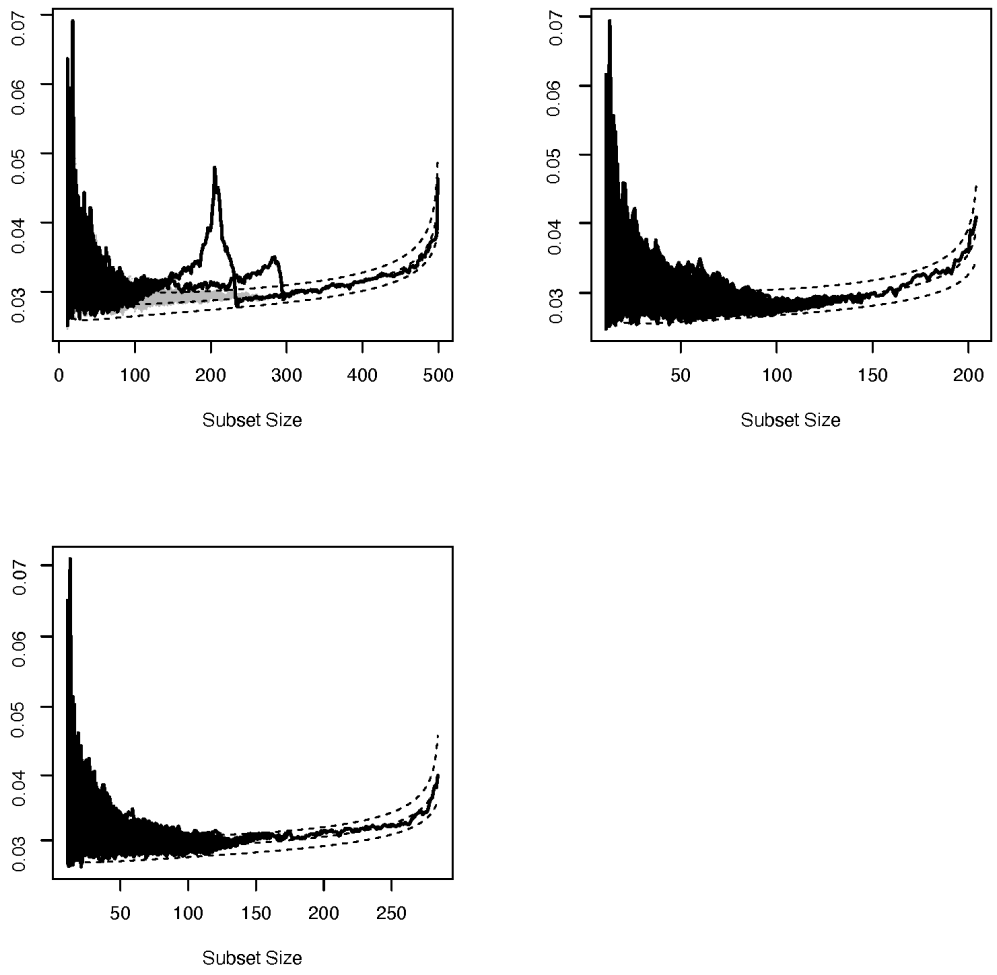


Figure 2.3.4: *Simulation 3: Forward plots of the* SRMR *computed from the "basic" sets during 100 Forward Searches. Dashed lines denote 1%, 50% and 99% simulation envelopes.*

2.3.4 Forward Search for Detecting Items Showing DIF in Multiple Group Analysis

Section 2.1.1 reviews a backward method of identifying DIF items within a latent variable mixture modelling framework (Sawatzky et al., 2018). We now search items lacking measurement invariance forwardly in the latent class modelling context.

Multiple-group latent class model (Equation 2.2.3, Clogg & Goodman, 1984) is used to simultaneously analyse multiple-group data with the purpose of interpreting the latent classes across these groups. The equivalence of latent class measurements requires the class-conditional item response probabilities to be equal across the observed groups (see Section 2.2.2), meaning that all individuals in the same latent class have the same expected responses on manifest variables, regardless of which observed group they belong to (McCutcheon, 2002).

As mentioned earlier in Section 2.2, the assumption of measurement invariance can be tested using goodness-of-fit test statistics or evaluated based on model selection criteria. If there is sufficient evidence against measurement invariance, the next step is to identify DIF items. For example, Masyn (2017) carried out a likelihood-ratio test to compare a model with complete invariance for all items with respect to the grouping variable to a model with class-specific DIF on every item in regard to the grouping variable. Once the test is rejected and there is evidence of measurement invariance (i.e. DIF is identified in at least one item), nested models assuming increasingly restrictive across-group invariance constraints on model parameters are compared (adjusting for multiple testing) until a model in which some items show DIF does not have a significantly worse fit than the model in which all items exhibit DIF.

In this section, we propose to use the FS to directly detect items exhibiting uniform DIF. A measurement-equivalent model is fitted to a sequence of data subsets established throughout the FS, and any significant deviation from the baseline model will be indicative of the inclusion of a DIF item in the subset. Unlike the previous

approach (Masyn, 2017) which makes sequence comparisons among nested models with different degrees of measurement non-equivalence, the FS produces a film of data subsets while the baseline model remains measurement-equivalent. By monitoring the fit between the baseline model and the sequence of subsets, one can assess the effect that each potential DIF item has on the measurement-equivalent model once it is included in the subset.

Data Generation

We simulated data with $N = 500$ observations and $J = 10$ items from a two-group two-class latent class model (Equation 2.2.3 in which $G = K = 2$), where class probabilities $\boldsymbol{B} = (c_1, c_2) = (60\%, 40\%)$ remain the same for both groups. Tables 2.3.6 and 2.3.7 give the conditional item response probabilities $\boldsymbol{\Lambda} = \{\pi_{jk}\}_{10\times 2}$ under Settings A and B respectively. In both settings, Items 1-6 are equivalent while Items 9 and 10 exhibit strong non-equivalence. Setting B differs from Setting A in that its conditional response probabilities for Items 7 and 8 are different for Groups 1 and 2, but the differences are not large enough to change the interpretation of the latent classes. In Setting A, the difference between the two groups in terms of Items 7 and 8 is large and the interpretation of latent classes is different for the two groups.

		Item1	Item2	Item3	Item4	Item5	Item6
Class 1 (60%)		0.06	0.12	0.02	0.11	0.21	0.36
Class 2 (40%)		0.49	0.47	0.24	0.38	0.53	0.62
		Item7	Item8	Item9	Item10		
Class 1 (60%)	Group 1	0.19	0.35	0.22	0.30		
	Group 2	0.75	0.80	0.70	0.65		
Class 2 (40%)	Group 1	0.55	0.55	0.65	0.41		
	Group 2	0.45	0.40	0.60	0.55		

Table 2.3.6: *Simulation 3 Setting A: Conditional item response probabilities used in the data-generating model.*

The FS is carried out with the purpose of detecting non-equivalent items (which deviate from the measurement equivalent baseline model). We keep using the notation from Section 2.3.1 for convenience.

		Item1	Item2	Item3	Item4	Item5	Item6
Class 1		0.06	0.12	0.02	0.11	0.21	0.36
Class 2		0.49	0.47	0.24	0.38	0.53	0.62
		Item7	Item8	Item9	Item10		
Class 1	Group 1	0.19	0.22	0.30	0.35		
	Group 2	0.30	0.40	0.70	0.65		
Class 2	Group 1	0.55	0.65	0.41	0.46		
	Group 2	0.45	0.60	0.60	0.55		

Table 2.3.7: *Simulation 3 Setting B: Conditional item response probabilities in the data-generating model.*

Step 1: Choosing the initial subset

The FS starts from an initial subset containing m items ($m < J$) well fitted by the measurement equivalent baseline model.

Let $\mathbf{h} = (h_1, \ldots, h_m)'$ be an m-dimensional vector of column indices in $\boldsymbol{Y}$. $S^m_{\mathbf{h}} = (y_{h_1}, \ldots, y_{h_m})$ denotes a candidate for the initial subset, where y_{h_1} is the h_1-th column of $\boldsymbol{Y}$ ($1 \leqslant h_1, \ldots, h_m \leqslant J = 10$ and $h_j \neq h_{j'}$).

The total number of possible subsets comprised of m items is $\binom{J}{m}$. Whether it is computationally feasible to investigate all of them depends on the number of items contained in the data. In this study, it is computationally feasible to compare the fit between the baseline model and all possible subsets containing 3 or 4 items (the number is 120 or 210). For data consisting of a large number of items (e.g., Simulation 4 in Section 2.3.2, a feasible way is to randomly select a sufficiently large number of subsets denoted by $(S^m_{\mathbf{1}}, \ldots, S^m_{\mathbf{H}})$ where H is much smaller than $\binom{J}{m}$ and choose the one that the model fits best according to some criterion.

Since the dataset in this example is associated with a sparse contingency table, we use limited-information goodness-of-fit test statistics as a criterion for selecting an initial subset. In addition to goodness-of-fit test statistics, item-specific fit statistics based on the comparison between observed and expected item responses are also useful.

A family of limited-information statistics denoted by M_r ($r < J$) was introduced by Maydeu-Olivares and Joe (2005) to identify the source of the model misfit through examining residuals based on the lower margins up to order r. M_r of the r-th order

under a latent model is given by

$$M_r = M_r(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}) = (\boldsymbol{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}))'(N^{-1}\boldsymbol{\Sigma}_r(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}))^{-1}(\boldsymbol{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}})), \quad (2.3.9)$$

where $N^{-1}\boldsymbol{\Sigma}_r$ is the asymptotic covariance matrix of the sample moments up to order r. $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\pi}_r$ are evaluated at model parameter estimates $(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}})$. $\boldsymbol{p}_r$ and $\boldsymbol{\pi}_r$ represent vectors of sample moments and expected moments up to r-th order, respectively. The limited information statistic of the second order M_2 is most often used for binary data. The M_2 statistic asymptotically follows a chi-squared distribution, where the degrees of freedom equal the difference between the number of moments up to the second order and the number of model parameters q:

$$M_2 \dot{\sim} \chi^2_{df_{M_2}}, \quad (2.3.10)$$

with $df_{M_2} = J + J(J-1)/2 - q$, where q refers to the number of parameters to be estimated. Maydeu-Olivares and Joe (2005) pointed out that for the asymptotic null distribution of M_2 to be valid, the model needs to be correctly specified, the second-order margins cannot be too sparse, and the sample size needs to be relatively large.

A significantly low p-value for the M_2 statistic indicates that one or more outliers are present in the "basic" set. We may not be able to investigate all subsets. Once we find a subset that yields a p-value greater than 5%, we can stop searching. The initial subset is denoted by S^m_*.

In our example, a 2-class model assuming measurement equivalence was fitted to the "basic" set whose size increased as the FS proceeded. The initial subset was comprised of $m = 4$ invariant items. There were $\binom{J}{m} = \binom{10}{4} = 210$ possible candidates for the initial subset. For computational convenience, we only compared 100 candidates, $S^m_{\mathbf{1}}, \ldots, S^m_{\mathbf{100}}$, and chose the one with the highest p-value (above 10%) for the M_2-statistic (Equation 2.3.10) as the "basic" set (denoted as S^4_*) to initialise the search.

Step 2: Progressing in the Forward Search

The "basic" set resulting from **Step 1** is denoted by S_*^m. The search moves to a larger "basic" set S_*^{m+1} by adding one of the $(J-m)$ items outside of the "basic" set. The measurement equivalent 2-class model is fitted to $(J-m)$ possible subsets and then these subsets are ordered based on their associated p-value for the M_2-statistic. The one with the highest p-value is the new "basic" set S_*^{m+1}. The search moves forward till the "basic" set includes every item in the data. A sequence of subsets are established as the search progresses: S_*^t, for $t = m+1, \ldots, J$.

Step 3: Monitoring the Forward Search

To indicate the presence of nonequivalent items in the "basic" set, we assessed how likely the given invariant measurement model could have generated the "basic" sets established throughout the search $(S_*^{m+1}, \ldots, S_*^J)$. Once a non-equivalent item is present in a "basic" set, the invariant measurement model is no longer compatible with the subset. As a result, we can see the change in the value for a goodness-of-fit or an item fit statistic.

We calculated p-value associated with the M_2-statistic, which has been described in **Step 1**, and root mean square error of approximation (RMSEA; Steiger, 1990) for the subsets established during the progress of the search. The RMSEA assesses the approximate fit of the model in the population. It ranges from 0 to 1, with smaller values indicating better model fit. A value of 0.06 or less is indicative of acceptable model fit. An unbiased estimator for the RMSEA associated with the M_2 statistic (Maydeu-Olivares & Joe, 2006) is given by

$$\widehat{\text{RMSEA}}_2 = \sqrt{\frac{\max\{\frac{M_2 - df_{M_2}}{N-1}, 0\}}{df_{M_2}}}, \tag{2.3.11}$$

where the degree of freedom $df_{M_2} = J + J(J-1)/2 - q$ and q refers to the number of parameters to be estimated (e.g., $q = K - 1 + JK$ under a latent class model).

Figure 2.3.5 shows forward plots under Settings A and B, where items newly added at each step of the search are labelled. In Setting A, the p-value for M_2-statistic drops rapidly below 5% while the RMSEA increases, albeit still lower than 0.06 when the first non-equivalent item is included in the subset. The p-value never recovers and the RMSEA eventually exceeds 0.06 during the subsequent steps, when three other non-equivalent items join the subset. In Setting B, the inclusion of moderately non-equivalent items 7 and 8 does not make the p-value fall below 5% or make the RMSEA exceed 0.06. When strongly non-equivalent Items 9 and 10 are included, the p-value immediately drops almost to zero and the RMSEA increases, albeit below 0.06. It seems that the p-value for the M_2 statistic is a more effective indicator of the misfit between the subsets and the baseline model.

2.3.5 Simulation Study: Forward Search for Detecting DIF in Factor Analysis

We have demonstrated how to use the FS to identify items that behave differently across different observed groups in the context of latent class analysis. We are interested in the performance of FS in detecting uniform DIF in high-dimensional data. In this section, the proposed FS procedure for detecting items deviating from the baseline model is applied to data generated from a multiple-group 2PL IRT model with a relatively large number of items.

As mentioned in Section 2.2.3, item invariance in IRT modelling means that the latent construct holds the same property across different groups of individuals. Parameters pertaining to item invariance are constrained to be equal even if they are estimated based on different groups. This assumption is often violated, however. In educational tests, for example, items could be more difficult for one group of test takers than another. Lack of measurement invariance invalidates comparisons between groups in regard to a latent construct because test takers with a certain ability level from one group have different response probabilities than those with the

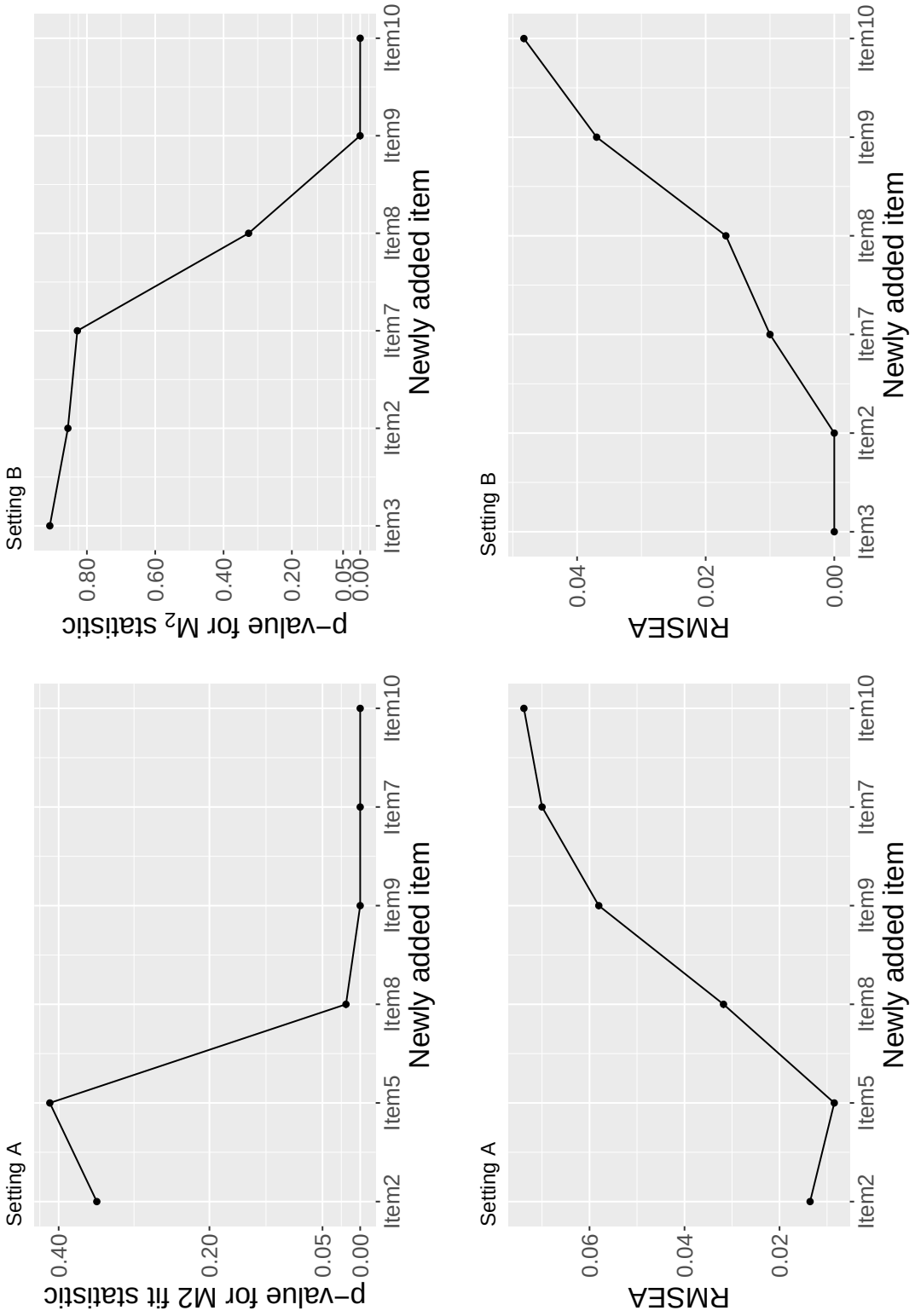


Figure 2.3.5: *Simulation 3: Forward plots of the evolution of the p-value of the M_2 statistic under two simulation settings.*

same ability level from another group.

Data Generation

The FS is applied to a relatively high-dimensional dataset generated from a two-group IRT model (Equation 2.2.8, where $G = 2$). The dataset contains $N = 100$ individuals and $J = 40$ binary items. Table 2.3.8 shows item difficulties for which data were simulated. Four items (i.e. Items 37-40) out of the 40 items are more difficult for Group 1 than Group 2, while the difficulty for the other 36 items remains the same for both groups. To simplify the problem, uniform DIF is considered for the data-generating model, where item discrimination parameters were set to be the same for both groups, as shown in Table 2.3.9.

	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9
	0.1692	-0.2943	0.1362	0.4041	-0.2777	-1.0902	-0.3839	0.1521	0.6393
	Item10	Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18
	1.0866	0.8877	0.5323	0.4614	0.4578	1.0892	0.2916	0.0196	-1.6562
	Item19	Item20	Item21	Item22	Item23	Item24	Item25	Item26	Item27
	0.3822	-0.0610	0.2582	-1.0083	0.6638	-1.1357	-0.9954	-1.5397	0.0716
	Item28	Item29	Item30	Item31	Item32	Item33	Item34	Item35	Item36
	-1.1907	1.1673	0.4559	1.0733	-0.0050	0.0451	0.9977	0.7672	-0.0867
	Item37	Item38	Item39	Item40					
Group 1	-0.4216	-0.7673	-0.2179	0.1226					
Group 2	-0.1054	-0.1918	-0.0745	-0.0306					

Table 2.3.8: *Simulation 4: Item difficulty parameter values used in the data-generating model.*

Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10
0.9899	1.3203	0.6747	1.1697	0.9172	0.7523	1.1471	0.7220	1.4808	0.7417
Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18	Item19	Item20
0.8903	0.9981	0.2858	1.4562	0.8140	0.8548	0.7222	0.9142	1.4763	1.1977
Item21	Item22	Item23	Item24	Item25	Item26	Item27	Item28	Item29	Item30
1.3619	0.9926	0.7286	0.8659	1.0375	1.3444	1.1365	1.6066	0.7914	1.4812
Item31	Item32	Item33	Item34	Item35	Item36	Item37	Item38	Item39	Item40
1.4055	1.1319	1.4715	0.4776	0.5548	0.8334	0.5924	0.9752	0.8925	0.6133

Table 2.3.9: *Simulation 4: Item discrimination parameter values used in the data-generating model.*

2.3.5.1 Implementation of Forward Search

The FS was applied to the data generated from the above multiple-group 2PL IRT model. The model fitted throughout the search was the IRT model assuming equivalence in item difficulty parameters for both groups. The initial subset was established by comparing the p-value for the M_2 statistic among 100 subsets, each containing 20 items. The M_2 statistic is in the form of

$$M_2 = M_2(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}) = \boldsymbol{e}_r(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}})'(N^{-1}\boldsymbol{\Sigma}_2(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}))^{-1}\boldsymbol{e}_2(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}}), \tag{2.3.12}$$

where $\boldsymbol{e}_r(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}})$ denotes the vector containing the model-implied first- and second-order residual proportions, and the matrix $\boldsymbol{\Sigma}_2(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Lambda}})$ involves an asymptotic covariance matrix of the first- and second-order residual proportions and a matrix of derivatives of the model-implied marginal probabilities up to the second order with respect to the model parameters. The asymptotic properties of the M_2 statistic in the context of IRT modelling are similar to those under latent class analysis.

In this example, 50% of the items were included in the initial subset because a large majority of items (36/40) were measurement equivalent. Since it is computationally infeasible to examine every subset containing 20 items, which would be $\binom{J}{m} = \binom{40}{20}$ in total, 100 subsets were sufficient to find a well-established initial subset free of non-equivalent items. We chose the one with the highest p-value as the "basic" set to initialise the FS.

The search proceeded by adding one least non-equivalent item to the "basic" set at each step according to the p-values for the M_2-statistic. Figure 2.3.6 shows the evolution of the p-value for the M_2 statistic based on the "basic" set during the search. It can be seen that when item 39 (one of the non-equivalent items) is added, the p-value drops below 5%, indicating that the model assuming item invariance is not compatible with the subset including item 39. The subsequent entries are three other non-equivalent items. All four non-equivalent items (i.e. Items 39, 49, 38 and 37) are identified hence.

Figure 2.3.6: *Simulation 4: Forward plot of p-value for the M_2 statistic computed whenever an item is added to the "basic" set. The labels on x-axis denotes the entrance of an item outside of the "basic" set.*

2.4 Case Study: European Social Survey Data

2.4.1 Data Description

In this section, we apply the FS procedures for detecting uniform DIF in survey data containing eight items (listed in Table 2.4.1) that measure public opinions on immigration and its effect on the economy and society overall in Germany, UK and Czech Republic from the 7th European Social Survey (ESS; European Social Survey, 2014). These three countries were chosen because they are representative of varying attitudes towards immigration from more welcoming, holding no strong opinions to less welcoming. In most of the 21 ESS countries, there are more respondents who tend to believe that cultural life is enriched by immigrants (Items 4) or that immigration makes the country a better place to live (Items 6) than those who believe that immigration is good for the economy (Item 5) or immigrants make

crime problems worse (Item 3). In terms of population-weighted average observed scores for the eight items, Germany is one of the countries with the highest average score for all the eight items. The Czech Republic, on the contrary, has one of the lowest average scores for all eight items. The average score of the UK falls somewhere in between. Therefore, we assumed that the Czech Republic, the UK and Germany can well represent the 21 ESS countries with distinguishable levels of preference for migration: Germany tends to be more positive towards migrants, the UK comes second, and Czechia tends to be less welcoming. The sample sizes vary across the three countries: the numbers of respondents from Germany, the UK and the Czech Republic are 3,045, 2,264 and 2,148, respectively. The number of incomplete responses goes from 127 for the Czech Republic and 110 for the other countries. A sample of size $N = 1,700$ was randomly drawn from the complete data for each of these three countries.

Table 2.4.1 lists original and binary scales for item responses. The original response categories for items are on a scale from 0 to 10. For the first six items, 0 stands for a negative attitude towards immigration and 10 stands for a positive attitude towards immigration. For items 7 and 8, 0 indicates a positive attitude towards immigrants and 10 indicates for a negative attitude towards immigrants. Item responses are dichotomised by coding ordinal categories 0 to 4 as 0 and categories 5 to 10 as 1. We reversed the scale for items 7 and 8 in order to align with the responses to Items 1-6. As shown in the last column, on the binary scale 1 indicates a positive opinion and 0 indicates a negative opinion towards immigration.

2.4.2 Multiple-group latent class model

Multiple-group latent class model is given by Equation 2.2.3 with $G = 3$ and K (the number of latent classes) to be decided. An *unconstrained latent class model*, where class prevalence, $\boldsymbol{B}^{(g)} = (c_1^{(g)}, \ldots, c_K^{(g)})$, and conditional item response probabilities, $\boldsymbol{\Lambda}^{(g)} = \{\pi_{jk}^{(g)}\}_{8\times K}$ (for $j = 1, \ldots, 8$, $k = 1, \ldots, K$), are freely estimated for each of the three countries.

Item	Description	Ordinal scale	Binary scale
1	Immigrants generally take jobs away or help to create new jobs	0 Take away - 10 Create new	1 positive - 0 negative
2	Immigrants take out more than they put in regarding taxes and welfare or not	0 Take out more - 10 Put in more	1 positive - 0 negative
3	Immigrants make country's crime problems worse or better	0 Worse - 10 Better	1 positive - 0 negative
4	The country's cultural life is undermined or enriched by immigrants	0 Undermined - 10 Enriched	1 positive - 0 negative
5	Immigration is bad or good for country's economy	0 Bad - 10 Good	1 positive - 0 negative
6	Immigrants make the country a worse or better place to live	0 Worse - 10 Better	1 positive - 0 negative
7	Mind if an immigrant of a different race or ethnic group was your boss	0 Not at all - 10 A lot	1 positive - 0 negative
8	Mind if an immigrant of a different race or ethnic group would marry close relative	0 Not at all - 10 A lot	1 positive - 0 negative

Table 2.4.1: *ESS Data: A list of binary items concerning public attitudes towards immigrants.*

Our first consideration is whether the same number of latent classes are required to adequately explain the associations among the eight items in each country. The number of latent classes in each country was selected based on the BIC. The smallest BIC values led to three classes for the UK and four classes for Germany and Czechia. A 3-class model ($K = 3$) was finally chosen for all three countries since for Czechia and Germany the smallest class in the 4-class model ($K = 4$) accounts for less than 5% and does not distinguish from one of the other classes.

Table 2.4.2 shows class prevalences and conditional item response probabilities estimated from the *unconstrained 3-class model*. Based on this result, we characterised the classes as those with negative perceptions, holding no strong views and with positive perceptions of immigrants. However, the differences in class-conditional item response probabilities across countries are striking, making it difficult to directly compare the three countries. To the extent that class-conditional item response probabilities differ across countries, the latent classes are likely to have different meanings for a different country.

In the next step of our analysis, we examine whether equivalence of measured items

	Positive	No strong views	Negative
CZ	0.3440	0.2927	0.3633
UK	0.4412	0.3560	0.2028
DE	0.6519	0.2687	0.0795

	Country	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Positive	CZ	0.6733	0.7297	0.6394	0.8599	0.8007	0.8639	0.7769	0.7304
	UK	0.9091	0.9055	0.7412	0.9576	0.9534	0.9868	0.9391	0.9382
	DE	0.9516	0.8534	0.4914	0.9752	0.9462	0.9730	0.9504	0.9296
No strong views	CZ	0.2511	0.3407	0.1951	0.2172	0.1797	0.1746	0.8378	0.8623
	UK	0.5800	0.4478	0.3636	0.4333	0.4681	0.3843	0.8546	0.8592
	DE	0.5782	0.3634	0.1504	0.5439	0.4529	0.2936	0.9449	0.9510
Negative	CZ	0.0771	0.1088	0.0693	0.1825	0.1382	0.1706	0.0961	0.1047
	UK	0.1261	0.0475	0.0852	0.0359	0.1163	0.0134	0.6269	0.6800
	DE	0.3927	0.2055	0.0557	0.2286	0.3214	0.1154	0.2342	0.2764

Table 2.4.2: *ESS Data fitted by the unconstrained 3-class model model assuming non-equivalence of items: Estimated class prevalences and conditional positive response probabilities by country.*

holds by comparing the *unconstrained 3-class model* with a *restrictive 3-class model* in which across-country equality constraints are imposed on the item parameters ($\boldsymbol{\Lambda}$, dropping g). Since our focus is not on *complete measurement equivalence* (see Section 2.2.2), class prevalences ($\boldsymbol{B}^{(g)}$) across three countries are freely estimated in both models.

Table 2.4.3 shows class prevalences and conditional positive response probabilities estimated from the *restrictive model*. Now we can clearly see three well-defined classes. The class associated with 'positive' attitude towards immigration is defined by generally positive responses to all items. Respondents holding 'no strong views' on immigration are characterised by negative to neutral responses to Items 1-6 and positive responses to Items 7 and 8. Respondents belonging to the 'negative views' class are characterised by negative responses to all items.

	Positive	No strong views	Negative
CZ	0.2406	0.3216	0.4379
UK	0.5067	0.4055	0.0878
DE	0.7123	0.2382	0.0495

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
Positive	0.8933	0.8376	0.5906	0.9480	0.9225	0.9493	0.9235	0.9021
No strong views	0.4091	0.3305	0.2255	0.3162	0.3122	0.2189	0.8866	0.9036
Negative	0.1292	0.1523	0.1069	0.2102	0.1831	0.1923	0.1417	0.1633

Table 2.4.3: *ESS Data fitted by the restrictive 3-class model assuming equivalence of items: Estimated class prevalences and conditional positive response probabilities.*

BIC is used to compare these two models. The BIC values for the *unconstrained model* and *restrictive model* are 42,578.95 and 43,090.78, respectively, suggesting the *unconstrained model* that allows non-equivalence in item parameters is preferable. Thus, the FS is carried out to detect the items exhibiting DIF. The baseline model fitted throughout the search is the *restrictive model* assuming equivalence in item parameters.

2.4.2.1 Implementation of Forward Search

The algorithm starts by searching an initial subset comprised of $p = 3$ least non-equivalent items, which is the "basic" set at the beginning of the FS. There are $V = \binom{J}{p} = \binom{8}{3} = 56$ possible choices for the initial subset denoted by $(S^3_{\mathbf{1}}, \ldots, S^3_{\mathbf{V}})$. The baseline model was fitted to all 56 possible initial subsets. The subset consisting of items 2, 4 and 6 has the highest p-value of the M_2-statistic (Equation 2.3.10) and is, therefore, the initial "basic" set, denoted by S^3_*.

The search moves forward by adding an item from the remaining five items to the "basic" set, (S^4_*). The baseline model was fitted to all possible S^4_* subsets and then ranked the subsets based on the p-value for the M_2-statistic. The subset including Item 7 produces the largest p-value and becomes the new "basic" set S^4_*. The next item that entered the "basic" set was Item 8, which is expected since group differences in Items 7 and 8 given the latent classes are not striking. Following the same progression criterion, a sequence of subsets of increasing sizes was established throughout the search: $S^4_*, S^5_*, S^6_*, S^7_*, S^8_*$.

Figure 2.4.1 visualises the progression of p-value for the M_2-statistic throughout the search. Once Item 5 was included in the "basic" set, the p-value for the M_2 statistic dropped below 5%, meaning that the baseline model that imposes equality constraints on class-conditional item response probabilities was no longer compatible with the subset. As Items 1 and 3 joined following Item 5, the p-value for the M_2 statistic became even lower. Therefore, Item 5 and the subsequently added Items 1 and 3 were flagged as non-equivalent or DIF items across the three groups of data.

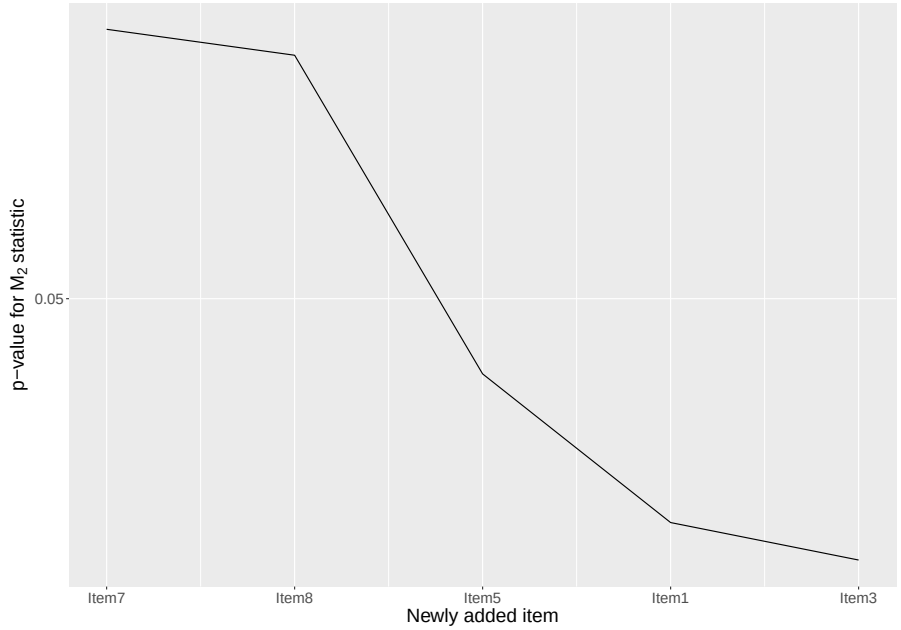


Figure 2.4.1: *ESS Data under latent class analysis: Forward plot of p-value for the M_2 statistic computed whenever an item is added to the "basic" set. The labels on x-axis denotes the inclusion of an item outside of the "basic" set.*

2.4.3 Multiple-group IRT Model

Since IRT models are widely applied to ESS survey data as well, we are interested to know if the detection result is consistent when the baseline model is a 2PL IRT model. Multiple group analysis of (2PL) IRT model is introduced in Section 2.2.3.

Before implementing the FS, one needs to evaluate which level of equivalence fits the data best. Testing for measurement invariance involves comparing models that impose more and more strict equality constraints on item parameters. Therefore, we consider comparing three specific IRT models: (a) a model assuming metric non-invariance (non-uniform DIF items), where item slopes and item intercepts are allowed to vary across groups, and (b) a model assuming scalar non-invariance (uniform DIF items), where item slopes are set to be equal and item intercepts are allowed to vary across groups, and (c) a measurement-equivalent model where item slopes and item intercepts are constrained to be equal across groups.

Model comparison is based on chi-square difference tests and the Bayesian information criterion. The chi-square test for comparing Models (a) and (b) is significant (p-value $< 1\%$) and hence suggests non-equivalence of item slopes across countries.

Model (b), on the other hand, has a significantly (p-value $< 0.1\%$) better fit than Model (c). The BICvalues for the three models are 41,437, 41,388 and 41,829, respectively, indicating that Model (b) is preferred due to its balance between model parsimony and goodness of fit. Based on the chi-square tests and the BIC values, it can be concluded that there exists scalar non-invariance and uniform DIF is identified in some items at least.

2.4.3.1 Implementation of Forward Search

Now we address the detection of uniform DIF items using the FS. Model (c) in which the item intercepts $\boldsymbol{\nu}$ and item slopes $\boldsymbol{\lambda}$ are constrained to be equal across the three countries is the baseline model being fitted during the progression of the FS. Uniform DIF items (with equivalent slopes but non-equivalent intercepts) were not well fitted by Model (c), and we, therefore, expect that their inclusion leads to model misfit.

Once again, the initial subset consisting of items 2, 4 and 6 was selected since its p-value for the M_2 statistic is the highest among all the 56 possible subsets consisting of three items. The forward plot 2.4.2 shows the evolution of the p-value for the M_2 statistic over the progression of the FS. Item 8 entered the "basic" set first, followed by Items 7 and 1. The p-value for the M_2 statistic dropped below 5% when item 5 was added to the "basic" set and dropped even further when the last item (item 3) was included. The forward plot indicates that the inclusion of Items 5 and 3 results in a poor fit of the baseline model assuming equal item intercepts. Therefore the FS identifies uniform DIF in Items 5 and 3. This result is largely consistent with the previous result when the baseline model is the latent class model assuming measurement equivalence. The previous result indicates three items showing uniform DIF, which are Items 5, 1 and 3 in order by the entrance.

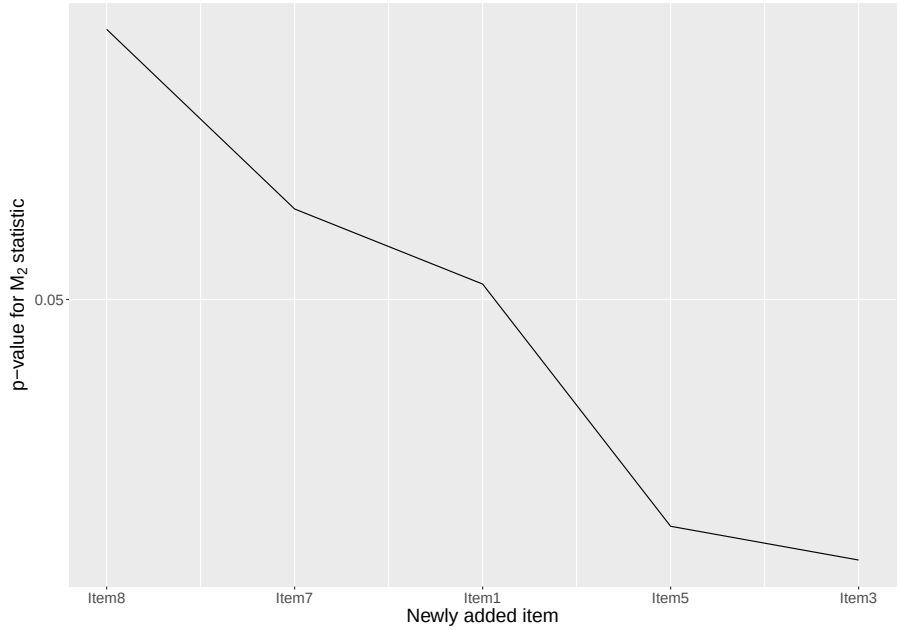


Figure 2.4.2: *ESS Data under a 2PL IRT model: Forward plot of p-value for the M_2 statistic computed whenever an item is added to the "basic" set. The labels on x-axis denotes the entrance of an item outside of the "basic" set.*

2.5 Concluding Remarks

In this chapter, we have explored the scope of the FS as a method of detecting individuals and items that deviate from a baseline model in the latent class and factor mixture modelling contexts.

The FS algorithm was first applied to detect outlying response patterns that depart from a given latent class model for binary data. The effect of the sequential addition of individuals on the fitted model was assessed at each step of the search by monitoring a fast-bootstrap p-value for the total bivariate residuals (TBVR). The fast-bootstrap p-value for the TBVR has proved to be less sensitive to sparseness than those associated with full-information goodness-of-fit statistics and more computationally efficient than other resampling-based p-values. The forward plot of the fast-bootstrap p-value for the TBVR during the FS provides insight into the hidden structure of binary data and deviations from the baseline model.

We also provided simulation studies to show the importance of selecting an (almost) "outlier-free" initial subset. Without a robust start, outliers may enter the "basic"

set in the early stage rather than during the last few steps of the search, which makes it difficult to separate them from the rest of the data. This leads to an alternative strategy, which is running multiple random-start forward searches instead of a single FS with a robust start. The alternative strategy was then applied to detect latent population heterogeneity in the form of latent classes in a factor mixture model. Multiple forward searches were simultaneously carried out and the resultant forward plots clearly indicate the existence of population heterogeneity. The number of latent classes is often determined using information criteria, for example, the Bayesian Information Criterion (BIC; Fraley & Raftery, 1998). The main difference between the BIC and FS in the detection of latent classes or mixture components is that the latter can inform the classification of individuals while the former cannot. Plus, the BIC could be sensitive to outliers (Atkinson & Riani, 2008) and tends to result in overfitting by assuming an excessive number of mixture components for the latent factor without sequential model comparison.

Another development of the FS made in this chapter was detecting variables showing DIF. The FS starts from a subset established by a relatively small number of items well fitted by a measurement-equivalent baseline model, and proceeds by adding the least non-equivalent items to the "basic" set. The presence of non-equivalent items in the "basic" set is expected to affect the fit of the baseline model assuming measurement equivalence. This effect is assessed by the p-value for a limited-information goodness-of-fit statistic. The FS algorithm for detecting non-equivalent items was applied to an ESS dataset consisting of items for public attitudes towards immigration among three European countries. The FS managed to identify non-equivalent items under two baseline models, including a measurement-equivalent latent class model assuming equivalent class-conditional item response probabilities and a measurement-equivalent 2PL IRT model.

The advantage of using the FS to detect DIF is that the baseline model fitted to subsets of the data is measurement-equivalent and thus leads to the least number of parameters to be estimated. Measurement invariance is conventionally tested by

evaluating how well a specified model fits the dataset and comparing a sequence of nested models with increasingly restrictive across-group equality constraints on item parameters (Masyn, 2017). As for the FS, there is no need to consider different specifications of measurement (non-)equivalence. Instead, the FS selects a film of data subsets and monitors the fit (or misfit) between these subsets and the measurement-equivalent baseline model.

In terms of future research, there are three areas that we would like to investigate. First, we have yet to address the multiple testing issues. When the FS is used to detect outliers, the issue of hypothesis testing appears whenever we declare individuals or items to be outliers. As a result, multiple hypotheses are being tested simultaneously when the FS is applied to detect individuals or items in a dataset. As Becker and Gather (1999) pointed out, it is not enough to test whether individual cases or items are outlying, we need to develop multiple outlier testing methods. Riani, Atkinson, and Cerioli (2009) developed a simultaneous test of outlyingness that is intended to find outliers in a proportion of the multivariate normal data.

Second, while the diagnostic statistics (e.g., the fast-bootstrap p-value for the TBVR, the p-value for the M_2 statistic) monitored in the FS proved to capture the effect of the addition of outliers on model fit in our study, we may need to further assess their reliability through comparing them to other types of residuals and goodness-of-fit statistics.

Finally, the current work on FS is expected to be adapted to other latent variable models for different types of data. The purpose of this method is to identify data subsets that have not been generated by a baseline model. The type of the baseline model does not change the procedure of the FS. What may need to be changed is the criteria used in the progression of the search, and more importantly, the diagnostic statistics used for assessing the effect of the sequential addition of individuals or items on the baseline model throughout the FS.

Chapter 3

Two-way Outlier Detection Model

3.1 Introduction

In Chapter 2, the Forward Search (FS) is employed to detect individuals and items deviating from a specified baseline model, including outlying response patterns, latent population heterogeneity and items exhibiting DIF. The FS is capable of detecting one-way outliers based on attributes of persons or items but not two-way outliers defined by attributes of both persons and items. It is often the case that item response data contain outliers among both the individuals and the items. Two-way outliers as such can lead to a substantial deviation from a carefully specified latent variable model which may be supported by substantive theory and historical data. On the other hand, as mentioned in Section 1.1, two-way outliers often provide valuable insight into the data, and thus, it is of substantive interest to detect them.

As mentioned earlier, two-way outliers may arise due to Differential Item Functioning (DIF; Holland & Wainer, 1993; Millsap, 2012), a phenomenon that is widely observed in educational testing, psychological measurement, as well as many other areas of social research. It happens when a subset of items does not measure subgroups of individuals in the same way. The subgroups are sometimes defined by observed variables such as gender, ethnicity, years of education etc, but they can also be unobserved due to unobserved sources of population heterogeneity. In the context

of educational testing, it is often the case that a subset of items is "easier" or "more difficult" for a certain subgroup than the others. In this case, responses from the subgroup of test takers to the subset of items can be viewed as two-way outliers, as they may be poorly fitted by a factor model that fits the rest of the data well.

A related, while more challenging, the problem is the detection of two-way outliers due to latent DIF (Cho, Suh, & Lee, 2016), for which not only the DIF items but also the group membership of individuals are not known a priori. One such example is the two-way detection of item compromise that benefits test takers through their preknowledge of compromised items in educational tests (Cizek & Wollack, 2017). Item compromise is a type of cheating behaviour which is increasingly common in computer-based tests. Since those tests are administrated on a regular basis (C. Wang, Xu, Shang, & Kuncel, 2018), items used in previous tests tend to be reused later. Test takers who participate in a test earlier may share test material with those who take it later (C. Wang, Zheng, & Chang, 2014). Newly issued items might also be leaked to online forums or test-prep or firms. Davey and Nering (2002) mentioned a typical item compromise scandal that occurred in 1994. Kaplan, a test-prep firm, sent its employees to take ETS tests. They then prepared their pupils for ETS tests based on the test items they had memorised and reconstructed. The leaked or exposed items are known as compromised items. Test takers with preknowledge of compromised items before the administration of an exam are expected to gain score inflation. Therefore, it is important to know which items are compromised and which test takers have preknowledge of them, and the failure to do so threatens the validity of test scores and the fairness of tests.

Looking beyond this, latent DIF may also exist in educational tests due to other reasons according to Cho et al. (2016). Similar problems associated with latent DIF also occur in other areas besides educational testing. For example, in political science, it has been well recognised that roll call voting data of the United States Congress can largely be described by a liberal-conservative latent dimension with some minor deviations (Poole & Rosenthal, 1991; Poole, Rosenthal, & Koford,

1991). Through a two-way outlier-detection formulation, i.e., by detecting outlying legislators and roll calls that do not fit the unidimensional model, one may gain a better understanding of the patterns of roll call voting that cannot be explained by the liberal-conservative dimension. Latent DIF may also exist in psychological measurement data in which a two-way outlier-detection formulation can facilitate the discovery of minor psychological traits and the relevant groups. The promising prospect for practical use further motivates us to develop a model-based method of detecting two-way outliers due to latent DIF.

A two-way outlier detection model is supposed to have a baseline model for standard or honest response behaviour and an additional model component for atypical response behaviour because of latent DIF. The choice for the baseline model is discussed as follows. As mentioned earlier in Section 2.2.4, latent factor models (Bartholomew, Knott, & Moustaki, 2011) have been widely applied to multivariate data, particularly item response data consisting of individuals' responses to a set of measured items. Unidimensional and multidimensional factor models, also known as Item Response Theory (IRT) models (Embretson & Reise, 2000; Reckase, 2009), are commonly used to model test takers' responses to items in educational tests. In this context, a latent factor is often interpreted as the ability that the test is designed to assess. In psychology, multidimensional factor models are typically used to describe respondents' answers to items in a psychological questionnaire (Wirth & Edwards, 2007), where the latent factors are interpreted as psychological traits (e.g., personality traits). In political science, similar models, also known as ideal point models, are used to model voting behaviours (Bafumi, Gelman, Park, & Kaplan, 2005), where the latent factors are typically interpreted as voters' political standing.

To account for latent DIF, an additional model component needs to be built upon the baseline model. While statistical methods have been well established for modelling and detecting DIF (Millsap, 2012), models for detecting two-way outliers due to latent DIF and procedures for statistical decisions remain to be developed. We, therefore, propose a two-way outlier detection model to fill the gap that adds a

latent class model component to a factor model. The factor model component serves as a baseline model for data without outliers, and a latent class model component is used to capture the two-way outliers. Specifically, the proposed model imposes latent class structures among both the individuals and the items, rather than only assuming latent classes among the individuals as in the classical latent class analysis (Allman, Matias, & Rhodes, 2009; Goodman, 1974; Lazarsfeld & Henry, 1968). The proposed model is closely related to, but also substantially different from, existing statistical models and methods for the detection of outliers in multivariate data (see e.g., Candès, Li, Ma, & Wright, 2011; Hadi, 1992; Mavridis & Moustaki, 2008, 2009; Reiser, 1996; C. Wang & Xu, 2015; C. Wang, Xu, & Shang, 2018; Zhou, Li, Wright, Candès, & Ma, 2010)

Under the proposed model, statistical decision theory is established for the detection of two-way outliers. Motivated by compound decision theory for multiple testing (Benjamini & Hochberg, 1995; Efron, 2004, 2008, 2012; Efron, Tibshirani, Storey, & Tusher, 2001; Robbins, 1951; Sun & Cai, 2007; C.-H. Zhang, 2003), we propose the local False Discovery Rate (FDR) and local False Non-discovery Rate (FNR) as compound risk measures for the detection of two-way outliers. Decision rules are developed based on these measures, for which optimality results are established. The statistical inference and decision-making are performed under a fully Bayesian framework, for which a Markov chain Monte Carlo (MCMC) algorithm is developed. Since our model involves many discrete latent variables, standard MCMC algorithms such as Gibbs and Metropolis-Hastings can suffer from slow mixing (e.g., Celeux, Hurn, & Robert, 2000; Richardson & Green, 1997). We tackle this problem by applying the parallel tempering technique (Geyer, 2011).

The proposed method is applied to cheating detection based on data from the single administration of a non-adaptive test. It simultaneously detects outlying test takers and items as potential cheaters and compromised items. The proposed method uses item response data and item response time data, which are often collected in computer-based testing, for improving outlier detection. As shown via our real

data analysis and simulation studies, incorporating response time information can improve outlier detection accuracy. Our simulation results further suggest that the proposed model is quite robust against various forms of model misspecification, even though it relies on some parametric assumptions that may not be satisfied perfectly in practice.

The detection of test takers who benefit from item preknowledge (cheaters) and compromised items has received much attention among quantitative researchers in education. Specifically, McLeod, Lewis, and Thissen (2003) proposed a person-fit index for the detection of cheaters in computerised adaptive testing, under an IRT model. For non-adaptive testing, Belov (2013) proposed a person-fit index for characterising the outperformance of a student on the compromised items, assuming that the set of compromised items is known. Under a similar setting, Sinharay (2017a) proposed likelihood-ratio and score tests for the detection of cheaters, and Segall (2002) and Shu, Henson, and Luecht (2013) proposed IRT models for item preknowledge and developed Bayesian classification procedures. For the detection of compromised items, O'Leary and Smith (2017) and X. Wang and Liu (2020) proposed methods based on data from the single administration of a non-adaptive test. These approaches require knowledge of a subset of non-compromised items to first identify a set of potential cheaters. The detection of compromised items relies on the identified cheaters in the first stage. Under an online setting where data from multiple tests are sequentially collected, Veerkamp and Glas (2000), J. Zhang (2014), Chen and Li (2019), and Chen, Lee, and Li (2020) formulated the detection of compromised items as a sequential change detection problem and proposed sequential procedures. We refer the readers to three edited volumes, Wollack and Fremer (2013), Kingston and Clark (2014) and Cizek and Wollack (2017), for a comprehensive review of related works. Note that most of the existing methods focus on the detection of either cheaters or compromised items, and often require prior information which is not always available, for example, a given subset of non-compromised items. In contrast, the proposed method can simultaneously detect

both test takers with item preknowledge and compromised items without such prior information.

The rest of the chapter is organised as follows. In the following section 3.2, we propose a statistical model for detecting two-way outliers in multivariate data and discuss its application to simultaneous detection of cheaters and compromised items due to item preknowledge in an educational assessment. Statistical decision theory is developed under a fully Bayesian framework in Section 3.3. In the next section 3.4, we describe the Bayesian inference procedures. The proposed two-way outlier detection model is applied to a real dataset from a computer-based licensure test in Section 3.5. Simulation studies are presented in Section 3.6, where we further evaluate the classification and detection performance of the proposed model under various situations. Concluding remarks are given in Section 3.7. The proof of a theoretical result from Section 3.3 and details of the developed MCMC algorithm are presented in Appendix C.1.

3.2 Models

3.2.1 A Two-Way Outlier Detection Model for Multivariate Data

3.2.1.1 Background and Notation

Consider N individuals responding to J items. Let Y_{ij} be individual i's response to item j. We focus on binary responses, i.e., $Y_{ij} = 0, 1$, where the two types of responses may correspond to incorrect and correct answers in educational testing, and "no" and "yes" responses in psychological measurement, among others. We use $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ to denote the response vector from individual i and use $\boldsymbol{Y} = (Y_{ij})_{N\times J}$ to denote the response matrix. When item-response data are collected digitally rather than by paper and pencil, which is becoming more and more popular these days, response time data may also be collected. Let T_{ij} denote the amount of

time individual i spends to answer item j and $\boldsymbol{T} = (T_{ij})$ denote the data matrix for response times.

In what follows, we discuss the two-way outlier detection model for $(\boldsymbol{Y}, \boldsymbol{T})$. We introduce a latent binary variable ξ_i that takes the value 1 when individual i is an outlier and 0 otherwise. Similarly, η_j is a latent binary variable that takes the value 1 when item j is an outlier and 0 otherwise. Table 3.2.1 illustrates how data are affected by the two-way outliers in the proposed model. (Y_{ij}, T_{ij}) are modelled with the outlier model if, and only if, both $\xi_i = 1$ and $\eta_j = 1$, which represents a typical phenomenon of latent DIF. In the cheating detection application, items with $\eta_j = 1$ correspond to the leaked/compromised items and individuals with $\xi_i = 1$ correspond to test takers who have preknowledge about all the compromised items before taking the test. In this context, a baseline model will capture the normal item-response behaviour, and the outlier model will capture the behaviour of the test takers with preknowledge when responding to the compromised items. In particular, the outlier model will allow test takers to have a higher probability of answering leaked items correctly and with a shorter response time (see, e.g. C. Wang, Xu, & Shang, 2018). The proposed model is described below.

		Item j	
		$\eta_j = 0$	$\eta_j = 1$
Person i	$\xi_i = 0$	Baseline Model	Baseline Model
	$\xi_i = 1$	Baseline Model	Outlier Model

Table 3.2.1: *The two-way outlier structure in the proposed model.*

3.2.1.2 Proposed model

We start with a relatively more general model and then give specific examples. We introduce $\boldsymbol{\theta}_i$ and $\boldsymbol{\tau}_i$ as the person-specific parameters, also known as the factors, that drive the item responses and the response times, respectively. Both $\boldsymbol{\theta}_i$ and $\boldsymbol{\tau}_i$ can be unidimensional or multidimensional, but their dimensions are typically assumed to be much smaller than J. We also introduce $\boldsymbol{\beta}_j$ and $\boldsymbol{\alpha}_j$ to denote the

item-specific parameters for the item responses and the response times, respectively. To simplify the notation, we use $\Theta_i = (\boldsymbol{\theta}_i, \xi_i, \boldsymbol{\tau}_i)$ and $\Delta_j = (\boldsymbol{\beta}_j, \eta_j, \boldsymbol{\alpha}_j)$ to denote the person- and item-specific parameter vectors, respectively.

The proposed model consists of two submodels, one for binary item responses and one for continuous response times. The item-response submodel takes the following logistic form

$$P(Y_{ij} = 1|\Theta_i, \Delta_j, \delta) := p(\Theta_i, \Delta_j, \delta) = \frac{\exp(h_1(\boldsymbol{\theta}_i, \boldsymbol{\beta}_j) + \xi_i\eta_j\delta)}{1 + \exp(h_1(\boldsymbol{\theta}_i, \boldsymbol{\beta}_j) + \xi_i\eta_j\delta)},$$

where δ is a non-negative parameter and $h_1(\cdot, \cdot)$ is a pre-specified function. Given $\xi_i = 0$ or $\eta_j = 0$,

$$p(\Theta_i, \Delta_j, \delta) = \frac{\exp(h_1(\boldsymbol{\theta}_i, \boldsymbol{\beta}_j))}{1 + \exp(h_1(\boldsymbol{\theta}_i, \boldsymbol{\beta}_j))}$$

is the baseline item-response submodel for non-outlying item responses. When $\xi_i = \eta_j = 1$, the term $\xi_i\eta_j\delta \neq 0$ captures the deviation from the baseline model. In particular, for our application, the parameter δ is set to be non-negative to let the probability of providing a correct answer (i.e., $Y_{ij} = 1$) increase when individual i and item j are outliers. In this context, δ may be interpreted as the advantage that a test taker gains from item preknowledge. In other applications, this sign constraint can be removed if such prior information is not available. To keep the model parsimonious, the parameter δ is assumed to be the same across all the outlying individuals and items. As will be discussed in Section 3.2.2, this assumption can be relaxed.

The function h_1 should be chosen based on knowledge about the baseline model from substantive theory and/or historical data. We give two parametric examples of h_1 below, but point out that h_1 can also take a non-parametric form as in non-parametric IRT models (Douglas, 1997; Duncan & MacEachern, 2008; Ramsay & Winsberg, 1991).

Example 1. *The Rasch model (Rasch, 1960) is one of the most popular IRT models in educational testing and is also widely used in many other areas. In particular, the*

licensure test to be studied in Section 3.5 is designed and scored under this model. The Rasch model assumes that both $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_j$ are unidimensional. With slight abuse of notation, we denote them by non-bold typeface θ_i and β_j, respectively. This model assumes that $h_1(\theta_i, \beta_j) = \theta_i - \beta_j$, which leads to

$$p(\Theta_i, \Delta_j, \delta) = \frac{\exp(\theta_i - \beta_j + \xi_i\eta_j\delta)}{1 + \exp(\theta_i - \beta_j + \xi_i\eta_j\delta)}. \tag{3.2.1}$$

In the context of educational testing, θ_i and β_j are interpreted as the ability of test taker i and the difficulty of item j, respectively. When there are no outliers, the probability of correctly answering an item is monotone increasing with one's ability θ_i and monotone decreasing with the item's difficulty β_j. When $\xi_i = 1$ and $\eta_j = 1$, $logit(P(Y_{ij} = 1|\Theta_i, \Delta_j, \delta)) = \theta_i - \beta_j + \delta$. That is, the item response function still takes a Rasch form, but the log odds increases by a constant drift δ. This Rasch-type item-response submodel given by Equation (3.2.1) will be further discussed in the rest of the chapter, given its suitability for our case study in Section 3.5.

Example 2. *It may be the case that the J items simultaneously measure K factors, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})$, for which a multidimensional factor model is needed. In that situation, we may set $h_1(\boldsymbol{\theta}_i, \boldsymbol{\beta}_j) = \beta_{j0} + \beta_{j1}\theta_{i1} + \cdots + \beta_{jK}\theta_{iK}$, where $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{jK})$ contains $K + 1$ item-specific parameters. The item-response submodel then becomes*

$$p(\Theta_i, \Delta_j, \delta) = \frac{\exp(\beta_{j0} + \beta_{j1}\theta_{i1} + \cdots + \beta_{jK}\theta_{iK} + \xi_i\eta_j\delta)}{1 + \exp(\beta_{j0} + \beta_{j1}\theta_{i1} + \cdots + \beta_{jK}\theta_{iK} + \xi_i\eta_j\delta)}. \tag{3.2.2}$$

When $\xi_i = 0$ or $\eta_j = 0$, (3.2.2) becomes the multidimensional two-parameter logistic model (Reckase, 2009) which includes the two-parameter logistic model (Birnbaum, 1968) as a special case when $K = 1$.

The response-time submodel is specified similarly to the item-response submodel. Specifically, we consider a log-normal model which assumes that

$$\log(T_{ij})|\Theta_i, \Delta_j, \gamma, \kappa \quad \sim \quad N\left(h_2(\boldsymbol{\tau}_i, \boldsymbol{\alpha}_j) - \xi_i\eta_j\gamma, \kappa\right),$$

where γ is another non-negative parameter that plays a similar role to δ in the item-response submodel, $h_2(\cdot, \cdot)$ is a pre-specified function, and $\kappa > 0$ is the variance of the normal distribution. In our application, the parameter γ is set to be non-negative to allow the response time for outlying individuals and items to be shorter (test takers with preknowledge tend to answer the compromised items faster). That is, when $\xi_i = 1$ and $\eta_j = 1$, the mean log-time is reduced from the baseline level $h_2(\boldsymbol{\tau}_i, \boldsymbol{\alpha}_j)$ to $h_2(\boldsymbol{\tau}_i, \boldsymbol{\alpha}_j) - \gamma$. In that context, γ may be interpreted as the reduction in response time due to item preknowledge. Similar to the discussion about parameter δ, the sign constraint on γ can also be removed if there is no such prior knowledge about the response times. We assume that the same γ and κ are shared by all the individuals and items for model parsimony, which can be relaxed.

The choice of function h_2 is similar to the choice of function h_1 in the item-response submodel. In what follows, we give a specific example, but also point out that other choices of h_2 are possible. In particular, one can choose h_2 so that the baseline response-time submodel is consistent with the one proposed in van der Linden (2007).

Example 3. *Similar to the Rasch-type model in Example 1, we let both $\boldsymbol{\tau}_i$ and $\boldsymbol{\alpha}_j$ be unidimensional and denote them by non-bold typeface τ_i and α_j. We let function h_2 take the form $h_2(\tau_i, \alpha_j) = \alpha_j - \tau_i$, which leads to*

$$\log(T_{ij}) | \Theta_i, \Delta_j, \gamma, \kappa \quad \sim \quad N\left(\alpha_j - \tau_i - \xi_i \eta_j \gamma, \kappa\right). \tag{3.2.3}$$

When $\xi_i = 0$ or $\eta_j = 0$, we obtain the baseline model for response times

$$\log(T_{ij}) | \Theta_i, \Delta_j, \gamma, \kappa \quad \sim \quad N\left(\alpha_j - \tau_i, \kappa\right).$$

In the context of educational testing, τ_i can be interpreted as the speed factor of test taker i and α_j can be interpreted as the time-consumingness of item j. When there are no outliers, the mean response time is monotone increasing with the item-specific time-consumingness α_j and monotone decreasing with the person-specific

speed factor τ_i. This response-time submodel will be applied to our case study in Section 3.5.

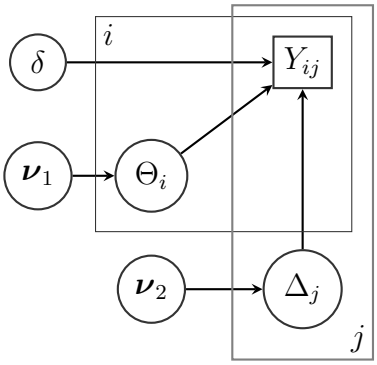
Like many other latent variable models, conditional independence assumptions are imposed. We first assume that (Y_{ij}, T_{ij}), $j = 1, \ldots, J$, are conditionally independent given Θ_i, Δ_j, δ, γ, and κ. Such a conditional independence assumption across items is often known as the local independence assumption. We further assume that Y_{ij} and T_{ij} are conditionally independent given Θ_i, Δ_j, δ, γ, and κ, meaning that all the person effects on the response and response time distribution are captured by the person parameters. Such conditional independence assumptions are commonly made in latent variable models for item responses and response times. We refer the readers to van der Linden (2007) for the substantive justifications.

We further adopt a Bayesian hierarchical modelling framework, under which parameters Θ_i, Δ_j, δ, γ, and κ are treated as random variables. Specifically, we let Θ_i, $i = 1, \ldots, N$, be independent and identically distributed (i.i.d.) samples from distribution $g_1(\Theta|\boldsymbol{\nu}_1)$ and $\Delta_j, j = 1, \ldots, J$, be i.i.d. samples from distribution $g_2(\Delta|\boldsymbol{\nu}_2)$, respectively, where g_1 and g_2 characterise the population of individuals and the domain of items, respectively. Both g_1 and g_2 are taken to be parametric distributions and use $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ as generic notations for the hyperparameters of the two distributions, respectively. This hierarchical modelling structure is visualised in Figure 3.4.1b using a graphical model representation. We showcase the specification of g_1, g_2, and the priors for $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2$, δ, γ, and κ in Section 3.4.1 under the specific model with item-response submodel (3.2.1) and response-time submodel (3.2.3).

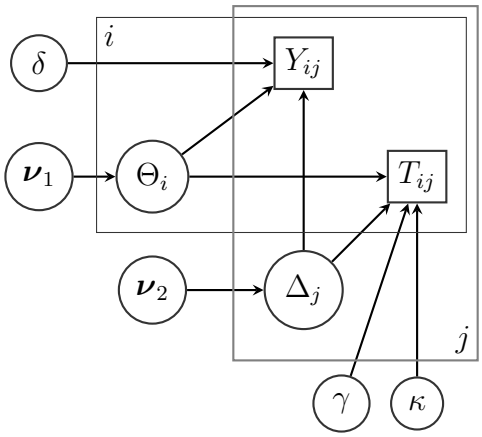
3.2.1.3 Model without response time data

Sometimes, response time information is not collected, for example, in paper-and-pencil-based educational tests. In that case, response times are missing completely at random and the proposed model reduces to a model for item responses. This *reduced model* only contains parameters from the item-response submodel and the

corresponding hyperpriors. The graphical representation of this reduced model is given in Figure 3.2.1a, where the reduced person and item parameters are denoted by $\Theta_i = (\boldsymbol{\theta}_i, \xi_i)$ and $\Delta_j = (\boldsymbol{\beta}_j, \eta_j)$, respectively, and the corresponding hyperparameters are still denoted by $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, respectively.



(a) *General graphical representation of the reduced model for item responses Y_{ij}'s when all the response times T_{ij}'s are missing completely at random. The reduced person parameters are denoted by $\Theta_i = (\boldsymbol{\theta}_i, \xi_i)$ and the reduced item parameters are denoted by $\Delta_j = (\boldsymbol{\beta}_j, \eta_j)$.*



(b) *Graphical representation of the proposed model for the joint distribution of item responses Y_{ij}'s and response times T_{ij}'s. The full person parameters are denoted by $\Theta_i = (\boldsymbol{\theta}_i, \boldsymbol{\tau}_i, \xi_i)$ and the full item parameters are denoted by $\Delta_j = (\boldsymbol{\beta}_j, \boldsymbol{\alpha}_j, \eta_j)$.*

Figure 3.2.1: *General graphical representation of the reduced model and the full model. The boxes are plates representing replicates. The two outer plates represent individuals and items, respectively, and the inner plate presents an item response in Figure 3.2.1a and item response and response time in Figure 3.2.1b.*

3.2.1.4 Application to the detection of item preknowledge

The proposed model framework requires that the baseline model is correctly specified. Any deviation from it is solely attributed to item preknowledge and not to other aberrant situations, such as more than one latent dimension (multidimensionality) needed to explain the associations among the items. Although this assumption might appear to be strong, it can still be examined using historical test data with no leaked items and test takers from the same population.

Furthermore, the validity of the model interpretation also depends on the extent to which our parametric assumptions hold. Section 3.7 discusses how some of those parametric assumptions can be violated in practice and can be also relaxed. In ad-

dition, as shown via simulation studies in Section 3.6, the proposed two-way outlier detection model tends to be robust against several forms of model misspecification.

Finally, we emphasise that, given the sensitivity of decisions regarding cheating in tests and the relatively strong assumptions of the proposed model, the latent classes resulting from the two-way detection should be interpreted with caution (i.e. leaked items and test takers with preknowledge). Results from our model can provide warnings to the test administrators, but the detected outlying cases should be further investigated and verified using additional sources of information.

3.2.2 Model Generalisations

Key components of the proposed two-way outlier detection model are the interaction terms $\xi_i\eta_j\delta$ and $\xi_i\eta_j\gamma$ in the item-response and response-time submodels, respectively. Specifically, the effects of the two-way outliers are characterised by the parameters δ and γ in the two submodels, respectively, and they are assumed to be the same across all the outliers. This assumption can be relaxed to allow for heterogeneity among the outliers. One way to relax this assumption is by assuming each drift parameter to be the sum of a person-specific parameter and an item-specific parameter. For example, one may replace $\xi_i\eta_j\delta$ by $\xi_i\eta_j(\delta_i + \delta'_j)$ in the item-response submodel, where δ_i and δ'_j are non-negative person- and item-specific drift parameters, respectively.

Moreover, in the current framework, the outlier model is essentially unidimensional, as a result of the imposed two-way latent class structure. In the application to the detection of item preknowledge, it means that a test taker has preknowledge of either all or none of the leaked items. However, when there are multiple sources of item leakage and test takers with item preknowledge have access to one or more of those sources, this assumption can be relaxed by assuming multiple latent classes among both the individual and item outliers.

3.2.3 Related Works

Factor analysis in the presence of outliers has received much attention in the literature, but mainly focuses on the detection of outlying cases/individuals rather than items as well. One line of research is on the robust estimation of factor models (Moustaki & Victoria-Feser, 2006; Pison, Rousseeuw, Filzmoser, & Croux, 2003; Yuan & Bentler, 1998, 2001). Another line of research focuses on the detection of outliers among the individuals who do not fit a baseline factor model, using residual-based procedures (Reiser, 1996) or forward search procedures (Hadi, 1992; Mavridis & Moustaki, 2008, 2009). All these works only consider outlying individuals. The proposed two-way outlier detection method is among the very few attempts to simultaneously classify individuals and items as outliers.

Although several models that combine factor and latent class modelling have been proposed for detecting aberrant behaviours (Bolt, Cohen, & Wollack, 2002; Boughton & Yamamoto, 2007; Goegebeur, De Boeck, Wollack, & Cohen, 2008; Shu et al., 2013; C. Wang & Xu, 2015; C. Wang, Xu, & Shang, 2018), none of them is about the two-way classification of individuals and items.

Another feature of our model is that it does not require any prior knowledge about the outlying individuals and items (e.g. a subset of compromised items). Shu et al. (2013) proposed a Deterministic, Gated item response theory Model (DGM) for data consisting only of item responses. This model makes similar assumptions to our item-response submodel, except that (1) the DGM assumes the known status of each item (i.e., whether each item is compromised or not), and (2) the drift parameter for cheating (i.e., δ in the current model) is assumed to be person-specific in the DGM. Our model is more closely related to C. Wang and Xu (2015) and C. Wang, Xu, and Shang (2018) who also assume a mixture of log-normal distribution for response times from normal and aberrant response behaviours. Like the proposed method, these works also do not require prior knowledge about the test takers with preknowledge of the leaked items. The main difference is that C. Wang and Xu

(2015) and C. Wang, Xu, and Shang (2018) focus on identifying person-item pairs for which aberrant behaviours are involved, rather than directly classifying test takers and items. Therefore, they allow aberrance in any person-item combination, by introducing a person-and-item specific latent variable to indicate the status of each response. By having person-and-item specific latent variables, the models of C. Wang and Xu (2015) and C. Wang, Xu, and Shang (2018) tend to be more flexible than the proposed model, in the sense that these models allow data to deviate from the baseline model along more directions. Consequently, these models may be preferred when data involve multiple types of aberrant behaviours, such as rapid guessing and cheating. On the other hand, unlike the proposed method, the models of C. Wang and Xu (2015) and C. Wang, Xu, and Shang (2018) do not directly lead to classifications of the test takers and items, let alone quantifying the uncertainty of the classifications. To detect test takers with preknowledge and leaked items, follow-up analysis is needed based on the posterior distributions of the person-and-item specific latent variables. Therefore, these methods are not as straightforward as the proposed one, if the main goal is to perform the two-way detection of individuals with preknowledge and leaked items.

3.3 Statistical Decision Theory for Two-way Outlier Detection

In what follows, we provide statistical decision theory for the detection of two-way outliers under the proposed model, assuming the model is correctly specified. We start with the classical Bayesian decision theory and then develop compound decision rules for the detection of outlying individuals and items.

3.3.1 Introduction to Bayesian Decision Theory

As individuals and items are essentially mathematically exchangeable, we only discuss the Bayesian decision theory for the detection of outlying individuals. We denote D_i as the decision on individual i, where $D_i = 1$ means flagging the individual as an outlier and $D_i = 0$ otherwise. A false positive error happens when $D_i = 1$ and $\xi_i = 0$ and a false negative error happens when $D_i = 0$ and $\xi_i = 1$. Decisions on the detection of outlying individuals involve a trade-off between these two types of errors, whose importance may be asymmetric. For example, in the application to the detection of item preknowledge, a false positive error corresponds to an innocent test taker being flagged as a cheater and a false negative error corresponds to a cheater not being flagged. These two types of errors have substantially different consequences (Skorupski & Wainer, 2017).

To apply Bayesian decision theory to this classification problem, we need to specify the relative cost of a false positive error, denoted by $\zeta \in (0, 1)$, which further implies that the relative cost of a false negative error is $1-\zeta$. Then the Bayes risk is defined as

$$\mathcal{R}(D_i) := \zeta P(D_i = 1, \xi_i = 0) + (1 - \zeta)P(D_i = 0, \xi_i = 1). \tag{3.3.1}$$

Following the classical Bayesian decision theory (see, e.g., Chapter 2, Shao, 2003), the optimal decision rule which minimises the Bayes risk is obtained by comparing the posterior probabilities with the relative cost ζ. That is, an individual is classified as an outlier if the posterior probability is larger than ζ.

This Bayesian decision rule depends on the relative cost ζ. However, this parameter may not be easy to specify in practice, as the relative importance of a false positive error is often hard to quantify. In what follows, we discuss how this parameter may be chosen adaptively based on a compound risk which is obtained by aggregating information from the entire set of individuals.

3.3.2 Compound Decision for Detecting Outlying Persons

We evaluate decision-making at an aggregated level for all individuals. This involves solving N decision problems simultaneously and thus is called a compound decision problem (Robbins, 1951; Sun & Cai, 2007; C.-H. Zhang, 2003). Given a decision rule, hypothetically, the results can be classified into four categories, as summarised in Table 3.3.1, where N_{00}, N_{01}, N_{10}, and N_{11} denote the numbers of true negative, false positive, false negative, and true positive, respectively. The quality of decisions can be quantified by two quantities. One is the False Discovery Proportion (FDP) $N_{01}/\max\{N_{\cdot 1}, 1\}$, which is the proportion of non-outliers among the detections. In the application to cheating detection, this gives the proportion of innocent test takers among those who are flagged as cheaters. The denominator is chosen so that this proportion is well-defined even when $N_{\cdot 1} = 0$. The other quantity is the False Non-discovery Proportion (FNP) $N_{10}/\max\{N_{\cdot 0}, 1\}$, which is the proportion of outliers among the non-detections. It is worth noting that, however, the FDP and FNP cannot be directly used because the outliers are not directly observable. As an alternative, we use the posterior means of the FDP and FNP, which are known as the local FDR and local FNR, respectively. Similar measures have been proposed for solving compound decision problems in Efron (2004, 2008, 2012); Efron et al. (2001), among others. Given data and a decision rule, the local FDR and local FNR are completely determined under the proposed model.

	Not flagged as outlier	Flagged as outlier	Total
Non-outlier	N_{00}	N_{01}	$N_{0\cdot}$
Outlier	N_{10}	N_{11}	$N_{1\cdot}$
Total	$N_{\cdot 0}$	$N_{\cdot 1}$	N

Table 3.3.1: *A summary of the outcomes of detecting outlying individuals. Note that this table is hypothetical, as in real applications, outliers and non-outliers are directly observable.*

Suppose that the consequence of a false positive error is more severe than that of a false negative error, which may be the case for the detection of cheaters in educational testing. Then a sensible decision criterion is to minimise the local FNR

while controlling the local FDR to be below a pre-specified threshold ρ. Given the practical meaning of local FDR, the threshold ρ should be much easier to specify than the relative cost in the Bayesian decision rule discussed previously. For instance, by setting $\rho = 0.01$, we approximately control the proportion of non-outliers to be below 1% among those who are detected as outliers.

Now consider a Bayesian decision rule with a relative cost ζ. We discuss how the optimal ζ is determined by the above decision criterion based on the local FDR and local FNR with a given threshold ρ. For ease of exposition, we use $\tilde{\boldsymbol{Y}}$ as a generic notation for the data, where $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$ when only item responses are collected and $\tilde{\boldsymbol{Y}} = (\boldsymbol{Y}, \boldsymbol{T})$ when both item responses and response times are available. Specifically, given relative cost ζ, the Bayesian decision for each test taker i can be written as

$$D_i(\zeta) = 1_{\{P(\xi_i=1|\tilde{\boldsymbol{Y}})>\zeta\}}. \tag{3.3.2}$$

Under our fully Bayesian setting and given threshold ζ, the local FDR becomes

$$\text{fdr}_\zeta(\tilde{\boldsymbol{Y}}) = \frac{\sum_{i=1}^N D_i(\zeta) P(\xi_i = 0|\tilde{\boldsymbol{Y}})}{\max\{\sum_{i=1}^N D_i(\zeta), 1\}}, \tag{3.3.3}$$

which only depends on the posterior probabilities $P(\xi_i = 1|\tilde{\boldsymbol{Y}})$, $i = 1, \ldots, N$. The local FNR can be obtained similarly as

$$\text{fnr}_\zeta(\tilde{\boldsymbol{Y}}) = \frac{\sum_{i=1}^N (1 - D_i(\zeta)) P(\xi_i = 1|\tilde{\boldsymbol{Y}})}{\max\{\sum_{i=1}^N (1 - D_i(\zeta)), 1\}}.$$

As summarised in **Proposition 1** below, the optimal relative cost is given by $\zeta^* = \inf\{\zeta : \text{fdr}_\zeta(\tilde{\boldsymbol{Y}}) \leqslant \rho\}$. That is, the decision rule given by $D_i(\zeta^*)$, $i = 1, \ldots, N$, minimises the local FNR under the constraint that the local FDR is below ρ. The proof of **Proposition 1** is given in Appendix B.

Proposition 1. *Given data $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$ or $(\boldsymbol{Y}, \boldsymbol{T})$, the local FDR $fdr_\zeta(\tilde{\boldsymbol{Y}})$ as a function of ζ is non-increasing and left-continuous, and the local FNR $fnr_\zeta(\tilde{\boldsymbol{Y}})$ is non-*

decreasing in ζ. *Thus,*

$$\zeta^* = \inf\{\zeta : fdr_\zeta(\tilde{\boldsymbol{Y}}) \leqslant \rho\} \tag{3.3.4}$$

solves the optimisation

$$\min_\zeta fnr_\zeta(\tilde{\boldsymbol{Y}}), \quad s.t. \quad fdr_\zeta(\tilde{\boldsymbol{Y}}) \leqslant \rho. \tag{3.3.5}$$

The corresponding optimal decision rule is $D_i(\zeta^*) = 1_{\{P(\xi_i=1|\tilde{\boldsymbol{Y}})>\zeta^*\}}$.

Given posterior probabilities $P(\xi_i = 1|\tilde{\boldsymbol{Y}})$, the optimal decision is easy to obtain. The computation is described in Algorithm 1.

Algorithm 1 (Optimal compound decision)**.** *Let the posterior probabilities* $P(\xi_i = 1|\tilde{\boldsymbol{Y}})$ *and threshold* ρ *for local FDR be given. The optimal relative cost* ζ^* *is given by the following steps.*

1. *Sort the posterior probabilities in an increasing order. Denote the sorted values as* $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(N)}$.

2. *Compute the cumulative means* $c_{(0)} \leqslant c_{(1)} \leqslant c_{(2)} \leqslant \cdots \leqslant c_{(N)}$, *where*

$$c_{(0)} = 0, \text{ and } c_{(i)} = \frac{\sum_{j=1}^{i} p_{(j)}}{i}, i = 1, \ldots, N.$$

3. *Let* $i^* = \max\{i : c_{(i)} \leqslant \rho\}$.

Then the optimal relative risk is given by $\zeta^* = p_{(i^*)}$, *if* $i^* > 0$, *and* $\zeta^* = 0$, *if* $i^* = 0$.

This local FDR control procedure can be viewed as the Bayesian version of the well-known Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995) for multiple hypothesis testing. The BH procedure is designed to control the FDR, which is defined as the unconditional expectation of the FDP. Unlike the proposed procedure that is based on posterior probabilities, the BH procedure achieves the control of FDR using p-values from multiple testing. Under the proposed Bayesian

framework, it seems more straightforward to control local FDR as in the proposed procedure. Also note that the FDR is automatically controlled by controlling local FDR, due to the relationship between conditional and unconditional expectations.

Another possible decision criterion is to control the posterior probability of making at least one false discovery, which corresponds to the Family-Wise Error Rate (FWER) under the frequentist setting. This FWER-type criterion exerts a more stringent control over false discovery than the proposed one by its definition. Therefore, the proposed procedure has greater power at the cost of increased rates of false positives. In this sense, the proposed local FDR control procedure is more suitable when having a large number of individuals and thus a large number of decisions need to be made simultaneously.

For certain applications, false negatives may have a more significant consequence than false positives. Then it may be more suitable to minimise the local FDR while controlling local FNR. As the definitions of local FDR and local FNR are mathematically symmetric, the above procedure can be easily adapted.

3.3.3 Compound Decision for Detecting Outlying Items

The compound decision theory developed above can be adapted to the detection of outlying items, based on the posterior probabilities for the item-specific binary indicators η_j. In the application to the detection of item preknowledge, when there is sufficient evidence suggesting that an item is compromised, it should be removed to maintain the quality of the item pool. This decision problem faces a trade-off between the financial cost of item pool replenishment and the need of maintaining the quality of the item pool. For high-stake tests, test fairness is usually the first priority and thus false negatives may have a more significant consequence than false positives. In that case, it becomes more sensible to minimise the local FDR, under the constraint that the local FNR is below a suitable threshold (e.g., 1%).

3.4 Bayesian Inference

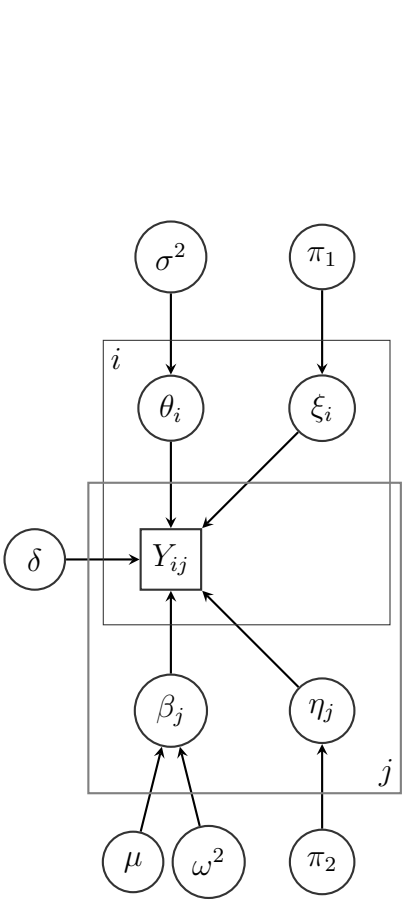
The proposed two-way outlier detection model is estimated under a fully Bayesian framework, where model parameters are all treated as random variables. In this section, we first specify the hierarchical models for the full model for item response and response times and the reduced model for item responses. We then propose a parallel tempering MCMC algorithm to sample model parameters from their joint posterior distribution. The details are presented in Appendix C.1. Finally, we discuss model comparison methods in the Bayesian framework.

3.4.1 Hierarchical Model Specification

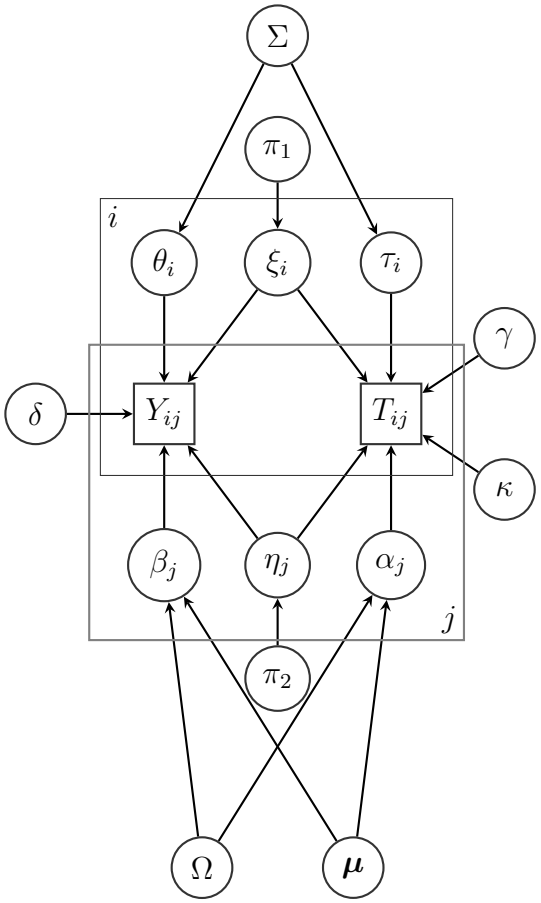
The hierarchical frameworks for the reduced model and the full model are shown in Figures 3.4.1a and 3.4.1b. In the full Bayesian framework, global parameters ($\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, δ, γ, and κ), person- (Θ_i's) and item-specific (Δ_j's) parameters are all treated as random effects (represented by circles in the figures below). In both figures, the boxes are plates representing replicates. The two outer plates represent persons and items, and the inner plate presents an item response.

3.4.1.1 Prior and Hyperprior Specification

We showcase the specification of prior and hyperprior distributions under the specific model with item-response submodel (3.2.1) and response-time submodel (3.2.3). We start with the specification of g_1, the joint distribution of $\Theta_i = (\theta_i, \xi_i, \tau_i)$. It is assumed that (θ_i, τ_i) follows a bivariate normal distribution $N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{2\times 2}$. Note that a person's ability and speed are typically correlated, which is why we assume a bivariate normal distribution for (θ_i, τ_i). Similar settings are adopted in existing models for item responses and response times; see e.g., van der Linden (2007). We further assume that the latent indicator ξ_i is independent of (θ_i, τ_i), following a Bernoulli distribution Bern(π_1). This independence assumption can be



(a) *Hierarchical framework for the two-way outlier detection model based on item response data only. The boxes with $i = 1, \ldots, N$ in the top-left corner indicates that each parameter inside is specific to a value of i. The same explanation is also applied to the boxes with $j = 1, \ldots, J$ in the bottom-right corner. Note that the mean of the person parameters θ_i's is fixed at 0 for the sake of identifiability.*



(b) *Hierarchical framework for the two-way outlier detection model that incorporates item responses and response times. Note that $\mu = (\mu_1, \mu_2)$ and Ω are the mean and covariance matrix for the two item parameter, (β_j, α_j), for $j = 1, \ldots, J$, while Σ is the covariance matrix for the two person parameters (θ_i, τ_i) for $i = 1, \ldots, N$. The mean for (θ_i, τ_i) is fixed at $(0, 0)$ for identifiability.*

Figure 3.4.1: Hierarchical framework for the reduced and the full models for detecting two-way outliers.

relaxed by modelling the conditional distribution of ξ_i given (θ_i, τ_i), for example, by a logistic regression model. This relaxation is left for future investigation.

We now specify g_2, the joint distribution of $\Delta_j = (\beta_j, \eta_j, \alpha_j)$. Similar to that of g_1, we first let (β_j, α_j) follow a bivariate normal distribution $N(\boldsymbol{\mu}, \Omega)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and $\Omega = (\omega_{ij})_{2\times 2}$. It is further assumed that η_j is an independent Bernoulli random variable, $\text{Bern}(\pi_2)$.

It remains to specify the prior for positive parameters δ, γ, κ, as well as the priors

for hyperparameters π_1, π_2, μ_1, μ_2, Ω, and Σ.

1. As δ is positive, we assume a half-Cauchy prior distribution with scale 2.5. This is regarded a weakly informative prior, following the suggestions given in Gelman (2006), Gelman, Jakulin, Pittau, and Su (2008) and Polson and Scott (2012).

2. γ is assumed to follow the same half-Cauchy prior distribution as δ.

3. κ is assumed to follow an inverse Gamma distribution, IG$(0.001, 0.001)$, where the shape and scale parameters are both set to 0.001. This choice follows the suggestion in Chapter 5, Lunn, Jackson, Best, Thomas, and Spiegelhalter (2012).

4. π_1 and π_2 are assumed to be i.i.d., following a beta distribution Beta$(2, 2)$. This prior distribution can be regarded as a weakly-informative prior given the sample and item sizes in our application. It is suggested in Agresti and Coull (1998) and Carlin and Louis (2000), Chapter 2, as the prior for a proportion parameter. We choose this distribution rather than a uniform distribution, because π_1 and π_2 are believed to not locate on the boundaries of the interval $[0, 1]$.

5. μ_1 and μ_2 are assumed to be i.i.d., following a normal distribution $N(0, 5^2)$. The standard deviation 5 is chosen based on the scales of μ_1 and μ_2 in the current application, under which this prior may be regarded as weakly informative.

6. Σ and Ω are assumed to be i.i.d., following an inverse Wishart distribution where the scale matrix, IW(Ψ, ν), $\Psi = ((2, 0)^\top, (0, 2)^\top)$, and the degree of freedom $\nu = 2$. This choice follows the suggestion in Chapter 6, Lunn et al. (2012). Under this prior distribution, σ_{11}, σ_{22}, ω_{11}, and ω_{22} marginally follow an inverse Gamma distribution IG$(1/2, 1)$.

3.4.1.2 Induced Priors and Hyperpriors for Reduced Model

For the reduced model of item response data, the priors and hyperpriors are induced by those for the full model. For completeness, we list the induced priors and hyperpriors below.

1. For $\Theta_i = (\theta_i, \xi_i)$, θ_i and ξ_i are independent, following normal distribution $N(0, \sigma_{11})$ and Bernoulli distribution $\text{Bern}(\pi_1)$, respectively.
2. Similarly, for $\Delta_j = (\beta_j, \eta_j)$, β_j and η_j are independent, following normal distribution $N(\mu_1, \omega_{11})$ and Bernoulli distribution $\text{Bern}(\pi_2)$, respectively.
3. δ follows a half-Cauchy prior distribution with scale 2.5.
4. π_1 and π_2 are i.i.d., following a beta distribution $\text{Beta}(2, 2)$.
5. μ_1 follows a normal distribution $N(0, 5^2)$.
6. σ_{11} and ω_{11} are i.i.d. $\text{IG}(1/2, 1)$.

3.4.2 Computation

Statistical inference is carried out under a full Bayesian setting. An MCMC algorithm is developed for the computation[1].

3.4.2.1 Parallel Tempering MCMC

This computation is non-trivial, due to the presence of many discrete variables and the interactions between them. More specifically, the model involves person- and item-specific binary latent variables ξ_i and η_j. The complexity of simulating these variables by MCMC is similar to the simulation of discrete systems like mixture models and Ising-type models. Such systems typically involve many well-separated

[1]The R code for the MCMC algorithm can be found on https://github.com/YanLu-stats/OD2WIRT.

local modes and thus suffer from the issue of slow mixing (e.g., Celeux et al., 2000; Katzgraber, Trebst, Huse, & Troyer, 2006; Richardson & Green, 1997). Tempering methods (Geyer, 2011) provide a powerful tool for exploring distributions with many local modes. Specifically, parallel tempering (PT), which is also known as the Metropolis-coupled Markov chain Monte Carlo or replica exchange MCMC sampling, is chosen for the computation of the proposed model. More precisely, parallel tempering simulates multiple MCMC chains simultaneously, and a Metropolis-Hastings sampler can be used for the MCMC sampling within each chain.

In parallel tempering, multiple MCMC chains interact in order to effectively explore the state space and thus improve the performance of mixing. The target distributions of these chains are obtained by tempering the original posterior density, i.e., raising the original posterior density to different powers $T^{-1} \in [0, 1]$, where T is known as the 'temperature'. The original posterior density is included by setting $T = 1$. A chain corresponding to a higher temperature tends to have a flatter target distribution, for which the MCMC sampler is less likely to be trapped at local modes and thus has fast mixing. In contrast, when the temperature is low, the MCMC chain is more likely to be trapped and thus suffer from slow mixing. Parallel tempering improves the mixing of the low-tempered MCMC chains, by exchanging information between chains with adjacent temperatures. That is, at each iteration, a pair of chains with adjacent temperatures is randomly chosen and a Metropolis-Hastings update is used to decide whether to swap their parameter states.

The use of the algorithm requires some tuning, including (a) the step sizes of random-walk Metropolis-Hastings updates within each chain, (b) the number of temperature levels, and (c) the temperature values. For (a), we follow the suggestion given by Roberts and Rosenthal (2001); that is, we tune the step sizes to achieve an acceptance rate of around 2.3. For (b) and (c), it is suggested to follow the theoretical guidance given in Atchadé, Roberts, and Rosenthal (2011). Further details of this algorithm are given in Appendix C.1. For the implementation of the decision procedures in Section 3.3, the posterior distributions of ξ_i and η_j are approximated by

the posterior samples from this MCMC algorithm.

3.4.2.2 Remarks: Thinning

MCMC chains are likely to be strongly autocorrelated and therefore produce clumpy samples that do not characterise the posterior distribution in question accurately if the chains are not long enough. To obtain a more accurate estimation of the target distribution, one should get rid of as much autocorrelation as possible.

Thinning can be used to reduce autocorrelation. M samples from a thinned chain with sufficiently small autocorrelation will almost certainly produce a more precise estimation of the posterior than M unthinned steps with high autocorrelation. However, to get M samples from a thinned chain, $M \times n$ steps are needed for keeping every n-th sample. With such a long chain, the autocorrelation has probably been averaged out anyway. As Link and Eaton (2012) mentioned, a longer unthinned chain usually yields better estimates of the true posterior than the shorter thinned chain. Furthermore, Link and Eaton (2012) also mentioned that thinning is not useful if we aim to reduce standard errors in parameter estimates from MCMC samples. Therefore, to reduce autocorrelation through thinning by n steps, we can simply run an unthinned chain n times as long. As Jackman (2009) pointed out, thinning is not a strategy for avoiding long runs required for obtaining more precise estimates, but a strategy for manipulating the otherwise overwhelming number of MCMC samples.

3.4.2.3 Remarks: Empirical Bayes Estimation

Instead of taking a fully Bayesian setting, it is also possible to adopt an empirical Bayes framework (Casella, 1985; Efron, 2014; Robbins, 1956), under which $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2$, δ, γ, and κ are treated as fixed parameters and estimated by maximum likelihood estimation, while the person- and item-specific parameters Θ_i and Δ_j are treated as random variables. However, due to the complex structure of the current model, the expectation-maximisation algorithm, which is a standard approach to empirical

Bayes inference, is computationally infeasible. Without specifying priors, the tricky thing is how to constrain certain parameters (e.g. π_1 and π_2) to avoid a local optimal solution. A tailored stochastic optimisation algorithm needs to be developed.

3.4.3 Model Comparison

We use model comparison methods to answer the following questions that may be of substantive interest in specific applications. That is, does our item response data show evidence of the existence of outlying individuals and items? If so, does item response time information help detect these outliers? These questions may be answered by Bayesian model comparison.

To answer the first question, we compare the proposed model for item responses with the baseline item-response model which does not contain outliers. These two models are the same, including the specification of the priors and hyperpriors, except that the hyperparameters π_1 and π_2 are set to be 0 and thus $\xi_i\eta_j\delta = 0$ for all i and j in the baseline model. The preference of the proposed model against the baseline model suggests the existence of outliers.

To answer the second question, we compare the proposed model for item responses and response times with a null model for the same data. This null model is the same as the proposed model, except that the drift parameter γ in the response-time submodel is set to be zero. When $\gamma = 0$, it means that there is no difference between the outlying and non-outlying individuals in their response-time distributions. If the proposed model is preferred to the null model, it suggests that response times contain information about the outliers. Thus, incorporating response time information may better inform the detection of outliers.

3.4.3.1 Introduction to Deviance Information Criterion

Deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) has been employed to find the most adequate and yet parsimonious model

among several candidate models given observed data. This model comparison criterion has also been applied to hierarchical IRT models, particularly. The DIC is proposed as an alternative for model comparison when it is difficult to determine the number of effective parameters. It is defined as the sum of a deviance measure and a penalty term for the effective number of parameters based on a measure of model complexity. In comparing two models, the one with a smaller DIC value is preferred.

The models mentioned earlier in this section are compared based on the DIC, and more specifically, a marginalised DIC in which the person- and item-specific parameters are treated as latent variables or random effects and integrated out. We choose the marginalised DIC, instead of the conditional DIC that incorporates the person- and item-specific variables in the focus of the analysis, because the marginalised DIC often performs better in comparing hierarchical models (e.g., Quintero & Lesaffre, 2018). The marginalised DIC is computed by MCMC sampling. The calculation of the marginal DIC (mDIC) is described below.

3.4.3.2 Calculation of Marginal DIC

The marginal deviance is defined as

$$\mathcal{D}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = -2 \log \mathcal{L}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2; \boldsymbol{y}) \tag{3.4.1}$$

where $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$ are the second-level model parameters. $\mathcal{L}(\cdot)$ represents the marginal likelihood function. In the context of our generalised modelling framework, the marginal likelihood is given by

$$\mathcal{L}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2; \boldsymbol{y}) = \int_{\boldsymbol{\Theta}, \boldsymbol{\Delta}} \cdots \int \prod_{i=1}^{N} \prod_{j=1}^{J} \Big\{ p(y_{ij}; \Theta_i, \Delta_j) g_1(\Theta_i | \boldsymbol{\nu}_1) g2(\Delta | \boldsymbol{\nu}_2) \, d\Theta_i \, d\Delta_j \Big\}. \tag{3.4.2}$$

This marginal likelihood leads to the second-level marginalised DIC since the first-level parameters (i.e. Θ_i's and Δ_j's) are all integrated out.

According to Spiegelhalter et al. (2002), the DIC is defined as the difference between doubled posterior expected deviance

$$\text{DIC} = 2\overline{\mathcal{D}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2)} - \mathcal{D}(\widehat{\boldsymbol{\nu}_1}, \widehat{\boldsymbol{\nu}_2}). \tag{3.4.3}$$

The posterior mean of the deviance is expressed as

$$\overline{\mathcal{D}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2)} = \frac{1}{T}\sum_{t=1}^{T}(-2)\sum_{i=1}^{N}\sum_{j=1}^{J}\log\left\{\frac{1}{K}\sum_{k=1}^{K} p(y_{ij}|\boldsymbol{\nu}_1^{(t)}, \boldsymbol{\nu}_2^{(t)}, \Theta_{i,rep}^{(t),k}, \Delta_{j,rep}^{(t),k})\right\}. \tag{3.4.4}$$

K denotes the number of replicates at each post-burnin MCMC iteration $t = m + 1, \ldots, T$, with m being the number of MCMC iterations till the burnin. K is typical as large as 1,000 to reduce the sampling error. $\Theta_{i,rep}^{(t),k}$ and $\Delta_{j,rep}^{(t),k}$ for $k = 1, \ldots, K$ are posterior samples drawn K times at each post-burnin MCMC iteration t.

Next, we calculate the point estimate of the deviance by evaluating Equation (3.4.1) at posterior mean estimates:

$$\mathcal{D}(\widehat{\boldsymbol{\nu}_1}, \widehat{\boldsymbol{\nu}_2}) = -2\log\mathcal{L}(\widehat{\boldsymbol{\nu}_1}, \widehat{\boldsymbol{\nu}_2}; \boldsymbol{y}), \tag{3.4.5}$$

where $\widehat{\boldsymbol{\nu}_1}$ and $\widehat{\boldsymbol{\nu}_2}$ represent posterior means computed from the MCMC sampler after convergence. The mDIC is then calculated based on its definition given by Equation (3.4.3).

It is worth noting that the above interpretations of model comparison results are obtained under the assumption that the two-way outlier detection model is correctly specified and the two-way outliers in the model correspond to cheaters and compromised items.

3.4.3.3 Remarks

While computationally convenient, the DIC suffers from several caveats, including the lack of model selection consistency and the possibility of selecting overfitted models (Spiegelhalter, Best, Carlin, & Van der Linde, 2014). Therefore, we note that the results based on the DIC need to be interpreted with caution in practice.

In future research, other model comparison criteria will be investigated and compared with the DIC, including the Bayes factor (Kass & Raftery, 1995) and the Bayesian information criterion (BIC; Schwarz, 1978) that approximates the logarithm of the Bayes factor. The BIC is well-known to the full Bayesian framework. It is calculated based on the likelihood with a penalty term as a measure of model complexity, which is determined by the effective number of model parameters. The calculation of the effective number of model parameters in the hierarchical modelling context is found to be difficult, however, since the presence of priors constrains the effective dimension of the parameter space, according to the work by Entink, Fox, and van der Linden (2009); Fox (2010). The Bayes factor may be theoretically more attractive because it yields consistent model selection under suitable regularity conditions. However, its computation tends to be less straightforward than the DIC. Algorithms remain to be developed for computation under the current modelling framework.

3.5 Case Study: Licensure Test Data

We apply the proposed method to a dataset from a computer-based non-adaptive licensure test. The test is designed and operated under the Rasch model, which is consistent with the proposed baseline item-response submodel. This dataset has also been analysed for the detection of cheating in several chapters of Cizek and Wollack (2017) and journal articles including Sinharay (2017a) and Sinharay (2017b). We point out that the methods in these analyses require prior information about the items' statuses. For example, Sinharay (2017a) and Sinharay (2017b) require to

know the compromised items a priori and focus on the detection of cheaters. Unlike these existing analyses, we focus on the simultaneous detection of cheaters and compromised items, without requiring such prior knowledge.

The test contains 170 binary-scored items ($J = 170$), for which test takers' item responses and response times are available. The dataset is preprocessed by removing 12 test takers with zero response time in one or multiple items, which is believed to be a data recording error. This leads to a final dataset containing 1,624 test takers ($N = 1,624$). The testing program flagged 41 among these 1,624 test takers as likely cheaters, through a combination of statistical analysis and a careful investigative process which brought in other pieces of information. By a similar investigation process and data forensics, the testing program also believed that 64 among the 170 items were compromised. The identity of the testing program remains confidential to protect the test takers. We were asked to respect the desire of the program to remain anonymous. Therefore, we are unable to get into detail about how the testing program flagged the individuals or items.

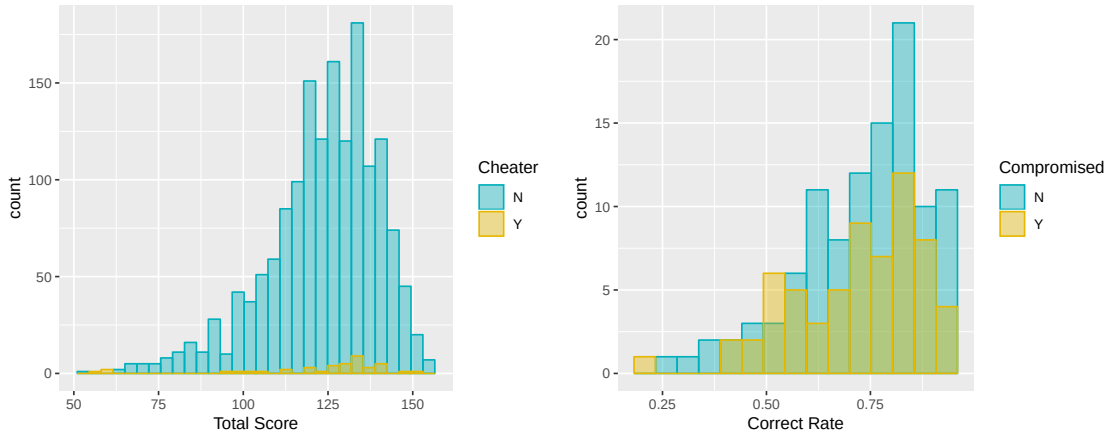
The labels of test takers and items will be used as partial truth for validating our data analysis results, but the proposed models do not rely on these labels at all. It is worth noting that these labels are not the ground truth and it is possible that there were test takers and items which ought to have been flagged but were not (Chapter 1, Cizek & Wollack, 2017). It is believed that the given labels are of good quality so that they can be used for the evaluation of detection methods. On the other hand, evaluation criteria based on these labels are not perfect, due to possible labelling errors.

The purpose of this analysis is two-fold. First, it is used to show the effectiveness of the proposed method, through a comparison between our results and the partial truth given by the testing program. Second, it is used to demonstrate the use of the proposed method in real tests, which may be of interest to practitioners.

3.5.1 Descriptive Analysis

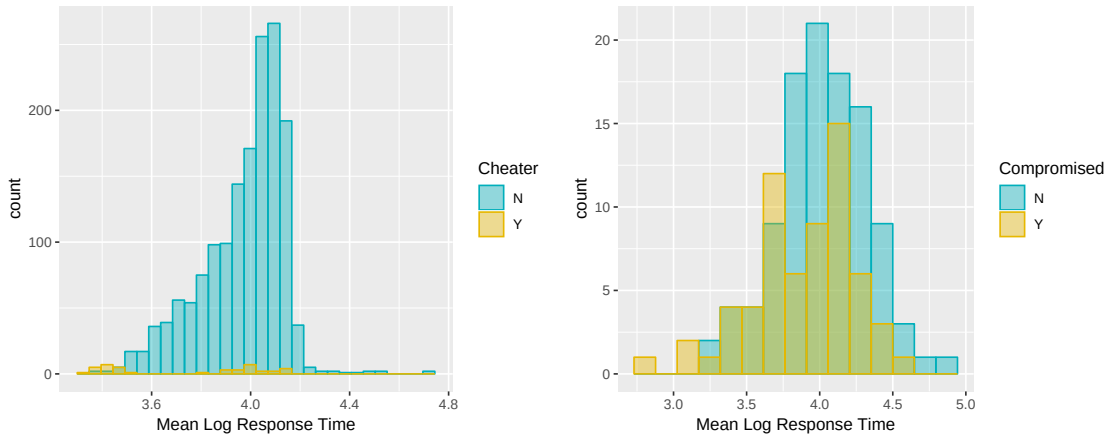
We start with a descriptive analysis to give an overview of the dataset. Panel (a) of Figure 3.5.1 shows the histogram of test takers' total scores by the testing program's cheating labels. Similarly, Panel (b) of Figure 3.5.1 gives the histogram of items' correct rates by the testing program's compromisation labels. Similarly, the two panels of Figure 3.5.2 show the histograms of the mean response time in the logarithm scale for test takers and items, respectively.

From these plots, it is not difficult to see that the corresponding summary statistics do not have much information about the labels on the test takers and items. In fact, the area under the curves (AUC) of the corresponding receiver operating characteristic (ROC) curves are 55.2% and 71.7% for the classification of the cheating labels based on the total score and mean log-time, respectively. Similarly, the corresponding AUCs for the classification of items are 52.4% and 60.6%, respectively. As we will see in the sequel, the proposed method substantially improves upon these benchmarks.



(a) **(b)**

Figure 3.5.1: *Licensure Test Data: Descriptive analysis. Panel (a): Histogram of test takers' total scores by the testing program's cheating labels. Panel (b): Histogram of items' correct rates by the testing program's labels of compromise status.*



(a) (b)

Figure 3.5.2: *Licensure Test Data: Descriptive analysis. Panel (a): Histogram of test takers' mean log response time by the testing program's labels of cheating. Panel (b): Histogram of mean log-time spent on items by the testing program's labels of compromise status.*

3.5.2 Detection based on Item Response Data

We start with analysing item responses using the reduced model (i.e. only analyse item responses and not time responses). Using the algorithm given in Appendix C.1., three MCMC chains were run with random starting points. Their convergence was assessed by trace plots and the Gelman and Rubin (GR) diagnostic statistic (Gelman & Rubin, 1992). The Gelman-Rubin R statistics applied to the parameters which are not person- or item-specific (see Table 3.5.2 for the list of these parameters) are below 1.20, suggesting that the chains converged to their equilibrium distributions after 10,000 iterations.

The inference is drawn based on 24,000 posterior samples from the three converged chains, where each contributes 8,000 samples. We first compare the fitted model with its null version using the DIC measure described in Section 3.4.3, to answer the question "does our item response data show evidence of cheating?". Recall that $\pi_1 = \pi_2 = 0$ in the null model, meaning that there are no cheaters or compromised items. The DIC value for the null model is also based on 24,000 posterior samples from an MCMC algorithm. The DIC values for the proposed and the null models are 138,282.6 and 218,308.4, respectively. The smaller DIC for the proposed model

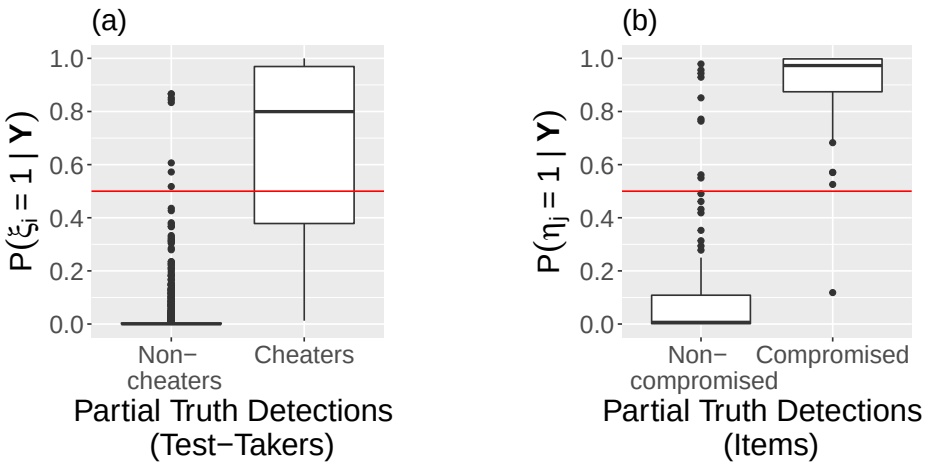


Figure 3.5.3: *Licensure Test Data: Box plots of posterior means of latent indicators from the reduced model for item responses. Panel (a): Box plots of the posterior means of ξ_i for the cheating and non-cheating groups (defined by the testing program). Panel (b): Box plots of the posterior means of η_j for the compromised and non-compromised items (defined by the testing program).*

suggests that item preknowledge is likely to exist among the test takers.

We then examine the classification results. Panel (a) of Figure 3.5.3 gives the box plots of the posterior means of ξ_i for the cheating and non-cheating groups (defined by the testing program), respectively. As we can see, the posterior means of ξ_i for the cheating group tend to be close to 1 and those for the non-cheating group tend to be close to 0, with some exceptions. Panel (b) of Figure 3.5.3 gives box plots of the posterior means of η_j for the compromised and non-compromised items (defined by the testing program), respectively. Similarly, the posterior means of η_j for the compromised items tend to be close to 1 and those for the non-compromised items tend to be close to 0. The corresponding ROC curves for the classification of the labels for cheaters and compromised items by the posterior means of ξ_i/η_j are presented in Figure 3.5.4. The AUCs for these two ROC curves are 0.868 and 0.836, respectively. They are substantially larger than the ones given by the summary statistics discussed in Section 3.5.1. We remark that these results on the detection accuracy should be interpreted with caution, due to possible labelling errors.

Now consider a yes or no question about multiple hypotheses of whether each person (or item) is flagged. We previously flagged all test takers who have a posterior

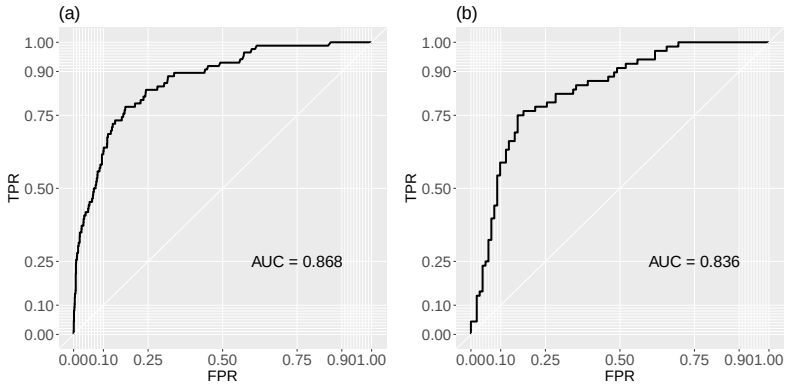


Figure 3.5.4: *Licensure Test Data: ROC curves for classification under the reduced model for item responses. Panel (a): ROC curve for the classification of cheaters (labelled by the testing program) by the posterior means of ξ_i. Panel (b): ROC curve for the classification of compromised items (labelled by the testing program) by the posterior means of η_j. The x- and y-axes of a ROC curve give the true positive rate (TPR) and false positive rate (FPR) for classification, respectively.*

probability of being involved in cheating greater than 0.5. In practice, we could afford to flag as many test takers as we can, but we need a principled approach to decide which test takers are worth flagging based on the information from the entire set of individuals rather than the information of each individual. The problem of hypothesis testing appears whenever we are attempting to classify individuals (or items) as potential cheaters (or compromised items). To solve the multiple testing issue, we apply the Bayesian decision framework proposed in Section 3.3.2 to false discovery rate control, a statistical procedure for dealing with multiple testing.

Moreover, Panel (a) of Figure 3.5.5 shows the local FDR and the local FNR as functions of the number of detections, respectively, when applying the proposed compound decision rule to test takers. As we can see, as the number of detections increases, the local FDR increases and the local FNR decreases. The same plot for items is given in Panel (b) of Figure 3.5.5. Specifically, the numbers of detections under different thresholds are given in Table 3.5.1, where we control local FDR for test takers and control local FNR for items. Again, we remark that the validity of the detection results depends on the extent to which our model assumptions hold.

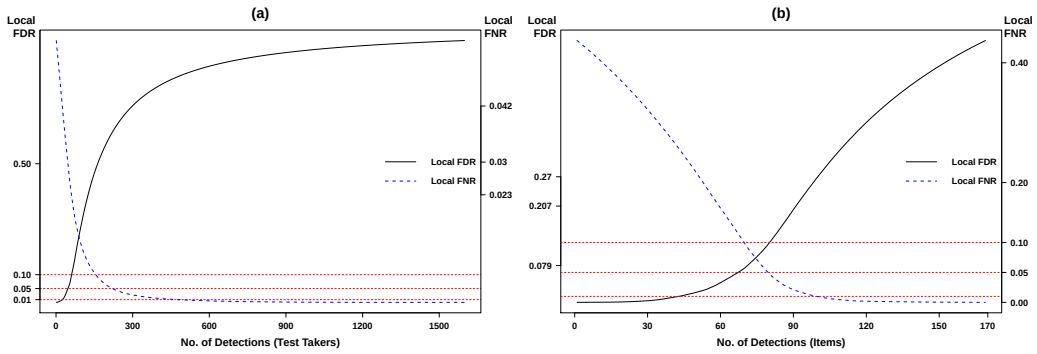


Figure 3.5.5: *Licensure Test Data: Detections based on the reduced model for item responses: The local FDR (represented by black solid curves) and the local FNR (represented by blue dashed curves) as functions of the number of detections.*

	1%	5%	10%
Test takers	25	46	61
Items	100	91	71

Table 3.5.1: *Licensure Test Data: The number of detections based on the reduced model for item response data. The first row shows the numbers of detections for test takers when controlling the corresponding local FDR at 1%, 5%, and 10% levels, respectively. The second row shows the numbers of detections for items when controlling the corresponding local FNR at 1%, 5%, and 10% levels, respectively.*

Therefore, we suggest treating such detection results as initial screening results, rather than as the final decisions.

Finally, posterior means and 95% credible intervals for the global parameters are presented in Table 3.5.2, where the global parameters refer to the parameters that are not person-specific or item-specific. In particular, the posterior mean of the proportion of cheaters is 2.8%, with a 95% credible interval (2.0%, 3.6%). This estimate is close to the proportion of 2.5% calculated based on the cheating labels from the testing program. The posterior mean of the proportion of compromised items is 40.1%, with a 95% credible interval (38.7%, 43.3%). This estimate is close to, but slightly higher than, the proportion of 37.6% given by the testing program. It may be the case that the testing program missed several compromised items during its labelling process. Furthermore, the posterior mean of δ is 0.895. That is, the odds ratio of correctly answering a compromised item is about $\exp(0.895) = 2.447$ when comparing a cheater and a non-cheater with the same ability level. Again, we point out that these interpretations depend on the extent to which our model holds

	σ_{11}	π_1	π_2	ω_{11}	μ_1	δ
EAP	0.285	0.028	0.401	0.685	-1.004	0.895
95% CI	$(0.261, 0.319)$	$(0.020, 0.036)$	$(0.387, 0.433)$	$(0.669, 0.854)$	$(-1.237, -0.912)$	$(0.758, 0.959)$

Table 3.5.2: *Licensure Test Data: Parameter estimation based on posterior samples from the reduced model for item responses. The row labelled "EAP" shows the posterior means of the global parameters, where EAP represents the Expected A Posteriori, and the row labelled "95% CI" provides the corresponding 95% credible intervals.*

and thus should be taken with caution.

3.5.3 Detection based on Item Responses & Response Times

We continue the modelling process by incorporating information from response times. The full model is applied to the dataset consisting of both item responses and response times. Three MCMC chains were used to fit the model. According to the GR statistics applied to the global parameters, the chains converged to their equilibrium distributions after 18,000 iterations.

The inference is drawn based on 24,000 posterior samples from the three chains after convergence. We compare this model with its null version by DIC, to answer the question "does item response time information help detect cheating?" Recall that these two models are the same, except that the response-time drift parameter $\gamma = 0$ in the null model. The DIC values for the proposed full model and its null version are 176,935.2 and 214,201.3, respectively. The smaller DIC value for the proposed full model suggests that response times contain substantial information about the cheating indicators.

The classification results are similar to those based only on item responses, and thus some plots shown above are omitted here. In particular, the ROC curves based on the posterior means of ξ_i and η_j have AUCs of 0.892 and 0.867, respectively, where these AUC values are slightly higher than those from the reduced model. In addition, the numbers of detections for test takers and items are shown in Table 3.5.3, where we still control local FDR for test takers and control local FNR for items. Comparing the results in Tables 3.5.1 and 3.5.3, generally more detections

	1%	5%	10%
Test takers	26	47	65
Items	101	89	74

Table 3.5.3: *Licensure Test Data: Detection based on the full model for item responses and response times. The first row shows the numbers of detections for test takers when controlling the corresponding local FDR at 1%, 5%, and 10% levels, respectively. The second row shows the numbers of detections for items when controlling the corresponding local FNR at 1%, 5%, and 10% levels, respectively.*

tend to be made under the full model. This is likely due to the fact that the posterior distributions tend to be more concentrated under the full model as it utilises more information.

Posterior means and 95% credible intervals for the global parameters are given in Table 3.5.4. Comparing Tables 3.5.2 and 3.5.4, we find that the estimates of the common parameters shared by the two models are close to each other. In particular, the 95% credit intervals overlap for each parameter. In addition, based on the posterior mean of Σ, the correlation between the ability and speed factors is as high as 0.410. This result indicates that test takers with higher abilities tend to answer the items faster. Such a high correlation between the two factors is not uncommon for high-stake tests. For example, C. Wang, Chang, and Douglas (2013) report a similar level of correlation between the ability and speed factors in a high-stake computerised adaptive test, under a similar Bayesian hierarchical model but without a cheating component. The estimated correlation between the two item-specific parameters is 0.237. This positive correlation suggests that solving more difficult items tends to take more time, which is consistent with our intuition.

3.6 Simulation Study

We now present two simulation studies for evaluating the finite-sample performance of the proposed method. The first study focuses on settings where our model is correctly specified, and the second study investigates the sensitivity of the proposed method under various forms of model misspecification.

	σ_{11}	π_1	π_2	ω_{11}	μ_1	δ
EAP	0.289	0.027	0.410	0.699	-0.867	0.807
95% CI	(0.259, 0.298)	(0.022,0.036)	(0.365, 0.432)	(0.626, 0.789)	$(-0.993, -0.795)$	(0.732, 0.852)

	σ_{22}	σ_{12}	ω_{22}	ω_{12}	μ_2	γ
EAP	0.248	0.110	0.397	0.125	-0.472	0.620
95% CI	(0.213, 0.285)	(0.986, 0.139)	(0.334, 0.427)	(0.082, 0.132)	$(-0.879, -0.291)$	(0.451, 0.907)

	κ
EAP	0.802
95% CI	(0.589, 1.037)

Table 3.5.4: *Licensure Test Data: Parameter estimation based on posterior samples from the full model for item responses and response times. The row labelled "EAP" shows the posterior means of the global parameters and the row labelled "95% CI" provides the corresponding 95% credible intervals.*

3.6.1 Study I

3.6.1.1 Settings

We consider simulation settings that mimic real educational tests. Specifically, we consider two settings for the sample size N and item size J, (1) $N = 2,000, J = 200$, and (2) $N = 4,000, J = 400$. This leads to two different settings, where the detection is expected to be more accurate under the second setting given its larger sample and item sizes. In what follows, these two settings are referred to as S1 and S2, respectively.

For each setting, we generate 50 independent datasets under the full model, with the global parameters fixed across the datasets. The proportion parameters π_1 and π_2 are set to 10% and 40%, respectively, the drift parameters δ and γ are both set to be 1.2, and the rest of the global parameters are set to be the same as the posterior means in Table 3.5.4 from the real data analysis above. For each dataset, we apply both the reduced model for item responses and the full model for item responses and response times. An additional simulation study is presented in Appendix D that shares a similar setting with the current study, except that the item size J is set to mimic educational tests with a smaller number of items. More specifically, this additional study considers two settings for N and J: (S1) $N = 2,000, J = 50$,

	Test taker				Item			
	S1		S2		S1		S2	
AUC	Reduced	Full	Reduced	Full	Reduced	Full	Reduced	Full
25%	0.953	0.954	0.951	0.959	0.951	0.959	0.951	0.959
50%	0.981	0.983	0.984	0.987	0.976	0.980	0.979	0.981
75%	0.990	0.994	0.993	0.993	0.992	0.994	0.997	0.996

Table 3.6.1: *Simulation Study I: Overall classification performance based on the posterior means of the person-specific latent indicator ξ_i and the item-specific latent indicator η_j. For each model, each setting, and each target (person/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 50 independent datasets.*

and (S2) $N = 4,000, J = 100$, and the rest of the settings remain the same. Similar results are observed in this additional study as those below from Study I.

3.6.1.2 Results

The analysis is conducted using our parallel tempering MCMC algorithm. For each dataset, we run 10,000 iterations, with the first 3,000 iterations as the burn-in. The results are based on the posterior samples from the last 7,000 iterations.

We first examine the classification results. For each model and each simulated dataset, we classify the test takers based on the posterior means of ξ_i and evaluate the classification based on the AUC value of the corresponding ROC curve. A larger AUC value implies higher classification accuracy. Similarly, the classification of the items is based on the posterior means of η_j and the accuracy is measured by the corresponding AUC value. These results are shown in Table 3.6.1. It can be observed that the classification is slightly more accurate under setting S2, due to the increased sample and item sizes. Moreover, the AUC values given by the full model tend to be slightly larger than those from the reduced model, thanks to the additional information from response times.

We further evaluate the proposed compound decision rules. For each dataset, we control local FDR and local FNR at levels 1%, 5%, and 10% for test takers and items, respectively. We evaluate each decision rule by examining the resulting FDP and FNP; see Section 3.3 for the definitions of FDP and FNP. The results are given in Tables 3.6.2 and 3.6.3 for the classifications of test takers and items, respectively.

	S1						S2					
	Reduced			Full			Reduced			Full		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.038	0.088	0.007	0.037	0.086	0.004	0.026	0.067	0.004	0.025	0.072
50%	0.012	0.048	0.091	0.011	0.048	0.092	0.007	0.031	0.079	0.006	0.029	0.083
75%	0.016	0.052	0.099	0.015	0.056	0.096	0.009	0.039	0.092	0.007	0.033	0.088

Table 3.6.2: *Simulation Study I: Local FDR control for individuals. For each model, each setting, and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 50 independent datasets.*

	S1						S2					
	Reduced			Full			Reduced			Full		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.049	0.091	0.010	0.045	0.089	0.006	0.024	0.063	0.007	0.025	0.065
50%	0.011	0.051	0.098	0.012	0.048	0.092	0.010	0.031	0.079	0.011	0.033	0.077
75%	0.013	0.059	0.104	0.012	0.057	0.096	0.015	0.039	0.091	0.012	0.037	0.089

Table 3.6.3: *Simulation Study I: Local FNR control for items. For each model, each setting, and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 50 independent datasets.*

According to these tables, the FDP is well-controlled for test takers and so is the FNP for items.

Finally, we show the results on the estimation of the global parameters, as these parameters have substantive interpretations in cheating detection. Specifically, we focus on the posterior mean estimator, for which bias and variance are estimated based on the results from 50 independent replications. These results are presented in Table 3.6.4. The bias, in general, tends to be close to zero for all the global parameters from both models and both settings. In addition, the estimation tends to be more accurate under setting S2, due to the increased sample and item sizes.

3.6.2 Study II

3.6.2.1 Settings

In this study, we investigate the sensitivity of the proposed method under several forms of model misspecification. For the clarity of simulation settings, we focus on the misspecification of the item-response submodel. That is, we generate item-response data from a misspecified model and then apply our reduced model to

S1	Reduced model						S2	Reduced model					
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ		π_1	π_2	σ_{11}	μ_1	ω_{11}	δ
Bias	0.13	0.09	-0.15	-0.19	-0.11	0.13	Bias	0.09	0.05	0.11	-0.08	-0.08	0.09
Variance	0.14	0.12	0.37	0.23	0.27	0.31	Variance	0.11	0.15	0.29	0.25	0.21	0.27

S1	Full model											
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ
Bias	0.11	-0.08	0.07	-0.24	-0.08	0.08	-0.12	-0.04	0.07	0.14	0.09	-0.16
Variance	0.16	0.11	0.34	0.19	0.32	0.35	0.09	0.07	0.12	0.13	0.08	0.77

S2	Full model											
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ
Bias	0.07	-0.03	0.11	-0.18	-0.05	0.11	-0.08	-0.06	0.09	0.15	0.05	-0.11
Variance	0.12	0.08	0.33	0.21	0.34	0.31	0.05	0.09	0.08	0.10	0.12	0.63

Table 3.6.4: *Simulation Study I: Accuracy of the posterior mean estimator of the global parameters. The bias and variance for the posterior mean estimator are calculated based on the 50 replications.*

Setting	Misspecification	N	J
S3	(1)	2,000	200
S4	(2)	2,000	200
S5	(3)	2,000	200
S6	(1)	4,000	400
S7	(2)	4,000	400
S8	(3)	4,000	400

Table 3.6.5: *Study II: Six simulation settings, where (1)-(3) correspond to three forms of model misspecification, including the misspecification of (1) the baseline model, (2) the relationship between ξ_i and θ_i, and (3) the common drift parameter.*

classify the test takers and items. The overall classification performance, as well as the performance of the proposed compound decision rules, is evaluated. We focus on three forms of model misspecification whose details are discussed below, including the misspecification of (1) the baseline model, (2) the relationship between ξ_i and θ_i, and (3) the common drift parameter. The three forms of model misspecification, together with two settings for N and J as in Study I, lead to six different settings as summarised in Table 4.5.1. For each setting, except for the misspecified part, the global parameters are set the same as those in Study I. For each setting, 50 independent datasets are generated.

We now discuss the three forms of model misspecification in detail. For the baseline model, we replace the Rasch model by the two-parameter logistic model, an IRT model that is widely used in educational testing. That is, the following item-response

submodel is assumed

$$P(Y_{ij} = 1|\Theta_i, \Delta_j, \alpha_j, \delta) = \frac{\exp(\alpha_j(\theta_i - \beta_j) + \xi_i\eta_j\delta)}{1 + \exp(\alpha_j(\theta_i - \beta_j) + \xi_i\eta_j\delta)},$$

where α_j is known as the discrimination parameter. Note that the proposed model can be viewed as a special case where $\alpha_j = 1$ for all j. In the misspecified model, we generate the discrimination parameters α_j from a uniform distribution $U[1, 1.5]$.

In the proposed model, θ_i and ξ_i are assumed to be independent, meaning that whether a test taker cheats or not is independent of his/her ability. This assumption may not hold and it is likely that these two variables are negatively associated, i.e., test takers with lower ability are more likely to cheat in an exam. To mimic this situation, we generate (θ_i, ξ_i) jointly from a Gaussian copula. That is, we first generate (θ_i, ξ_i^*) from a bi-variate normal distribution, with mean vector $(-0.867, 0)^\top$ and covariance matrix $((0.289, -0.134)^\top, (-0.134, 1)^\top)$. Under this bivariate normal distribution, the correlation between θ_i and ξ_i^* is -0.25. We then let $\xi_i = 1_{\{\xi_i^* \geqslant z_{0.9}\}}$, which is obtained by truncating ξ_i^* at $z_{0.9}$, the 90% quantile of the standard normal distribution, so that $P(\xi_i = 1) = 0.1$. Under this Gaussian copula model, the marginal distributions of θ_i and ξ_i remain the same as those in Study I, while a negative association is introduced between the two variables.

For model parsimony, it is also assumed in the proposed model that the drift parameter δ is common across all the test takers and items. This assumption may not hold in practice. Therefore, in this misspecified model, instead of using a constant drift, we assume the drift parameter to be both item- and person-specific. That is, we assume

$$P(Y_{ij} = 1|\Theta_i, \Delta_j, \delta_{ij}) = \frac{\exp(\theta_i - \beta_j + \xi_i\eta_j\delta_{ij})}{1 + \exp(\theta_i - \beta_j + \xi_i\eta_j\delta_{ij})},$$

where the drift parameters δ_{ij} are generated i.i.d. from a uniform distribution $U[1, 1.5]$.

	Test takers						Item					
AUC	S_3	S_4	S_5	S_6	S_7	S_8	S_3	S_4	S_5	S_6	S_7	S_8
25%	0.904	0.889	0.945	0.912	0.922	0.947	0.921	0.903	0.949	0.947	0.937	0.945
50%	0.965	0.927	0.975	0.969	0.950	0.981	0.965	0.936	0.964	0.962	0.951	0.976
75%	0.983	0.989	0.994	0.980	0.984	0.993	0.981	0.971	0.988	0.985	0.984	0.991

Table 3.6.6: *Simulation Study II: Overall classification performance based on the posterior means of ξ_i and η_j. For each model, each setting, and each target (test taker/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 50 independent datasets.*

3.6.2.2 Results

We evaluate the proposed method under the six settings above. Similar to Study I, we evaluate the overall classification performance by AUC and the performance of the compound decision rules by the corresponding FDP and FNP. The results are given in Tables 3.6.6 through 3.6.8. Specifically, the AUC values in Table 3.6.6 are comparable to those from the correctly specified model in Table 3.6.1, though the AUCs from settings S3, S4, S6, and S7 are slightly smaller. As further shown in Tables 3.6.7 and 3.6.8, the compound decision rules tend to control the corresponding FDP and FNP under the targeted levels, except when the target level is 1%. That is, when controlling the local FDR and local FNR to be below 1% for test takers and items, respectively, the resulting FDP and FNP tend to exceed the targeted level under all six settings. This is likely due to the fact that the posterior probabilities cannot be accurately obtained when they are close to 0 or 1, under model misspecification.

Overall, the proposed method is reasonably robust against several forms of model specification, though the performance may be slightly affected. However, under potential model misspecification, the method should be used with caution if we aim to control local FDR or local FNR to be below a very small threshold (e.g., 1%).

	S3			S4			S5		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.013	0.031	0.061	0.018	0.042	0.059	0.012	0.028	0.067
50%	0.015	0.039	0.072	0.023	0.048	0.065	0.026	0.039	0.072
75%	0.019	0.046	0.074	0.026	0.059	0.070	0.029	0.043	0.079

	S6			S7			S8		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.004	0.029	0.075	0.014	0.036	0.059	0.009	0.025	0.061
50%	0.007	0.041	0.089	0.017	0.043	0.066	0.014	0.041	0.064
75%	0.012	0.045	0.093	0.024	0.047	0.072	0.023	0.052	0.075

Table 3.6.7: *Study II: Local FDR control for individuals. For each model, each setting, and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 50 independent datasets.*

	S3			S4			S5		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.007	0.021	0.061	0.019	0.032	0.078	0.015	0.039	0.072
50%	0.014	0.032	0.074	0.022	0.037	0.073	0.028	0.043	0.081
75%	0.017	0.038	0.076	0.024	0.041	0.085	0.030	0.047	0.087

	S6			S7			S8		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.024	0.045	0.012	0.031	0.072	0.014	0.036	0.064
50%	0.011	0.025	0.059	0.015	0.036	0.079	0.026	0.046	0.073
75%	0.014	0.032	0.082	0.025	0.041	0.086	0.029	0.054	0.083

Table 3.6.8: *Simulation Study II: Local FNR control for items. For each model, each setting, and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 50 independent datasets.*

3.7 Concluding Remarks

In this chapter, we propose a Bayesian hierarchical model for detecting two-way outliers due to latent DIF in item-response-type multivariate data. The proposed method is able to simultaneously detect outlying individuals and items that deviate from a given baseline model. Furthermore, a compound decision theory is proposed for the detection of two-way outliers under a Bayesian decision framework. Statistical inference is carried out under a fully Bayesian framework for which a parallel tempering MCMC algorithm is developed. The algorithm presented in Appendix C.1 is a powerful tool for improving the sampling of multi-modal posterior distributions. It is relatively easy to implement, as the algorithm can be performed on multiple processors to improve computational efficiency. The algorithm can fit a dataset consisting of 4,000 persons and 400 items in 2 minutes 27 seconds on an Intel machine (2.2GHz Intel Core i7) with 8 threads.

The proposed two-way outlier detection model is largely motivated by, and applied to, the simultaneous detection of test takers who benefit from item preknowledge and compromised items in educational tests. The proposed method does not rely on prior knowledge about either the test takers with item preknowledge or the compromised items, and thus is directly applicable to operational tests as a monitoring tool or more generally about outlying cases and items in other applications.

The proposed method is successfully applied to data from a licensure test which is known to suffer from item preknowledge. In this study, two models are applied, including the reduced model for item responses and the full model for item responses and response times. Both models accurately detect the potential cheaters and compromised items identified by the testing program, suggesting their usefulness in practice. In addition, the full model performs slightly better than the reduced one, suggesting that response-time information may help detect cheating. However, it should be noted that the labels provided by the testing program in this example are not the ground truth and thus the accuracy measures may be compromised.

The validity of the proposed method remains to be checked through applications to other educational tests. We note that a simple model, such as the one applied in the case study, may be preferable for the detection of cheating in educational tests, even though more general models are available as discussed in Section 3.2. This is because the numbers of test takers with preknowledge and compromised items are usually small in an educational test, which makes the effective sample sizes small for estimating parameters related to the outlier classes. In that case, a more complex model may lead to a high variance in the estimation, which further yields inaccurate classifications.

Limitations of the proposed method have been discussed in Section 3.2.1.4 as well as its robustness against model misspecification in Section 3.6.2. Another limitation is that it only models a specific type of cheating, i.e., preknowledge due to item leakage. It does not handle other types of cheating behaviours, such as copying others' answers, electronic transmission of data, hiring stand-ins, and bribing test administrators to correct one's answers. To investigate different types of cheating behaviours, different sources of information are needed and suitable statistical methods remain to be developed. For example, to detect copying behaviour, a statistical model is needed to characterise the similarity between the item responses from two test takers, possibly taking into account their response process information (e.g., response time), and seat locations in a test centre, etc. We leave these problems for future investigation.

Missing data are widely encountered in educational tests that may be informative for the detection of cheating, though not observed in our real data example. For an educational test with cheating test takers, the missingness of response likely depends on whether the test taker is cheating and whether the item is compromised. If many missing responses are observed, then the current framework should be extended by modelling the probabilities of responding. This problem is left for future development, for which ideas may be borrowed from latent variable models for non-ignorable missingness (e.g., Kuha, Katsikatsou, & Moustaki, 2018; O'Muircheartaigh

& Moustaki, 1999).

Besides the applications to cheating detection in educational tests, future research will be conducted to investigate the computation, model evaluation and comparison in other areas of application, such as voting behaviours and psychological measurement. Specifically, MCMC algorithms for the Bayesian inference of the proposed two-way classification model will be further explored. Although our parallel tempering algorithm works well for the current analysis, its performance will be evaluated under more settings, especially large-scale settings (larger numbers of individuals and items). In addition, other tempering methods, such as simulated tempering, can be explored. Moreover, goodness-of-fit issues and model selection will be further studied. In particular, the use of Bayes factors and BIC for comparing the proposed model with several relevant models will be investigated.

Chapter 4

Explanatory Two-way Outlier Detection Model

4.1 Introduction

We have proposed a method for detecting two-way outliers in multivariate data without any prior knowledge of the outlying status of individuals and items. The two-way outlier detection model proposed in the previous chapter incorporates a double mixture structure with an IRT-type model. In the absence of two-way outliers, the model reduces to an IRT model for fitting the standard item-response behaviour. In the presence of two-way outliers, the latent class component captures the effect of outliers while the IRT model component still captures the standard item-response behaviour.

The outlier detection has thus far been informed by item responses (and response times in the full model case), without relying on any information contained by external covariates for individuals and items. It is worth mentioning that in a wide range of studies in social research, education or behavioural science, covariate information is routinely gathered through questionnaires or institutional records in addition to item responses. Covariates often contain demographic and contextual characteristics such as gender, ethnicity, country of residence, and education level. The licence

test dataset previously used in Section 3.5, for example, contains external variables indicating one's attempt count, education background and item usage, to name a few. These external variables may provide more insights into the difference between compromised and uncompromised items, and between test takers with and without prior access to the compromised items. While the observed covariates may not be directly associated with item responses, they can indirectly affect the response probabilities through their relations with latent indicators of outlyingness. Moreover, they may also affect the distributions of the latent person-specific (e.g. person ability) and item-specific parameters (e.g. item difficulty). Therefore, it remains to be seen whether the inclusion of covariates would improve the classification of persons and items and hence inform the two-way outlier detection.

Another limitation of the two-way outlier detection model is that it assumes the independence between the latent indicators of outlyingness and the latent person and item parameters. This assumption may be too restrictive. In the example of detecting cheating due to item preknowledge, it is often the case that more challenging items are more likely to be leaked because they are often considered to be more beneficial to test takers with preknowledge in a sense that they serve to distinguish top candidates amongst those others. The dependence also goes for the person-specific latent indicator and latent trait. Research (Cizek & Wollack, 2016; Simha & Cullen, 2012) suggests that person ability and one's chance of cheating are related. Therefore, we have a practical reason to relax this assumption. The latent parameters can be viewed as latent covariates that are used to predict the latent class memberships of individuals and items along with observed covariates.

In this chapter, we extend the previously proposed two-way outlier detection model for item response data in an explanatory framework. The extension aims to incorporate covariates into the two-way outlier detection model through their relations with the latent class indicators as well as the distributions of latent person and item parameters. The assumption that the latent class indicators do not depend on the latent parameters is also relaxed within the explanatory framework.

Review of Covariate Inclusion in (Mixture) IRT Models

The two-way outlier detection model can be viewed as a double-mixture IRT model that reduces to an IRT-type model in absence of two-way outliers. In what follows we review research of relevance to covariate inclusion in IRT and mixture IRT contexts.

An IRT model in its basic form does not incorporate covariates, but extensions can be made to allow for covariates. Covariates can be included in IRT models (and mixture IRT models) in different ways depending upon whether they are person-specific or item-specific, whether they are observed or latent, and if they are observed covariates, whether they are directly associated with item responses (conditioning on the latent variables) or indirectly associated with item responses through affecting the distributions of latent parameters.

De Boeck and Wilson (2004) presented a doubly explanatory item response theory (EIRT) modelling framework incorporating the effects of observed and latent covariates on person and item sides with an IRT-type model. Wilson, De Boeck, and Carstensen (2008) applied the EIRT framework to Rasch and 2PL IRT models. This leads to a doubly explanatory model which allows person-specific observed and latent covariates, and item-specific observed covariates to be directly associated with item responses. It is worth noting that item parameters in this specification are treated as fixed effects as opposed to random effects or latent variables, as item difficulty is assumed to be perfectly predicted by item-specific observed covariates. Janssen, Schepers, and Peres (2004) relaxed this assumption and proposed an EIRT model to further account for random item variation. The full EIRT model is estimated using a Gibbs sampling method in a Bayesian setting.

The explanatory analysis has also been used outside the scope of the conventional IRT modelling framework, including mixture IRT and other mixture modelling frameworks. Recall that mixture IRT models are built upon the IRT models with latent class analysis and are known for their capacity to handle population heterogeneity and latent DIF (Cohen & Bolt, 2005). Covariates can be incorporated in

mixture IRT models through their relations with person- or item-specific latent class indicators, which is found to be useful in improving model parameter estimation and explaining members in the latent classes (Smit, Kelderman, & van der Flier, 1999).

Rost (1990) addressed a mixture Rasch model (MRM) in the case when person-specific covariates may be associated with an individual's latent class membership. The classification of individuals was substantially improved by relating latent class membership to external covariates. Cohen and Bolt (2005) extended the MRM to identify items exhibiting DIF across both observed and latent groups of individuals. Dai (2013) directly included person-specific covariates in a mixture Rasch model (MRM) under a logistic regression to estimate proportions of latent classes and hence more accurately recovered the latent classes and improved parameter estimation. Park, Xing, and Lee (2018) extended the explanatory framework to a cognitive diagnostic model (CDM) and proposed to integrate observed and latent predictors on both person and item levels into a CDM. Recall that a CDM consists of a latent class model component for allocating individuals to different skill profiles (also known as attributes) and an IRT model component for assessing the diagnostic efficiency of individuals based on their responses to measured items. In the proposed explanatory CDM, observed covariates are directly associated with the response probabilities and also indirectly associated with the responses by affecting the attributes.

The studies mentioned above, with the exception of Park et al. (2018), used a two-step approach to estimate explanatory mixture IRT models. According to the two-step approach, a mixture IRT model without involving covariates is estimated first to obtain estimates for latent variables, including person- or item specific- latent class memberships and latent parameters, and then the estimates for latent class memberships and latent parameters are regressed on covariates. However, the explanatory framework allows latent class membership and latent parameters to be estimated while at the same time accounting for their relationships with covariates. Smit, Kelderman, and van der Flier (2000) showed that the classification can benefit substantially from incorporating covariates during the estimation process of the

latent class membership. Park et al. (2018) also emphasised the value of simultaneously estimating latent variables and covariate effects within the explanatory CDM framework.

The study in this chapter builds upon the previous research on the explanatory analysis of IRT and mixture IRT models and differs from them in the sense that person and item covariates are included in a double-mixture IRT model to estimate their relations with the latent person and latent item parameters, and along with the latent parameters, to predict latent class memberships for persons and items. We anticipate that covariate inclusion will improve the detection of two-way outliers and be helpful in explaining the relations between latent variables and person and item characteristics. In addition, performances of the two-way outlier detection model with covariates and the two-way outlier detection model without covariates are compared to examine the possible advantages of including covariate information.

The rest of this chapter is structured as follows. In Section 4.2, we review the previously proposed two-way outlier detection model and propose to extend the model in an explanatory framework. In the following section 4.3, we specify the details of the Bayesian inference procedures and revisit compound decision rules and model comparison under the proposed explanatory two-way outlier detection framework. In Section 4.4, we revisit the licensure test data and use the proposed model for cheating detection. In order to see whether covariate inclusion and relaxation of the independence assumption would improve cheating detection, the classification performance under the current model is compared with that under the reduced model proposed in the previous chapter. We also look at the interpretation of the estimates for parameters that characterise the relationship between latent class memberships and covariates. Simulation studies are conducted in Section 4.5 to examine the stability of parameter estimation and classification under the proposed model in simulation settings with different sample sizes, item sizes and levels of the outlier effect. Furthermore, the classification and detection results under the proposed model are compared with those under its two submodels, in which covariate effects on per-

son ability, item difficulty, and latent class membership for persons and items are not fully accounted for, or not incorporated at all. Finally, key findings and future work are discussed in Section 4.6. The details for the parallel tempering MCMC algorithm can be found in Appendix C.2.

4.2 Model Setup

Recall that in Section 3.2, we propose a two-way outlier detection model for item response data (i.e. the reduced model). The item response data consist of binary outcomes $\boldsymbol{Y}_{N\times J}$ in response to J item from N persons. In this section, we review the reduced model, the assumptions that the model relies on, and why some model assumptions may be restrictive. The need for relaxing some assumptions, together with the necessity of covariate inclusion, leads to the proposal of the explanatory two-way outlier detection model.

4.2.1 Review of the Two-way Outlier Detection Model

The two-way outlier detection model relies upon an IRT model component as the baseline model for fitting standard item-response behaviour. The baseline model used in the previous chapter is the Rasch model, which connect the response from person i to item j with their ability level θ_i $(i = 1, \ldots, N)$ and item difficulty β_j $(j = 1, \ldots, J)$. To capture atypical item-response behaviour, a latent class model or a double mixture component is added to the baseline model component. The latent class component introduces two latent indicators of whether person i or item j is outlying, denoted as ξ_i and η_j, respectively. The inclusion of the latent class component essentially makes the problem of two-way outlier detection a two-way classification problem. The effect of two-way outliers on item responses is modelled by a drift parameter δ, which is assumed to be constant for all outlying individuals and items. The drift can be made person- or/and item-specific, non-positive or non-negative, depending on context.

The item response function (IRF) with an additional latent class component for capturing the effect of two-way outliers is given by Equation (3.2.1):

$$P(Y_{ij} = 1|\theta_i, \xi_i, \beta_j, \eta_j) = \frac{\exp(\theta_i - \beta_j + \xi_i\eta_j\delta)}{1 + \exp(\theta_i - \beta_j + \xi_i\eta_j\delta)}. \tag{4.2.1}$$

The model is applied to the detection of cheating due to item preknowledge in a licensure test in Section 3.5.2. In absence of cheating (i.e. $\xi_i = 0$ and $\eta_j = 0$), the model reduces to the baseline model for fitting standard item-response behaviour. In the presence of cheating (i.e. $\xi_i = 1$ and $\eta_j = 1$), test takers with preknowledge are more likely to give correct responses to compromised items. To reflect the potential boost in their response probability as a result of cheating, the drift parameter δ is constrained to be non-negative.

All parameters in the two-way outlier detection model are simultaneously estimated under a fully Bayesian framework. The hierarchical structure of the two-way outlier detection is shown in Figure 4.3.1a. The classifications of persons and items are made based on posterior probabilities of ξ_i and η given observed data: $P(\xi_i = 1|\boldsymbol{Y})$ and $P(\eta_j = 1|\boldsymbol{Y})$, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$.

There are several key assumptions are made for the two-way outlier detection model. First, item responses are conditionally independent given person- and item-specific latent parameters θ_i, β_j, ξ_i, and η_j. Second, for either a person or an item, the latent indicator of outlyingness (ξ_i or η_j) is assumed to be independent of the latent parameter (θ_i or β_j). The second assumption, however, may not be widely supported by empirical evidence. In the context of cheating detection, one's chance of cheating is believed to be associated with one's ability level. Both Cizek and Wollack (2016) and He, Meadows, and Black (2020) suggested that one's ability may be negatively associated with the chance of cheating on compromised items, meaning that less capable test takers are more likely to have the motivation to cheat on compromised items to get by. Cizek and Wollack (2016) also provided an alternative view that above-average examinees are more likely to get prior access to difficult items. The

potential relation between item leakage and item difficulty cannot be overlooked as well. Studies undertaken by Cizek and Wollack (2016) and Wagner-Menghin, Preusche, and Schmidts (2013) suggested that more challenging exam material is more likely to be leaked and meanwhile item preknowledge on easier items is more difficult to detect. Therefore, the independence assumption needs to be relaxed to account for the potential relationship between latent indicators of outlyingness and continuous latent parameters.

4.2.2 Proposed Model

To improve the classification and further inform the detection of two-way outliers, we propose to extend the two-way outlier detection model (4.2.1) in an explanatory framework, where latent indicators of outlyingness are allowed to be associated with covariates and latent predictors which include individual ability and item difficulty.

Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be a p-dimensional column vector of covariates specified for person i and $\boldsymbol{z}_j \in \mathbb{R}^q$ be a q-dimensional column vector containing covariates for item j. We assume that covariates are not associated with the item responses Y_{ij}'s directly, but through the latent indicators of outlying status, ξ_i's and η_j's.

The conditional relationships between latent indicators of outlyingness and covariates given continuous latent variables can then modelled by logistic regressions:

$$\begin{aligned} P(\xi_i = 1|\boldsymbol{x}_i, \theta_i) &= \frac{\exp[\boldsymbol{a}^\top \tilde{\boldsymbol{x}}_i + g_1(\theta_i, \boldsymbol{\lambda})]}{1 + \exp[\boldsymbol{a}^\top \tilde{\boldsymbol{x}}_i + g_1(\theta_i, \boldsymbol{\lambda})]} \\ P(\eta_j = 1|\boldsymbol{z}_j, \beta_j) &= \frac{\exp[\tilde{\boldsymbol{z}}_j^\top \boldsymbol{b} + g_2(\beta_j, \boldsymbol{\phi})]}{1 + \exp[\tilde{\boldsymbol{z}}_j^\top \boldsymbol{b} + g_2(\beta_j, \boldsymbol{\phi})]}, \end{aligned} \tag{4.2.2}$$

where $\tilde{\boldsymbol{x}}_i^\top = (1, \boldsymbol{x}_i^\top)$, $\tilde{\boldsymbol{z}}_j^\top = (1, \boldsymbol{z}_j^\top)$, $\boldsymbol{a} = (a_0, a_1, \ldots, a_p)$, and $\boldsymbol{b} = (b_0, b_1, \ldots, b_q)$. a_0 and b_0 denote intercepts and their values correspond to the proportions of outlying individuals and items in the data. $a_1, \ldots, a_p$, $b_1, \ldots, b_q$ are regression coefficients corresponding to observed predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$, respectively.

Functions g_1 and g_2 serve to flexibly determine the functional shapes of the rela-

tionships between the latent indicators and latent parameters θ_i and β_j. g_1 and g_2 can be represented using regression splines. Specifically, the regression spline of a latent predictor can be comprised of a linear combination of known basis functions and unknown regression parameters, λ and ϕ. A common choice for the basic functions is polynomial functions. In the application of cheating detection, if the average test takers are believed to be more likely to cheat than the struggling or top test takers, then g_1 takes a quadratic function of θ_i: $g_1(\theta_i, \boldsymbol{\lambda}) = \lambda_1\theta_i + \lambda_2\theta_i^2$, where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. If the chance of an item being compromised is positively correlated with item difficulty, it is not unreasonable to assume $g_2(\beta_j, \boldsymbol{\phi}) = \phi_1\beta_j$, where $\boldsymbol{\phi} = (\phi_1)$.

The explanatory model assumes that item responses are conditionally independent given person- and item-specific latent variables, θ_i, β_j, ξ_i, and η_j, while allowing for the dependence between the latent indicators and latent parameters, and between the latent parameters and covariates. The explanatory model still assumes that outlying individuals and items are not dominant in the data, which can be guaranteed by imposing weakly informative priors on the intercept parameters a_0 and b_0. Different from the previous models, the explanatory model allows the means of person- and item-specific latent parameters to depend on relevant covariates. By doing this, the distributions of person ability can vary across the demographic groups. More details are provided in the following Section 4.3.1.

4.3 Bayesian Inference and Compound Decision

The explanatory two-way outlier detection model (Equations 4.2.1 & 4.2.2) is estimated under a full Bayesian framework, where model parameters, including θ_i's, ξ_i's, β_j's, η_j's, δ, $\boldsymbol{a}$, $\boldsymbol{\lambda}$, $\boldsymbol{b}$, and $\boldsymbol{\phi}$ are all treated as random variables. A parallel tempering MCMC algorithm is applied to the explanatory model to sample model parameters from their joint posterior distribution. The details are presented in Appendix C.2.

Convergence is assessed based on trace plots and the Gelman-Rubin (GR) statistic

(Gelman & Rubin, 1992). When the trace plots show that the deviance of each chain is stabilised, and the juxtaposition of multiple chains with different starting points makes it clear that the convergence is reached, the GR statistic is computed for the global parameters which are not person- or item-specific. The GR statistic value is at around 1.01 for all the global parameters, suggesting these chains converged to their equilibrium distributions.

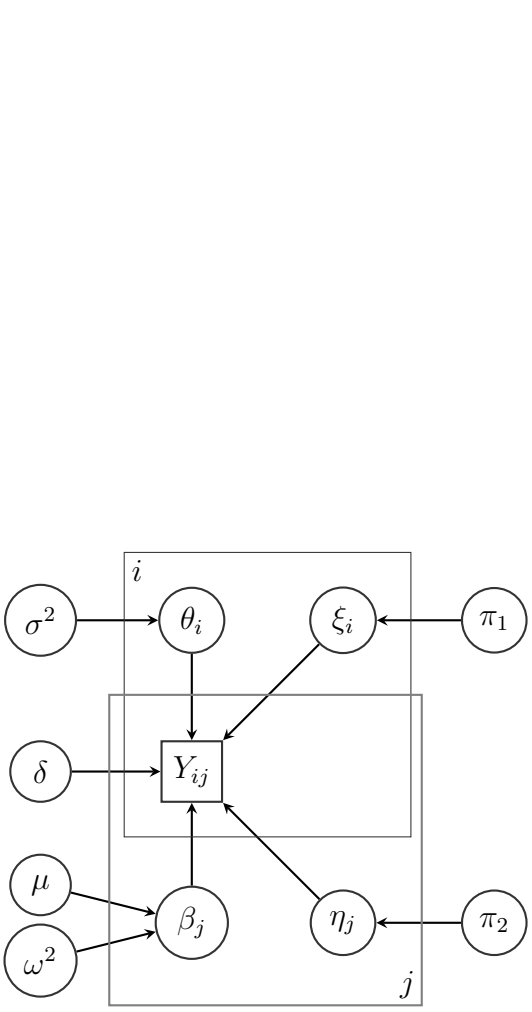
In this section, we first specify priors and hyperpriors and then briefly revisit the compound decision framework for the detection of two-way outliers under the explanatory model.

4.3.1 Hierarchical Model Specification

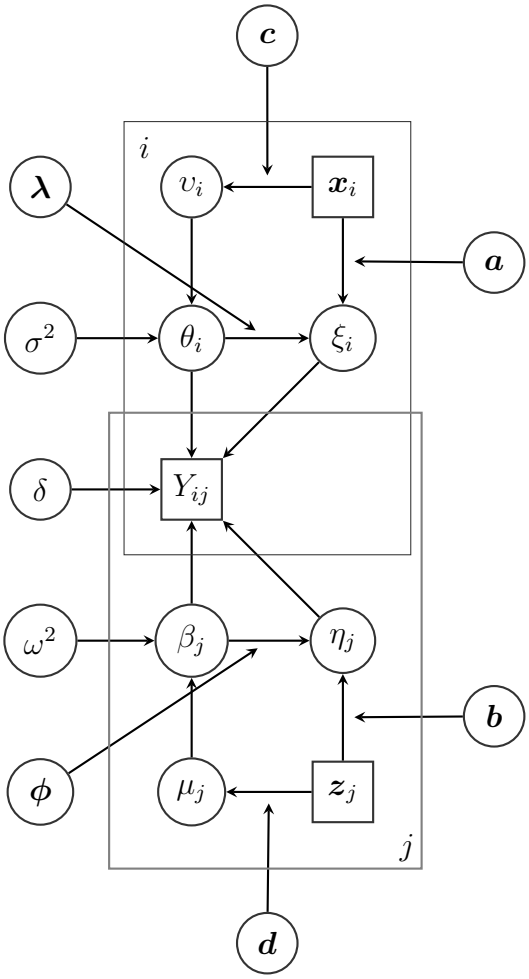
The hierarchical structure of the explanatory two-way outlier detection model is displayed in Figure 4.3.1b. Two outer plates represent persons and items and the inner plate represents item responses. We specify the priors and hyperpriors for the nodes in Figure 4.3.1b under a full Bayesian setting.

We start by specifying the prior distribution for person-specific parameters. The latent indicator, ξ_i, depends on the person ability parameter, θ_i. The conditional distribution of ξ_i given θ_i is defined by the model (4.2.2). The previous outlier detection model assigns a normal prior, denoted by $\mathrm{N}(0, \sigma^2)$, to θ_i. The normal prior is also independent of person-specific covariates. Under the explanatory framework, we can take into account the ways in which θ_i differs among groups of individuals or changes with observed covariates. Thus, the hierarchical structure of the explanatory model shown in Figure 4.3.1b includes the relation between the latent parameter and covariates for indicating individual or contextual characteristics.

The prior for θ_i is assumed to be normal and the prior mean is allowed to depend on person-level covariates $\boldsymbol{x}_i$: $\theta_i \sim \mathrm{N}(\upsilon_i, \sigma^2)$, where $\upsilon_i = \boldsymbol{c}^\top \tilde{\boldsymbol{x}}_i$, $\tilde{\boldsymbol{x}}_i^\top = (1, \boldsymbol{x}_i^\top)$, and parameters $\boldsymbol{c} = (c_0, c_1, \ldots, c_p)$. The prior mean of θ_i, for $i = 1, \ldots, N$, is subject to an identifiability constraint $\sum_i \boldsymbol{c}^\top \tilde{\boldsymbol{x}}_i = 0$ or $c_0 = -\boldsymbol{c}^\top \bar{\boldsymbol{x}}_i$. This can be achieved



(a) *Two-way Outlier Detection Model using item response data.*



(b) *Explanatory Two-way Outlier Detection Model using covariates and item response data.*

Figure 4.3.1: *Hierarchical framework for the two-way outlier detection model in Section 4.2.1 and the explanatory two-way outlier detection model in Section 4.2.2. The box with $i = 1, \ldots, N$ in the top-left corner indicates that each parameter inside is specific to a value of i. The same explanation is also applied to the box with $j = 1, \ldots, J$ in the bottom-right corner.*

by mean-centring person-specific covariates $\boldsymbol{x}_i$ and leaving the intercept, c_0, out of model fitting. In the special case where the person-specific covariate column vector $\boldsymbol{x}_i$ is empty, the model assumes that θ_i does not depend on the covariates: $\theta_i \sim \mathrm{N}(0, \sigma^2)$. We can also make θ_i heteroscedastic by allowing its variance to depend on relevant covariates if the variance varies across regions and subpopulations, or changes over time. This possible extension is discussed in Section 4.6.

The similar covariate-dependent prior can be specified for the item parameter, β_j. The conditional distribution of η_j given β_j is also defined in Equation (4.2.2). β_j is

assumed to follow a normal prior with the prior mean depending on the item-specific covariates $\boldsymbol{z}_j$: $\beta_j \sim \mathrm{N}(\mu_j, \omega^2)$, where $\mu_j = \boldsymbol{d}^\top \tilde{\boldsymbol{z}}_j$, $\tilde{\boldsymbol{z}}_j^\top = (1, \boldsymbol{z}_j^\top)$, and parameters $\boldsymbol{d} = (d_0, d_1, \ldots, d_q)$. In the special case where the item-specific covariate column vector $\boldsymbol{z}_i$ is empty, the model assumes β_j to be independent of the covariates: $\beta_j \sim \mathrm{N}(\mu, \omega^2)$, where $\mu = d_0$.

It remains to assign the hyperpriors to parameters that do not vary across individuals and items. The same weakly informative priors previously used in Section 3.4.1 are assigned to ω^2, σ^2 and δ again.

1. σ^2 and ω^2 independently follow an inverse-Gamma distribution with shape 0.5 and scale 1, denoted as $\mathrm{IG}(0.5, 1)$.

2. Outliers may positively or negatively affect item response probabilities, depending on context. This needs to be reflected by the prior for the drift parameter. A half-Cauchy distribution with a scale of 2.5 is assigned to δ in the application of cheating detection because cheating on compromised items would increase the positive response probabilities.

The intercepts a_0 and b_0 are constrained to be strictly negative since the proportions of outlying individuals and items are on the interval $(0, 1)$ and assumed to be lower than 50%. Ideally, the prior distribution for the intercepts is expected to be slowly approaching zero, because the proportion of outlying individuals or items should avoid the boundaries of the $(0, 1)$ interval. Therefore, a Gamma distribution with shape parameter 3 and scale parameter 1, denoted as $\mathrm{Gamma}(3, 1)$, is independently assigned to $-a_0$ and $-b_0$.

Finally, we move on to regression coefficients in Equation 4.2.2, $\boldsymbol{a}$, $\boldsymbol{\lambda}$, $\boldsymbol{b}$ and $\boldsymbol{\phi}$ bar a_0 and b_0. Without sufficient information about whether these coefficients are positive or negative, we have good reason to assume a weakly informative hyperprior centred at zero. Moreover, we would like to restrict these coefficients away from extremely large values and meanwhile do not want to avoid any possible large value. Therefore, a Normal distribution with mean 0 and variance 25 is assumed:

$a_1, \ldots, a_p, b_1, \ldots, b_q, \lambda_1, \lambda_2, \phi_1 \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, 5^2)$. The same choice is made for $\boldsymbol{c}$ and $\boldsymbol{d}$: $c_1, \ldots, c_p, d_0, d_1, \ldots, d_q \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, 5^2)$.

4.3.2 Compound Decision

The detection of outlying individuals and items is assessed within a compound decision based on the item responses and the covariate information from all individuals and all items. The classification is made according to the posterior probabilities of the latent indicators given item responses and covariates: $P(\xi_i = 1|\boldsymbol{Y}, \boldsymbol{x}_i)$ and $P(\eta_j = 1|\boldsymbol{Y}, \boldsymbol{z}_j)$, where $i = 1, \ldots, N$ and $j = 1, \ldots, J$. However, as mentioned earlier on, the posterior probability that a person or an item is an outlier is an individual-wise measure. To evaluate decision-making at an aggregated level of all individuals and all items, the compound decision theory developed in Sections 3.3.2 and 3.3.3 is needed. The quality of decisions is again determined by the False Discovery Proportion (FDP) and the False Non-discovery Proportion (FNP). Since both FDP and FNP cannot be directly obtained without knowing the status of each person and each item, the posterior means of the FPP and FNP, which are known as the local FDR and local FNR, are used instead. Given data and a decision rule, the local FDR and local FNR are completely determined under the proposed model (Efron, 2008; Efron et al., 2001; Robbins, 1951).

4.3.3 Model Comparison

The model comparison is carried out using the deviance information criterion (DIC; Spiegelhalter et al., 2002), which can be computed while estimating marginal likelihood using MCMC methods. Again we base the DIC on the marginal (log-)likelihood in which the person- and item-specific parameters are treated as latent variables or random effects and integrated out.The calculation details are already provided in Section 3.4.3. The model with a smaller DIC value would be more compatible with a replicated dataset of the same structure as the observed data and is therefore pre-

ferred. In the case study, the marginal DIC is used to determine whether covariates provide substantial information about the latent indicators for persons and items.

4.4 Case Study: Licensure Test Data

The proposed explanatory outlier detection model is applied to the same computer-based licensure test dataset introduced in Section 3.5.1. The data have been applied to the previously proposed two-way outlier detection models to simultaneously detect test takers with preknowledge and compromised items without knowing a priori the status of each test taker and each item. As described in Section 3.5.1, the dataset contains $N = 1624$ test takers and $J = 170$ test items. There is one person whose attempt count was recorded as zero. Since a score of 1 indicates that the person is a first-time test taker, a score of zero is likely to be an input error. The final version of the dataset consists of the responses from 1623 test takers ($N = 1623$). The testing program flagged a proportion of test takers and items as suspects. These flagged candidates or items may not be the actual cheaters or compromised items, but again we use this information as partial truth to evaluate classifications based on the explanatory model.

4.4.1 Description of Potential Covariates

The Licensure test data contain the background information of each candidate, including their attempt count, the country where they were schooled, the state where they applied for a licence, test centre, test training institution etc. The description and coding for external variables of interest are presented in Table 4.4.5.

The first person-specific observed covariate under consideration is attempt count since a test taker's performance in terms of total scores is negatively correlated to their number of attempts, according to Cizek and Wollack (2016). Table 4.4.1 shows the frequency for attempt counts ranging from 0 to 4 and more than 4. Since the

median and the mode are 1, we used a binary variable for the attempt count: $x_{1i} = 0$ if person i is a first-time test taker, and $x_{1i} = 1$ if person i has attempted more than once and is a repeat examinee.

The second person-specific observed covariate of our interest indicates the country where a test taker was educated. Table 4.4.2 lists the countries and their corresponding frequencies. We grouped this variable into a binary indicator denoted by x_{2i}: 0 for "USA" and 1 for "non-USA".

Cizek and Wollack (2016) mentioned that one's pass rate tends to be related to the country where they went to schools across their previous attempts. Among those who were educated in the United States, the pass rates tend to be much lower for repeat test takers than those for first-time candidates. By contrast, pass rates tend to remain similar across attempts for those who were schooled in Asian countries, regardless of their attempt counts. Therefore, an interaction between the attempt count and the country is incorporated to account for the potential difference between first-time and repeat candidates given the countries where they were schooled.

No. Attempts	0	1	2	3	4	4+
Frequency	1	1161	207	104	47	104

Table 4.4.1: *Licensure Test Data: Frequency table of the number of attempts made by test takers. A score of 1 indicates that candidate is a new, first-time examinee. Score for any person sitting for the test for the fourth time or more is marked as 4+.*

Country	USA	Non-USA
Frequency	1239	384

Table 4.4.2: *Licensure Test Data: Frequency table of countries where test takers were educated.*

Observed covariates on the item side include item usage and variables related to the location of the testing centre. The latter is way too sparse and barely informative. Items usage is considered by previous research (Cizek & Wollack, 2017) to be helpful in explaining the chance of leakage. So our focus is on item usage as a predictor to inform the latent indicator of outlying items.

Table 4.4.3 shows the frequencies of the times the $J = 170$ items have been used. There are 78 items that have been used more than twice, 69 items that have been

used twice, and only 23 items that have not been used in previous tests. We used the median ("2") as the cutting-off point and regrouped the item usage using two dummy variables z_1 and z_2, as shown in the first two panels in Table 4.4.4. Panel 4.4.4c shows that the two dummy variables indicate three categories from "never used before" to "repeatedly used": "less than twice" if $z_1 = 0$ and $z_2 = 0$; "exactly twice" if $z_1 = 1$ and $z_2 = 0$; "more than twice" if $z1 = 0$ and $z_2 = 1$.

Item Usage	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	23	69	37	19	5	7	3	1	3	1	1	1

Table 4.4.3: *Licensure Test Data: Frequency table of the times items have been used in tests.*

Item Usage	$\neq 2$	$= 2$
z_1	0	1
Frequency	101	69

(a) *Item usage coded by a single dummy variable z_1.*

Item Usage	$\leqslant 2$	> 2
z_2	0	1
Frequency	92	78

(b) *Item usage coded by a single dummy variable z_2.*

Item Usage	< 2	$= 2$	> 2
z_1	0	1	0
z_2	0	0	1
Frequency	23	69	78

(c) *Item usage coded by two dummy variables z_1 and z_2.*

Table 4.4.4: *Licensure Test Data: Frequency table of item usage.*

Finally, we include two continuous covariates relating to response times: averaged response time taken by each test taker and averaged response time spent on each item, denoted by x_{i3} and z_{j2}, respectively. Cizek and Wollack (2016) based the detection of compromised items on response times. The comparison between the reduced and the full model in Chapter 3 also suggests that response time data contain substantial information about the test takers with preknowledge and compromised items. We, therefore, expect that the inclusion of response times would improve the prediction of latent class memberships of individuals and items.

4.4.2 Model Specification

We first specify the relationships between the latent indicators of outlyingness, ξ_i and η_j, and predictors on both person and item levels. As mentioned in Section 4.2.2, the latent indicators of test takers with preknowledge and compromised items may also depend on the latent person and item parameters, in addition to observed

covariates. Person ability has been used to predict the chance of cheating (Cizek, 1999). Cizek and Wollack (2016) gave the empirical evidence that candidates at the top are less likely to be flagged by the testing program. Simha and Cullen (2012), however, pointed out that the average test takers are more likely to cheat than the struggling or top candidates. For these reasons, we use a quadratic function to represent the relationship between person-specific latent indicator ξ_i and person parameter θ_i.

Item difficulty is believed to be associated with the chance of being compromised. Presumably, most test takers are able to correctly respond to easy questions and therefore don't really have the motive for gaining inappropriate access to this content. It is the difficult test items that distinguish top candidates from the rest. Therefore, difficult items are more likely to be compromised. Under the assumption of a positive association between item difficulty and an item's chance of being compromised, the difficulty level is included in the explanatory model as a linear predictor. Cizek and Wollack (2016) also pointed out item difficulty level might change as item usage increases. Therefore, it might be also worth investigating whether the impact of item difficulty exerts on the chance of being compromised varies from new items to frequently recycled items.

In Table 4.4.5, we describe observed and latent predictors and their associated coefficients (Equation 4.2.2).

Predictor	Coef.	Definition	Variable Type
x_{i1}	a_1	No. of attempts made by person i	Binary (1 for "more than once")
x_{i2}	a_2	Country where person i was educated	Binary (0 for US, 1 for non-US)
$x_{i1} \cdot x_{i2}$	a_3	Interaction effect between attempt count and country for person i	Binary
x_{i3}	a_4	Average response time taken by person i	Continuous
θ_i	λ_1	Ability of person i	Continuous, latent
θ_i^2	λ_2	Squared ability of person i	Continuous, latent
z_{j1}	b_1	Times item j has been used in tests	Binary (1 for 'twice')
z_{j2}	b_2	Times item j has been used in tests	Binary (1 for 'more than twice')
z_{j3}	b_3	Average response time spent on item j	Continuous
β_j	ϕ_1	Difficulty level of item j	Continuous, latent

Table 4.4.5: *Licensure Test Data: A list of predictors along with their corresponding coefficients in the explanatory model.*

We now move on to specify the distributions of continuous latent variables, conditional on covariates. One's attempt count (x_{i1}) is likely to be indicative of one's ability. One's proficiency in test items may also be related to the country where a candidate was schooled (x_{i2}) since different solution strategies had been taught through different education experiences and may have varying degrees of effectiveness. Moreover, the average time taken by a test taker (x_{i3}) may be associated with their ability as those who spend more time on test items are likely to be less capable. Thus, θ_i is assumed to follow a normal linear model with a constant variance in the form of

$$\theta_i \sim \mathrm{N}(c_1 x^*_{i1} + c_2 x^*_{i2} + c_3 x^*_{i3}, \sigma^2), \tag{4.4.1}$$

where x^*_{i1}, x^*_{i2} and x^*_{i3} refer to the mean-centred covariates so that the arithmetic mean of the prior means of θ_i, for $i = 1, \ldots, N$, can be fixed at zero.

Item difficulty may depend on item usage since a decrease in mean item difficulty is likely to happen when test items are recycled, according to Wood (2009). Furthermore, the amount of time spent on an item (z_{j3}) is likely to be associated with the item's difficulty, as the items which cost test takers more time are likely to be more difficult. The item difficulty, β_j, is therefore assumed to follow a normal linear model with a constant variance given by

$$\beta_j \sim \mathrm{N}(d_0 + d_1 z_{j1} + d_2 z_{j2} + d_3 z_{j3}, \omega^2). \tag{4.4.2}$$

Table 4.4.6 presents the covariates used for characterising the distributions of latent parameters, θ_i and β_j, under the explanatory model.

Predictor	Coef.	Definition	Variable Type
x_{i1}	c_1	No. of attempts made by person i	Binary (1 for "more than once")
x_{i2}	c_2	Country where person i was educated	Binary (0 for US, 1 for non-US)
x_{i3}	c_3	Average response time taken by person i	Continuous
z_{j1}	d_1	Times item j has been used in tests	Binary (1 for 'twice')
z_{j2}	d_2	Times item j has been used in tests	Binary (1 for 'more than twice')
z_{j3}	d_3	Average response time spent on item j	Continuous

Table 4.4.6: *Licensure Test Data: A list of covariates and their associated coefficients used for characterising the distributions of latent parameters.*

4.4.3 Results

The item response data with covariates were analysed using the proposed explanatory model. Based on the MCMC algorithm given in Appendix C.1, we ran 10 MCMC chains with random starting points. For those parameters which also appear in the outlier detection model, we used the posterior means obtained in the previous chapter as the baseline for their initial values. Through inspecting trace plots for the deviance of each chain and calculating the Gelman-Rubin (GR) statistic for global parameters (GR is below 1.116 for all the global parameters), we concluded that these chains converged to their equilibrium distributions after 35,000 iterations.

The inference is made based on 100,000 posterior samples from the ten converged chains, where each chain contributes 10,000 samples. Posterior means and 95% credible intervals for global parameters (i.e. the parameters that do not vary across different individuals or items) are presented in Tables 4.4.7, 4.4.8 and 4.4.9. The posterior means are used as parameter estimates and the credible intervals summarise the uncertainty in parameter estimates.

We are particularly concerned about the intercepts a_0 and b_0 since they are related to the proportions of compromised individuals and items. Notice that the proportions of test takers and items flagged by the internal testing program are approximately 2.5% and 37.6%, respectively. The priors specified for the intercepts, a_0 and b_0, restrain their values to be negative so that the proportions can be constrained below 50%. The details for weakly-informative priors are already provided in Section 4.3.1. The interpretation of posterior mean estimates for the parameters in Equation 4.2.2 is described as follows.

1. The posterior mean for a_0 is -3.2564, meaning that the odds that a first-time test taker who was schooled in the US cheats is $\exp(-3.2564) = 0.0385$, without taking into account θ and the average response time x_3.

2. The posterior mean for a_1 can be interpreted as the difference in the log-odds between repeat and first-time test takers for those who were schooled in the

US, holding the person ability and the average response time at constant levels or values. Table 4.4.7 shows that there is a 95% probability that the "true" effect of x_2 lies within an interval containing zero, given the observed data.

3. The posterior mean for a_2 can be interpreted as the difference in the log-odds between those who were educated outside the US and those who were educated in the US among first-time candidates, holding the person ability and the average response time constant. The posterior mean for a_2 is 2.2648 and therefore the odds are $\exp(2.2648) = 9.6292$, indicating that getting prior access to compromised items is more than nine times as likely for those who were educated outside the US while controlling for other covariates, given the observed item response data.

4. The posterior mean for a_3 can be interpreted as the additional difference in the log-odds between those whose education was based outside the US and those who were educated in the US if they are repeat test takers rather than first-time takers.The posterior mean for a_3 is -1.4376 and therefore the odds are $\exp(-1.4376) = 0.2375$, indicating that obtaining prior access to the compromised item is over four times as likely for the repeat test takers who were also educated outside the US while controlling for other covariates, given the observed item response data.

5. The posterior mean for a_4 is -0.1374, and the 95% credible interval does not include zero, meaning that for an extra second spent on test items, the expected change in the log odds ratio is -0.1374. This leads to a nearly 13% decrease in the odds of cheating ($\exp(-0.1374) = 87.16\%$) for first-time examinees who were educated in the US while holding other predictors constant.

6. The posterior mean for b_0 can be interpreted as the log-odds for items that have been reused less than twice if values for other continuous predictors are fixed at zero. Table 4.4.7 shows that the odds for items that have been reused less than twice is $\exp(-0.7438) = 0.4753$, meaning that the estimated π_2 would

be 0.3195 if we fix z_1 and z_2 at their reference categories (i.e. items are used less than twice) and ignore the effects of β and the average response time z_3.

7. The posterior mean for $b_0 + b_1$ can be interpreted as the log-odds for items that have been reused twice, controlling for other predictors. The posterior mean for $b_0 + b_2$ can be interpreted as the log-odds for items that have been reused more than twice, controlling for other predictors. According to Table 4.4.7, there is a 95% probability that the "true" effect estimate b_1 or b_2 lies within an interval containing zero, given the observed item response data.

8. The posterior mean for b_3 can be interpreted as the expected change in the log-odds for an extra second spent on items. The posterior mean is -0.0343 and the 95% does not include zero, indicating that holding other predictors constant, for an extra second spent answering test items, the expected change in the odds is $\exp(-0.0343) = 0.9663$. This change implies that items are slightly less likely to be compromised as the average response time increases.

Note that the following inferences drawn about subpopulation groups informed by covariates containing demographic information (e.g. x_2) are used as a guide as to how to interpret a covariate effect on latent class membership. It is necessary to be cautious about making inferences about the subpopulations indicated by demographic variables while making high-stakes decisions (Adams, Wilson, & Wu, 1997).

Table 4.4.8 shows the estimation results for parameters that define the relationships between the latent traits and covariates. According to the table, we are 95% convincing that the "true" effect that each covariate exerts on the means of person ability and item difficulty falls in an interval excluding zero. The posterior means for c_1, c_2 and c_3 are negative, meaning that the three person-specific covariates are negatively associated with the latent indicator of test takers with preknowledge. To be more specific, while holding other parameters constant, repeat test takers, those who were schooled outside the US, and those who spent more time on average in response to test questions tend to be less proficient in contrast to first-time test tak-

ers, those who were schooled in the US, and those who respond to test items faster. The posterior means for d_1 and d_2 are negative, suggesting that reusing items is negatively associated with item difficulty. The posterior mean for d_3 is positive, meaning that items which cost test takers more time on average tend to be more difficult. The estimation of the relationships is in line with our expectations.

Predictor	Coefficient	Posterior Mean	95% CI
	a_0	-3.2564	(-3.7378, -2.8288)
x_1 (attempt)	a_1	-0.9333	(-2.7905, 0.3436)
x_2 (country)	a_2	2.2648	(1.3026, 3.2954)*
$x_1 \cdot x_2$ (interaction)	a_3	-1.4376	(-1.9427, -0.9596)*
x_3 (time.person)	a_4	-0.1374	(-0.1739, -0.1034)*
θ (ability)	λ_1	0.3265	(0.0210, 0.6694)*
θ^2 (ability^2)	λ_2	-0.1624	(-0.2833,-0.0658)*
	b_0	-0.7438	(-1.3174, -0.6366)
z_1 (usage1)	b_1	-0.3540	(-0.8600, 0.1484)
z_2 (usage2)	b_2	-0.2694	(-0.7682, 0.2267)
z_3 (time.item)	b_3	-0.0343	(-0.0577, -0.0128)*
β (difficulty)	ϕ_1	0.3941	(0.0115, 0.7907)*

Table 4.4.7: *Licensure Test Data: Posterior means and 95% credible intervals for parameters in the explanatory part of the model. The superscript $*$ for 95% credible interval indicates that given the observed data, we can be 95% sure that the "true" effect that the predictor exerts on the log-odds ratio falls within the range excluding zero.*

Predictor	Coefficient	Posterior Mean	95% CI
x_1 (attempt)	c_1	-0.5678	(-0.7162, -0.4195)*
x_2 (country)	c_2	-1.0849	(-1.1876, -0.9822)*
x_3 (time.person)	c_3	-0.1027	(-0.1125, -0.0114)*
	d_0	-0.4216	(-0.9984, -0.1976)
z_1 (usage1)	d_1	-0.7831	(-1.1387, -0.4275)*
z_2 (usage2)	d_2	-0.8823	(-1.2163, -0.5483)*
z_3 (time.item)	d_3	0.0346	(0.0174,0.1223)*

Table 4.4.8: *Licensure Test Data: Posterior means and 95% credible intervals for parameters characterising the relationship between θ_i and $\boldsymbol{x}_i$, β_j and $\boldsymbol{z}_j$. The superscript $*$ for 95% credible interval indicates that we can be 95% sure that the "true" effect that the predictor exerts on the prior means of θ_i and β_j falls within the range excluding zero given the observed covariates.*

The classification performance is assessed based on the AUCs under the ROC curves for the two-way classification while tactically using the labels provided by the testing program as the "true" status. Figure 4.4.1 displays the ROC curves for the classification of test takers and items by posterior means of latent indicators ξ_i's and

Parameter	Posterior Mean	95% CI
σ^2	0.3083	(0.1823, 0.4151)
ω^2	0.4978	(0.3941, 0.7680)
δ	0.6582	(0.5420, 0.8359)

Table 4.4.9: *Licensure Test Data: Posterior means and 95% credible intervals for parameters in the measurement model.*

η_j's. Compared with the AUCs (86.8% for individuals and 83.6% for items) based on the reduced model with the cheating effect but without covariate effects given by Equation (3.2.1), the AUC values under the explanatory model are slightly higher for test takers (87.1%) but substantially higher for items (91.0%).

We then assess the detections made under the compound decision-making framework. The goal of the compound decisions is to flag as many suspicious test takers as possible while ensuring that no more than 1%, 5% or 10% of the test takers are mistakenly flagged. Put another way, if a person is detected, their probability of cheating is at least 99%, 95% or 90%. As for the detection of compromised items, the goal is to control the quality of the remaining items, while not removing too many items. Figure 4.4.2 shows how the local FDR or the local FNR changes as the number of detections for test takers or items increases. Table 4.4.10 shows the numbers of detections in regard to test takers and items when the local FDR and the local FNR are set at three different thresholds (1%, 5% and 10%). We use an example to explain the table in regard to the curves above. When the local FDR is controlled at 1%, 19 test takers are detected as cheaters. This is reflected by the intersection between the local FDR (the black curve) and the 1% threshold (the red dashed line) in Panel (a), Figure 4.4.2, when applying the proposed compound decision rule to test takers. As we can see in Panel (a), as the number of detections increases, the local FDR increases. The number of detections for individuals under the reduced model without the structural model component in Chapter 3 are 25, 46 and 61, respectively, while controlling the local FDR at 1%, 5% and 10% levels. In contrast to the previous result, the current model detects slightly fewer individuals when the FDR is set at 1% and 5%, but slightly more when the FDR is set at

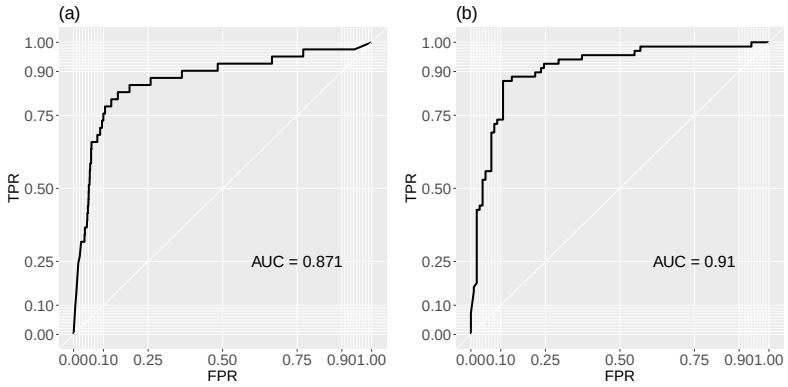


Figure 4.4.1: *Licensure Test Data: ROC curves for classification under the explanatory model for item response data. Panel (a): ROC curve for the classification of cheaters (labelled by the testing program) by the posterior means of ξ_i. Panel (b): ROC curve for the classification of compromised items (labelled by the testing program) by the posterior means of η_j. The x- and y-axes of a ROC curve give the true positive rate (TPR) and false positive rate (FPR) for classification, respectively.*

10%. The same plot for items is given in Panel (b), where we control local FNR for items. Panel (b) shows that the local FNR decreases as the number of detected items increases. The results under the current model (i.e. 103, 92, 71) are close to the results under the reduced model without covariates in Chapter 3 (i.e. 100, 91 and 71) It seems that the inclusion of covariate effects and the dependence of classification on the continuous latent variables changes the decisions on flagging individuals, but does not contribute much to the decisions on flagging items. Again, we remark that the validity of the detection results depends on the extent to which our model assumptions hold. Therefore, we suggest treating such detection results as initial screening results, rather than as the final decisions.

	1%	5%	10%
Test takers	18	48	70
Item	103	92	71

Table 4.4.10: *Licensure Test Data: The number of detections for test takers while the local FDR is controlled at 1%, 5% and 10%, and the number of detections for items while controlling the local FNR at 1%, 5% and 10%.*

We then compare marginal DIC values in Table 4.4.11 between three models, namely

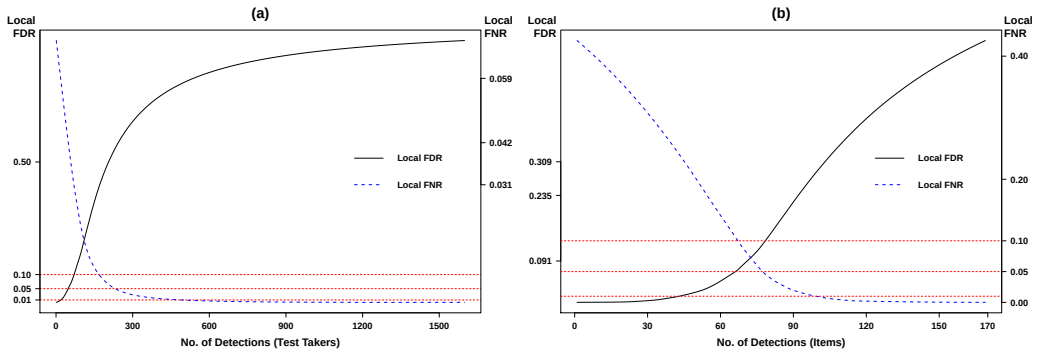


Figure 4.4.2: *Licensure Test Data: Detections based on the explanatory model for item responses: The local FDR (represented by black solid curves) and the local FNR (represented by blue dashed curves) as functions of the number of detections.*

the null model (without the cheating effect), the reduced model (with the cheating effect, without the explanatory framework) and the explanatory model (with the cheating effect, within the explanatory framework). The marginal DIC for the null model (218,867) is considerably higher than the marginal DIC for the other two, suggesting that cheating in the form of item preknowledge is likely to exist among some test takers. This result is also consistent with the analysis in Section 3.5.2. Between the reduced model and the explanatory model, the latter is preferred due to a smaller marginal DIC, meaning that the explanatory model is more compatible with replicated data of the same structure as the licensure test data.

	$\boldsymbol{a}=\mathbf{0}$ & $\boldsymbol{b}=\mathbf{0}$ $\boldsymbol{\lambda}=\mathbf{0}$ & $\boldsymbol{\phi}=\mathbf{0}$	$\boldsymbol{a}\neq\mathbf{0}$ & $\boldsymbol{b}\neq\mathbf{0}$ $\boldsymbol{\lambda}\neq\mathbf{0}$ & $\boldsymbol{\phi}\neq\mathbf{0}$
$\xi_i\eta_j=0$	218,867	–
$\xi_i\eta_j=1$	141,026	135,295

Table 4.4.11: *Licensure Test Data: Marginal* DIC *values under different specifications of the proposed model.*

4.5 Simulation Study

4.5.1 Settings

We use a simulation study for assessing the performance of the proposed model in terms of classification and model adequacy under different sizes of the data and

varying degrees of separation between classes. Table 4.5.1 displays the four simulation settings with respect to the sample size, the number of items and the drift parameter value. We consider two different sample sizes and the number of items: $N = 1000, 2000$, and $J = 100, 200$. Class separation is set to be low and high by fixing the drift parameter δ in the measurement model at 0.5 and 1.0, respectively. $\delta = 0.5$ corresponds to a limited advantage that a test taker can benefit from their prior access to compromised items. $\delta = 1.0$ corresponds to a bigger advantage taken by a test taker's preknowledge of compromised items.

For each setting 100 independent datasets were generated from an explanatory model that accounts for covariate effects on the latent indicators and the continuous latent variables. To be more specific, we consider two binary covariates: x_{1i}, for person $i = 1, \ldots, N$, and z_{1j} for item $j = 1, \ldots, J$. The explanatory framework in the data-generating model is specified below.

$$\begin{aligned} P(\xi_i = 1|\boldsymbol{x}_i, \theta_i) &= \frac{\exp(a_0 + a_1 x_{1i} + \lambda_1 \theta_i)}{1 + \exp(a_0 + a_1 x_{1i} + \lambda_1 \theta_i)} \\ P(\eta_j = 1|\boldsymbol{z}_j, \beta_j) &= \frac{\exp(b_0 + b_1 z_{1j} + \phi_1 \beta_j)}{1 + \exp(b_0 + b_1 z_{1j} + \phi_1 \beta_j)} \end{aligned}, \tag{4.5.1}$$

where the means of person and item parameters depend on covariates: $\theta_i \sim \mathrm{N}(c_1 x^*_{i1}, \sigma^2)$, where the person-specific covariate is mean-centred, and $\beta_i \sim \mathrm{N}(d_0 + d_1 z_{j1}, \omega^2)$.

The model component for item responses is given by Equation (4.2.1). Aside from the drift parameter δ, the values for parameters are set at the posterior means obtained from the licensure test data analysis (Tables 4.4.7, 4.4.8 and 4.4.9).

For each dataset, we apply (A) the explanatory model, (B) the explanatory model without accounting for covariate effects on the continuous latent variables θ_i and β_j, and (C) the explanatory model without accounting for covariate effects on the continuous latent variables and the latent indicators (i.e. the reduced model proposed in Chapter 3). The comparisons are made in regard to the classifications and the decisions across the four settings under the explanatory model. In order to see whether the inclusion of covariate effects can improve classifications and contribute

to the decision-making, we also compare the performance of the explanatory model (A) and its two submodels (B, C) under each setting.

Setting	δ	N	J
S1	0.5	1000	50
S2	1.0	1000	50
S3	0.5	2000	100
S4	1.0	2000	100

Table 4.5.1: *Simulation study: Four settings which differ by the benefits from cheating and the sample and item sizes.*

4.5.2 Results

The MCMC algorithm (Appendix C.1) is applied. For each simulated dataset, we ran 15,000 iterations, with the first 5,000 iterations being discarded as the burn-in. The estimation and classification results are based on the posterior samples from the last 10,000 iterations.

We first present the estimation results of the global parameters in the explanatory model. Table 4.5.2 shows the bias and variance of the posterior mean estimator for each global parameter based on the 100 replicated datasets. In general, the bias is close to zero for all global parameters under the four settings. We also find that the estimation under S3 tends to be more accurate in contrast to S1 because of the increased sample size and number of items The same conclusion can be drawn when comparing the estimation results under S2 and S4.

We assess the performance of the explanatory two-way outlier detection model at all possible classification thresholds by comparing posterior samples of person- and item-specific latent indicators, ξ_i and η_j, and their true status, the values used for simulating the data. The area under the ROC curve (AUC) provides an overall measure of the classification performance across all classification thresholds. As mentioned in Chapter 3, the AUC indicates the probability that the model ranks a random person or item with the compromised status more highly than a random person or item without the compromised status. Table 4.5.3 shows the first, the second

	a_0	a_1	λ_1	b_0	b_1	ϕ_1	c_1	d_0	d_1	σ^2	ω^2	δ
(S1)												
Bias	-0.12	-0.04	0.18	0.08	0.24	-0.18	0.07	0.12	0.13	0.18	-0.21	0.13
Var	0.49	0.46	0.24	0.56	0.51	0.31	0.18	0.28	0.17	0.34	0.32	0.31
(S2)												
Bias	0.14	-0.15	0.07	-0.08	0.11	0.10	0.06	-0.14	0.09	0.11	0.13	-0.08
Var	0.51	0.44	0.21	0.52	0.44	0.25	0.12	0.29	0.19	0.32	0.34	0.34
(S3)												
Bias	0.17	0.15	-0.09	-0.05	0.07	0.06	-0.01	0.06	0.08	-0.13	-0.25	-0.11
Var	0.36	0.38	0.18	0.27	0.25	0.19	0.07	0.22	0.13	0.29	0.27	0.27
(S4)												
Bias	0.17	0.15	-0.09	-0.05	0.07	0.06	-0.01	0.06	0.08	0.09	-0.17	-0.09
Var	0.25	0.31	0.14	0.23	0.26	0.15	0.09	0.14	0.09	0.23	0.26	0.29

Table 4.5.2: *Simulation Study: The bias and variance for the posterior mean of the second-level parameters in the explanatory model based on the 100 replicated datasets under four simulation settings.*

and the third quartiles of AUC values for persons and items under the explanatory model and its two submodels in the four settings. From S1 to S3 or from S2 to S4, as the sample and item sizes increase, the overall classification for both items and persons is more satisfactory when the drift is held constant. By comparing the AUCs under S1 and S2 (or under S3 and S4), when the drift gets larger, meaning that individuals would benefit more from their preknowledge of compromised items, the overall classification gets better in general when the sample and item sizes are held constant. Both tendencies are in line with what we expected; that is, a larger number of outlying individuals or items, and a stronger outlier effect would better inform the classification.

We further apply the compound decision rules to estimate whether the test takers or the items are compromised in aggregate. The compound decision rules are applied under the explanatory model and its two sub-models. For each independent replicated dataset, the local FDR and local FNR are controlled at 1%, 5% and 10%, respectively, for all individuals and items. Each decision rule is evaluated by examining the resultant FDP and FNP which are shown in Tables 4.5.4 and 4.5.5. In general, both the FDP for the classification of individuals and the FNP for the classification of items are well controlled under the explanatory model. The only exceptions occur when the local FDR and local FNR are controlled to be below 1%

(A)	Individuals				Items			
AUC	S1	S2	S3	S4	S1	S2	S3	S4
25%	0.925	0.935	0.933	0.947	0.921	0.933	0.941	0.961
50%	0.947	0.969	0.962	0.978	0.956	0.970	0.968	0.979
75%	0.984	0.991	0.987	0.993	0.982	0.990	0.994	0.995
(B)	Test takers				Items			
AUC	S1	S2	S3	S4	S1	S2	S3	S4
25%	0.917	0.923	0.916	0.938	0.918	0.924	0.923	0.938
50%	0.929	0.957	0.961	0.978	0.958	0.947	0.965	0.980
75%	0.978	0.990	0.984	0.989	0.974	0.980	0.979	0.992
(C)	Test takers				Items			
AUC	S1	S2	S3	S4	S1	S2	S3	S4
25%	0.905	0.915	0.913	0.931	0.909	0.921	0.925	0.938
50%	0.918	0.954	0.949	0.965	0.947	0.959	0.956	0.975
75%	0.946	0.963	0.971	0.985	0.972	0.979	0.984	0.995

Table 4.5.3: *Simulation Study: Overall classification performance based on the posterior means of ξ_i and η_j under (A) the explanatory model, (B) the explanatory model without accounting for covariate effects on the continuous latent variables, and (C) the reduced model. For each setting, and each target (test taker/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 100 replicated datasets.*

for individuals and items, respectively, the resultant FDP and FNP slightly exceed the target level in the settings where the sample and item sizes are smaller. In S4, where the sample and item sizes are 2000 and 100, and the outlier effect is stronger, both the FDP and the FNP are below 1%. Under the model without covariate effects on person ability and item difficulty and the reduced model, the FDP and FNP are not as well-controlled as they are under the explanatory model. That said, their values are still close to or under the target levels, especially in S3 and S4, where the size of the data is larger.

4.6 Concluding Remarks

In this chapter, we have extended the Bayesian hierarchical two-way outlier detection model in an explanatory framework. In doing so, covariates are linked with individuals' and items' latent class memberships, and the distributions of continuous latent parameters. The explanatory framework also enables us to relax the assumption of the independence between latent indicators and latent traits, which

(A)	S1			S2			S3			S4		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.008	0.039	0.072	0.009	0.037	0.065	0.007	0.029	0.063	0.006	0.032	0.069
50%	0.011	0.043	0.084	0.009	0.041	0.078	0.009	0.031	0.071	0.008	0.035	0.080
75%	0.013	0.052	0.096	0.011	0.048	0.083	0.012	0.045	0.079	0.009	0.044	0.085
(B)	S1			S2			S3			S4		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.008	0.038	0.089	0.009	0.041	0.086	0.007	0.036	0.082	0.008	0.034	0.081
50%	0.012	0.047	0.095	0.012	0.047	0.094	0.007	0.043	0.091	0.009	0.041	0.088
75%	0.014	0.054	0.104	0.013	0.056	0.108	0.010	0.051	0.098	0.011	0.048	0.101
(C)	S1			S2			S3			S4		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.013	0.047	0.091	0.010	0.046	0.089	0.008	0.041	0.083	0.008	0.038	0.080
50%	0.014	0.051	0.098	0.012	0.054	0.095	0.011	0.047	0.091	0.010	0.046	0.093
75%	0.017	0.055	0.111	0.016	0.058	0.109	0.012	0.051	0.104	0.012	0.052	0.103

Table 4.5.4: *Simulation Study: Local FDR control for individuals under (A) the explanatory model, (B) the explanatory model without accounting for covariate effects on the continuous latent variables, and (C) the reduced model. For each setting and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 100 independent datasets.*

(A)	S1			S2			S3			S4		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.008	0.039	0.076	0.008	0.041	0.078	0.007	0.038	0.074	0.007	0.033	0.069
50%	0.010	0.044	0.085	0.009	0.043	0.085	0.008	0.042	0.083	0.007	0.043	0.082
75%	0.011	0.054	0.092	0.012	0.051	0.091	0.010	0.047	0.086	0.009	0.048	0.088
(B)	S1			S2			S3			S4		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.007	0.042	0.088	0.009	0.039	0.089	0.008	0.042	0.084	0.007	0.039	0.085
50%	0.009	0.049	0.093	0.012	0.048	0.092	0.008	0.047	0.089	0.009	0.046	0.091
75%	0.013	0.056	0.101	0.012	0.055	0.106	0.011	0.052	0.097	0.011	0.051	0.094
(C)	S1			S2			S3			S4		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.011	0.044	0.092	0.012	0.047	0.090	0.008	0.045	0.087	0.007	0.044	0.086
50%	0.013	0.052	0.095	0.013	0.052	0.097	0.010	0.049	0.092	0.011	0.048	0.094
75%	0.014	0.058	0.102	0.014	0.056	0.104	0.013	0.053	0.098	0.012	0.053	0.099

Table 4.5.5: *Simulation Study: Local FNR control for items under (A) the explanatory model, (B) the explanatory model without accounting for covariate effects on the continuous latent variables, and (C) the reduced model. For each setting and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 100 replicated datasets.*

is held by the previous outlier detection model.

Statistical inference is carried out under a full Bayesian setting for which a parallel tempering MCMC algorithm is used, and the compound decision rules are applied to the detection of two-way outliers under the Bayesian decision framework. The explanatory model is applied to the licensure test data which is known to contain individuals who benefit from preknowledge of compromised items and covariate in-

formation. We compare the explanatory model and the reduced model proposed in Chapter 4 in terms of classification results, detections of individuals and items, and model adequacy. Overall, the explanatory model performs better than the previously proposed two-way outlier detection model, suggesting that covariate inclusion and less restrictive model assumptions may be helpful in improving outlier detection.

The performance of the explanatory model is evaluated under four simulation settings with different degrees of outlier effect, sample size and item size. The results show that the overall classification under the explanatory model tends to be more accurate, and the FDP and FNP tend to be better controlled, when the sample size and item size are larger, and when outliers have a stronger effect on response probabilities. The comparisons are also drawn among the explanatory model and its two sub-models, i.e the explanatory model without covariate effects on continuous latent variables, and the explanatory model without covariate effects on both continuous latent variables and latent indicators of outlyingness under the four settings. In conclusion, the overall classification results for both individuals and items given by the explanatory model are better than those under the two sub-models. Furthermore, the FDP and FNP are better controlled under the explanatory model in every simulation setting.

There are several directions that we believe are important and require developments in future studies. As mentioned in Section 4.3.1, the explanatory model allows the distributions of latent traits to depend on relevant covariates. In the case where the dispersion of latent traits varies across different demographic groups, or changes over regions and time, we can make the latent traits θ_i's (and β_j) heteroscedastic by assuming their variance to depend on relevant covariates as well: $\log \sigma_i^2 = \boldsymbol{h}^\top \tilde{\boldsymbol{x}}_i$, for $i = 1, \ldots, N$.

Another research direction would be developing measures for evaluating the effects of including covariates on two-way outlier detection. Covariates are involved in outlier detection through their relationships with latent class memberships of individuals and items. In Section 4.4, covariates are selected based on relevant research on the

topic of item preknowledge. In the presence of a large number of external covariates, it is necessary to simultaneously estimate latent class memberships while carrying out the variable selection. This can be achieved by imposing shrinkage priors on regression coefficients for covariates under a Bayesian framework. Shrinkage priors serve to lead the estimates of coefficients for covariates toward zero. One type of priors used to achieve the lasso-style shrinkage is spike-and-slab priors (Ishwaran & Rao, 2005). Lu, Chow, and Loken (2016) developed the Bayesian structural equation modelling with spike-and-slab priors (BSEM-SSP). The BSEM-SSP specifically quantifies the probabilities that single or multiple factor loadings should be included by evaluating whether the posterior probability of the inclusion of each of these parameters exceeds a pre-determined threshold. This SSP method accounts for the uncertainty in model selection and therefore provides more reliable parameter estimates. It would be of interest to develop an effective Bayesian variable selection method for the explanatory two-way outlier detection model. It is also worthwhile to investigate the sensitivity of the two-way classification to the prior specifications of model parameters other than regression coefficients.

Chapter 5

Conclusions

The purpose of this thesis is to develop model-based approaches for detecting one-way and two-way outliers in multivariate data without relying on prior knowledge of the outlying status of individuals and items. One-way outliers are defined as individuals or items deviating from a baseline model specified for the majority of the data, while two-way outliers are defined as item responses that deviate from a given baseline model due to atypical attributes of both individuals and items. In what follows we summarise key findings of our one-way and two-way outlier detection approaches, and discuss the limitations of the current study and directions for future research.

The Forward Search (FS) is used to detect one-way outliers. The FS bases statistical inference and parameter estimation on a sequence of augmented subsets built during the progression of the FS. The baseline model assumed for the majority of the data is fitted and problem-tailored diagnostic statistics are calculated whenever new individuals or items are introduced to the subset. The effect that one-way outliers have on the fitted model can be revealed by inspecting the evolution of the problem-tailored model fit measures. While in the literature the FS was applied to detect isolated and clustered outliers in multivariate modelling contexts (e.g., factor models and multivariate normal mixture models), its potential has not been explored in latent class and factor mixture modelling contexts with a general purpose

of understanding the hidden structure of the data.

In this thesis, we have applied the FS to detect outlying response patterns deviating from specified latent class models and mixture IRT models. The effect of sequentially adding individual responses on the baseline model was assessed through monitoring a fast-bootstrap p-value for a limited-information goodness-of-fit statistic, which proved to be computationally efficient and less sensitive to sparse binary data. In simulation studies, we addressed the importance of choosing an "outlier-free" initial subset that leads to the robust estimation of the baseline model. When it is computationally infeasible to find an "outlier-free" initial subset, it is important to adopt an alternative strategy; that is, running a sufficiently large number of searches with randomly selected initial subsets. This alternative way of conducting the FS was found to be useful in detecting latent population heterogeneity, e.g., latent classes in a factor mixture model, and determining latent class memberships of individuals.

While the FS in the literature has been primarily used to detect individuals, we have proposed to detect items on the basis of their atypical attributes using the FS. Our contribution to the literature on the FS is to adapt the FS to detect items showing DIF in multiple-group data. In the detection of items deviating from a latent class or an IRT model assuming measurement equivalence, the FS starts from an initial subset formed by equivalent items and proceeds by including the least non-equivalent items at each step of the search. In the simulation and real case studies, the p-value for a limited-information goodness-of-fit statistic was monitored throughout the search to indicate the effect of items lacking measurement equivalence on the measurement-equivalent baseline model.

In summary, the FS has provided valuable insight into the hidden structure of multivariate data in latent variable modelling contexts. It is helpful in determining which part of the data deviates from a baseline model and its assumptions. The proposed FS procedures for detecting outlying individuals, individuals belongings to latent groups, and items exhibiting DIF can be easily adapted to different modelling contexts for different types of data by specifying the baseline model and defining

problem-specific diagnostic statistics for assessing the effect of the sequential addition on the baseline model.

It is concluded that future research needs to address the reliability of different diagnostic statistics monitored during the progression of the FS and the multiple testing issues. Multiple testing arises at the application level and it is important to find an acceptable trade-off between the number of detected outliers and the power of the FS (Riani et al., 2009). It is worth looking into measures such as a p-value adjusted for multiple tests (i.e. false discovery rate; Benjamini & Hochberg, 1995) under simulation settings with different sample sizes and proportions of outliers.

The limitation of the FS as an outlier detection method is that it is unable to simultaneously identify outlying individuals and DIF items. The simultaneous detection of outlying individuals and DIF items is essentially a problem of unsupervised two-way classification, where the status of individuals and items is unknown, and the data are not dominated by either outlying individuals or DIF items. Therefore, we proposed a model-based two-way outlier detection method in which a latent class model component for capturing two-way outliers is built upon an IRT-type baseline model component specified for standard item response behaviour free from outliers. Depending on whether response time data and external covariates are used, three specifications of the two-way outlier detection model have been proposed in the thesis: the reduced model for item responses, the full model for item responses and response times, and the explanatory model for item responses and observed covariates. Aside from incorporating covariate effects on latent class indicators and distributions of latent parameters, the explanatory model also relaxes the assumption of the independence between latent class indicators and latent parameters.

We have proposed a parallel-tempering MCMC algorithm to carry out Bayesian inference for the hierarchical two-way outlier detection models. The reduced model, the full model and the explanatory model were applied to a licensure test dataset known to suffer from item pre-knowledge and containing response times and covariate information. To formulate simultaneous detection of individuals with preknowl-

edge and compromised items in a Bayesian decision framework, we further proposed compound decision rules that control the local false discovery rate for persons or local false non-discovery rate for items. The classification of persons and items under the three models are close to the labels provided by the dataset, suggesting the proposed models provide accurate detection results, although the validity of the proposed models remains to be further checked through extensive applications to data gathered from many other educational tests in the future.

We used simulation studies to investigate the performance of the reduced model, the full model and the explanatory model under varying conditions of sample and item sizes, model misspecifications, and strength of the outlier effect indicated by the drift parameter. It is concluded that the three proposed models are robust against most of the model misspecifications. The classification tends to be more accurate as the sample and item sizes increase and when the signals of two-way outliers are stronger. The two-way classification and detection results were compared across the three proposed models under the same simulation settings. The results indicate that the inclusion of response time data or covariate effects on the latent class memberships and the distribution of latent parameters has improved the classification and better controlled the FDP and FNP.

In terms of future research on model-based two-way outlier detection, we have several directions in mind. First, the proposed two-way outlier detection method is designed to tackle one type of two-way outliers due to latent DIF, e.g., test takers with prior knowledge of compromised items. In the case of multiple sources of outliers, or multiple types of cheating in an educational testing example, a two-way outlier detection method is needed to be developed.

The second research area is concerned with missing data. If a great proportion of item responses are missing in observed data, we need to investigate whether the missing data mechanism is related to covariates and latent class memberships of individuals and items. As missing data might be informative for the detection of two-way outliers, the model-based two-way outlier detection method can be generalised

by the inclusion of covariates and latent class memberships of persons and items in a missing data model.

Another future development is to evaluate the effect of covariate inclusion on the detection of two-way outliers, particularly when there are a large number of external variables in the dataset. The variable selection procedure under a Bayesian framework is needed to determine which covariates should be used to estimate the distributions of latent parameters and latent class memberships of individuals and items. The relevant idea of Bayesian feature selection is discussed in Section 4.6.

References

Acuna, E., & Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*, *1*, 25.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of educational and behavioral Statistics*, *22*(1), 47–76.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*, 3099–3132.

Atchadé, Y. F., Roberts, G. O., & Rosenthal, J. S. (2011). Towards optimal scaling of metropolis-coupled markov chain monte carlo. *Statistics and Computing*, *21*, 555–568.

Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, *89*(428), 1329–1339.

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. Springer Science & Business Media.

Atkinson, A. C., & Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics*,

15(2), 460–476.

Atkinson, A. C., & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis*, *52*(1), 272–285.

Atkinson, A. C., & Riani, M. (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japan Statistical Society*, *38*(1), 3–14.

Atkinson, A. C., Riani, M., & Cerioli, A. (2013). *Exploring multivariate data with the forward search*. Springer Science & Business Media.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171–187.

Barnett, V., & Lewis, T. (1984). Outliers in statistical data. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. West Sussex, UK: John Wiley & Sons.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*(1), 1–15.

Bartholomew, D. J., Steele, F., & Moustaki, I. (2008). *Analysis of multivariate social science data*. CRC press.

Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association*, *94*(447), 947–955.

Belov, D. I. (2013). Detection of test collusion via Kullback–Leibler divergence. *Journal of Educational Measurement*, *50*, 141–163.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an ex-

aminee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397 – 472). Oxford, England: Addison-Wesley.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331–348.

Boughton, K. A., & Yamamoto, K. (2007). A hybrid model for test speededness. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). New York, NY: Springer.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, *58*, 1–37.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Boca Raton, FL: Chapman & Hall.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*, 83–87.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*, 957–970.

Cerioli, A., & Riani, M. (1999). The ordering of spatial data and the detection of multiple outliers. *Journal of computational and graphical statistics*, *8*(2), 239–258.

Chen, Y., Lee, Y.-H., & Li, X. (2020). Item quality control in educational testing: Change point model, compound risk, and sequential detection. *arXiv preprint arXiv:2008.10104*.

Chen, Y., & Li, X. (2019). Compound sequential change point detection in multiple data streams. *arXiv preprint arXiv:1909.05903*.

Chen, Y., Lu, Y., & Moustaki, I. (2022). Detection of two-way outliers in multivariate data and application to cheating detection in educational tests. *Ann. Appl. Statist.*, *16*(3), 1718-1746. doi: 10.1214/21-AOAS1564

Cho, S.-J., Suh, Y., & Lee, W.-y. (2016). An NCME instructional module on latent

DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*, 48–61.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Routledge.

Cizek, G. J., & Wollack, J. A. (2016). *Handbook of quantitative methods for detecting cheating on tests.* Taylor & Francis.

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests.* New York, NY: Routledge.

Clogg, C. C. (1995). Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). Springer.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*(388), 762–771.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Dai, Y. (2013). A mixture rasch model with a covariate: A simulation study via bayesian markov chain monte carlo estimation. *Applied Psychological Measurement*, *37*(5), 375–396.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. *Computer-based testing: Building the foundation for future assessments*, 165–191.

Davies, L., & Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, *88*(423), 782–792.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized*

linear and nonlinear approach. Springer Science & Business Media.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7–28.

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*, 41–66.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, *99*(465), 96–104.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, *23*, 1–22.

Efron, B. (2012). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge, UK: Cambridge University Press.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, *29*, 285–301.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*, 1151–1160.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Entink, R. K., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48.

European Social Survey. (2014). *Ess round 7: European social survey round 7 data (2014)*. Data file edition 2.2. NSD – Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. doi: 10.21338/NSD-ESS6-2012

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fraley, C., & Raftery, A. E. (1998). How many clusters? which clustering method?

answers via model-based cluster analysis. *The computer journal*, *41*(8), 578–588.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–534.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*, 1360–1383.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Geyer, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 295–311). Boca Raton, FL: Chapman & Hall/CRC.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*, 761–771.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American statistical association*, *88*(424), 1264–1272.

He, Q., Meadows, M., & Black, B. (2020). An introduction to statistical techniques used for detecting anomaly in test results. *Research Papers in Education*, 1–19.

Holland, P., & Wainer, H. (1993). *Differential item functioning*. New York, NY: Lawrence Erlbaum Associates.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural*

equation modeling: a multidisciplinary journal, *6*(1), 1–55.

Iglewicz, B., & Martinez, J. (1982). Outlier detection using robust measures of scale. *Journal of Statistical Computation and Simulation*, *15*(4), 285–293.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846). John Wiley & Sons.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In *Explanatory item response models* (pp. 189–212). Springer.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research,*, *36*(3), 347–387.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Katzgraber, H. G., Trebst, S., Huse, D. A., & Troyer, M. (2006). Feedback-optimized parallel tempering monte carlo. *Journal of Statistical Mechanics: Theory and Experiment*, *2006*, P03018.

Kelley, M. E., & Anderson, S. J. (2008). Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. *Statistics in medicine*, *27*(18), 3674–3688.

Kingston, N., & Clark, A. (2014). *Test fraud: Statistical detection and methodology.* New York, NY: Routledge.

Kuha, J., Katsikatsou, M., & Moustaki, I. (2018). Latent variable modelling with non-ignorable item nonresponse: multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *181*, 1169–1192.

Lawrance, A. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 181–189.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* New York, NY:

Houghton Mifflin Co.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in ecology and evolution*, *3*(1), 112–115.

Lu, Z.-H., Chow, S.-M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate behavioral research*, *51*(4), 519–539.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, *10*(1), 21.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180–197.

Mavridis, D., & Moustaki, I. (2008). Detecting outliers in factor analysis using the forward search algorithm. *Multivariate behavioral research*, *43*(3), 453–475.

Mavridis, D., & Moustaki, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of Computational and Graphical Statistics*, *18*(4), 1016–1034.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: a unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage.

McCutcheon, A. L. (2002). Basic concepts and procedures in single-and multiple-group latent class analysis. *Applied latent class analysis*, 56–88.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection

of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*, 121–137.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*(3), 391–411.

Moustaki, I., & Victoria-Feser, M.-P. (2006). Bounded-influence robust estimation in generalized linear latent variable models. *Journal of the American Statistical Association*, *101*, 644–653.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*, 177–194.

O'Leary, L. S., & Smith, R. W. (2017). Detecting candidate preknowledge and compromised content using differential person and item functioning. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 151–163). New York, NY: Routledge.

Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. *Applied psychological measurement*, *42*(5), 376–392.

Pison, G., Rousseeuw, P. J., Filzmoser, P., & Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, *84*, 145–172.

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*, 887–902.

Poole, K. T., & Rosenthal, H. (1991). Patterns of congressional voting. *American Journal of Political Science*, *35*, 228–278.

Poole, K. T., Rosenthal, H., & Koford, K. (1991). On dimensionalizing roll call votes in the US Congress. *The American Political Science Review*, *85*, 955–976.

Quintero, A., & Lesaffre, E. (2018). Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine*, *37*, 2440–2454.

Ramsay, J., & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, *56*, 365–379.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen, Denmark: Nielsen and Lydiche.

Reckase, M. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. *Psychometrika*, *61*(3), 509–528.

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 85–107.

Riani, M. (2004). Extensions of the forward search to time series. *Studies in Nonlinear Dynamics & Econometrics*, *8*(2).

Riani, M., & Atkinson, A. C. (2001). A unified approach to outliers, inuence, and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics*, *10*(3), 513–544.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: series B (statistical methodology)*, *71*(2), 447–466.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with

an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*, 731–792.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In J. Neyman (Ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability.* Berkeley, CA: University of California Press.

Robbins, H. (1956). An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (pp. 157–163). Berkerly, CA: University of California Press.

Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, *16*, 351–367.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Rost, J., & von Davier, M. (1995). Mixture distribution rasch models: Foundations, recent developments, and applications. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 257–268). New York, NY: Springer.

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2).

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research*, *27*(7), 1745–1755.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*(2), 461–464.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, *27*, 163–179.

Shao, J. (2003). *Mathematical statistics*. New York, NY: Springer.

Shu, Z., Henson, R., & Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, *78*, 481–497.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Simha, A., & Cullen, J. B. (2012). A comprehensive literature review on cheating. *International Journal of Cyber Ethics in Education (IJCEE)*, *2*(4), 24–44.

Sinharay, S. (2017a). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics*, *42*, 46–68.

Sinharay, S. (2017b). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Applied Psychological Measurement*, *41*, 403–421.

Skorupski, W. P., & Wainer, H. (2017). The case for Bayesian methods when investigating test fraud. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 214–231). New York, NY: Routledge.

Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed rasch models. *Methods of Psychological Research Online*, *4*(3), 19–32.

Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online*, *5*(4), 31–43.

Solaro, N., & Pagani, M. (2007). The forward search for metric multidimensional scaling. In *Convegno cladag 2007* (pp. 381–384).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *76*, 485–493.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173–180.

Sun, W., & Cai, T. T. (2007). Oracle and adaptive compound decision rules for

false discovery rate control. *Journal of the American Statistical Association*, *102*, 901–912.

Thissen, D., Steinberg, L., & Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models.* Lawrence Erlbaum Associates, Inc.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association.

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

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2015). Assessing model fit in latent class analysis when asymptotics do not hold. *Methodology*.

van Kollenburg, G. H., Mulder, J., & Vermunt, J. K. (2018). The lazy bootstrap. a fast resampling method for evaluating latent class model fit. *arXiv preprint arXiv:1801.09519*.

Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, *25*, 373–389.

von Davier, M., & Lee, Y.-S. (Eds.). (2019). *Handbook of diagnostic classification models.* New York, NY: Springer.

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the rasch model. *International Scholarly Research Notices*, *2013*.

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). Irt modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, *39*(8), 583–597.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144–168.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and

response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 223–254.

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501.

Wang, C., Zheng, Y., & Chang, H.-H. (2014). Does standard deviation matter? using "standard deviation" to quantify security of multistage testing. *Psychometrika*, *79*(1), 154–174.

Wang, X., & Liu, Y. (2020). Detecting compromised items using information from secure items. *Journal of Educational and Behavioral Statistics*, *45*, 667–689.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research*, 281–324.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. *Assessment of competencies in educational contexts*, 91–120.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, *12*(1), 58.

Wollack, J. A., & Fremer, J. J. (Eds.). (2013). *Handbook of test security.* New York, NY: Routledge.

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education*, *14*(4), 465–473.

Yuan, K.-H., & Bentler, P. M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *51*, 63–88.

Yuan, K.-H., & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *54*, 161–175.

Yuan, K.-H., Fung, W. K., & Reise, S. P. (2004). Three mahalanobis distances and their role in assessing unidimensionality. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 151–165.

Yuan, K.-H., & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological methods*, *15*(4), 335.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *The Annals of Statistics*, *31*, 379–390.

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Applied Psychological Measurement*, *38*, 87–104.

Zhou, Z., Li, X., Wright, J., Candès, E., & Ma, Y. (2010). Stable principal component pursuit. In *2010 ieee international symposium on information theory* (pp. 1518–1522).

Appendices

A. Simulation Studies in Chapter 2

In Chapter 2, we describe the FS algorithm for detecting outlying response patterns, latent classes and items lacking measurement invariance in the latent variable modelling framework. The R code for the FS Algorithm applied to simulation examples in Section 2.3 can be found at `https://github.com/YanLu-stats/FSLVM`.

B. Proof of Proposition 1

Proof. We let $p_{(1)} < \cdots < p_{(n)} \in (0,1)$ be all the distinct values for $P(\xi_i = 1|\boldsymbol{Y})$, $i = 1, ..., N$, where n is less than or equal to N as there might be ties. We further let $p_{(0)} = 0$ and $p_{(n+1)} = 1$. Then by the form of local FDR in (3.3.3), it is easy to verify that $\text{fdr}_\zeta(\boldsymbol{Y})$ is a step function of ζ, where $\text{fdr}_\zeta(\boldsymbol{Y})$ is a constant in interval $[p_{(t-1)}, p_{(t)})$, for any $t = 1, ..., n+1$. Therefore, $\text{fdr}_\zeta(\boldsymbol{Y})$ is left continuous in ζ.

We further show that $\text{fdr}_\zeta(\boldsymbol{Y}) > \text{fdr}_{\zeta'}(\boldsymbol{Y})$, when $\zeta \in [p_{(t-1)}, p_{(t)})$ and $\zeta' \in [p_{(t)}, p_{(t+1)})$, for any $t = 1, ..., n$. When $t < n$, we have

$$\text{fdr}_{\zeta'}(\boldsymbol{Y}) = \frac{\sum_{i=1}^N (1 - P(\xi_i = 1|\boldsymbol{Y})) 1_{\{P(\xi_i=1|\boldsymbol{Y}) \geqslant p_{(t+1)}\}}}{\sum_{i=1}^N 1_{\{P(\xi_i=1|\boldsymbol{Y}) \geqslant p_{(t+1)}\}}},$$

and

$$\text{fdr}_{\zeta}(\boldsymbol{Y}) = \frac{\left(\sum_{i=1}^N (1 - P(\xi_i = 1|\boldsymbol{Y})) 1_{\{P(\xi_i=1|\boldsymbol{Y}) \geqslant p_{(t+1)}\}}\right) + \left(\sum_{i=1}^N (1 - p_{(t)}) 1_{\{P(\xi_i=1|\boldsymbol{Y}) = p_{(t)}\}}\right)}{\sum_{i=1}^N 1_{\{P(\xi_i=1|\boldsymbol{Y}) \geqslant p_{(t+1)}\}} + \sum_{i=1}^N 1_{\{P(\xi_i=1|\boldsymbol{Y}) = p_{(t)}\}}}.$$

As $1 - p_{(t)} > 1 - P(\xi_i = 1|\boldsymbol{Y})$ when $P(\xi_i = 1|\boldsymbol{Y}) \geqslant p_{(t+1)}$, $\text{fdr}_\zeta(\boldsymbol{Y}) > \text{fdr}_{\zeta'}(\boldsymbol{Y})$. When $t = n$, it is easy to see that $\text{fdr}_\zeta(\boldsymbol{Y}) > \text{fdr}_{\zeta'}(\boldsymbol{Y})$ as $\text{fdr}_{\zeta'}(\boldsymbol{Y}) = 0$. This completes the proof for the properties of $\text{fdr}_\zeta(\boldsymbol{Y})$. The proof for the non-decreasing property of $\text{fnr}_\zeta(\boldsymbol{Y})$ is similar and thus is omitted here.

By the left-continuity of $\text{fdr}_\zeta(\boldsymbol{Y})$, we have

$$\text{fdr}_{\zeta^*(\boldsymbol{Y};\rho)}(\boldsymbol{Y}) \leqslant \rho.$$

In addition, by the construction of $\zeta^*(\boldsymbol{Y};\rho)$, $\zeta' > \zeta^*(\boldsymbol{Y};\rho)$ for any $\zeta' \neq \zeta^*(\boldsymbol{Y};\rho)$ also satisfying $\text{fdr}_{\zeta'}(\boldsymbol{Y}) \leqslant \rho$. Then by the non-decreasing property of $\text{fnr}_{\zeta}(\boldsymbol{Y})$, $\text{fnr}_{\zeta'}(\boldsymbol{Y}) \geqslant \text{fnr}_{\zeta^*(\boldsymbol{Y};\rho)}(\boldsymbol{Y})$. Therefore, $\zeta^*(\boldsymbol{Y};\rho)$ solves the optimisation problem (3.3.5). □

Note that the posterior distributions in the compound decision framework are further conditional on person- and item-specific covariates, $\boldsymbol{X}$ and $\boldsymbol{Z}$, under the explanatory two-way outlier detection model proposed in Chapter 4: $P(\xi_i = 1|\boldsymbol{Y}, \boldsymbol{X})$ and $P(\eta_j = 1|\boldsymbol{Y}, \boldsymbol{Z})$.

C.1. Parallel Tempering MCMC Algorithm

As mentioned in Section 3.4.2, the standard MCMC algorithms, such as the Metropolis-Hastings algorithm, suffer from slow-mixing for our problem, due to the presence of many discrete variables and the interactions between these discrete variables in the current problem.

Let Ξ be a generic notation for the parameters and hyperparameters to be sampled. Note that $\Xi = \{\theta_i, \xi_i, \beta_j, \eta_j, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \delta : i = 1, ..., N, j = 1, ..., J\}$ for the reduced model, and $\Xi = \{\theta_i, \xi_i, \tau_i, \beta_j, \eta_j, \alpha_j, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \delta, \gamma, \kappa : i = 1, ..., N, j = 1, ..., J\}$ for the full model, respectively. Recall that $\tilde{\boldsymbol{Y}}$ is used as the generic notation for data, where $\tilde{\boldsymbol{Y}} = \boldsymbol{Y}$ and $(\boldsymbol{Y}, \boldsymbol{T})$ for the reduced model and the full model, respectively. We use $f(\Xi|\tilde{\boldsymbol{Y}})$ as a generic notation for the posterior distribution of interest. The goal is to sample Ξ from the target posterior distribution $f(\Xi|\tilde{\boldsymbol{Y}})$.

The algorithm involves sampling K MCMC chains with tempered target distributions. More specifically, let $0 < \psi_1 < \psi_2 < \ldots < \psi_K = 1$ be a pre-specified sequence of temperature levels. Then the kth chain has a target distribution $f_k(\Xi|\tilde{\boldsymbol{Y}}) \propto (f(\Xi|\tilde{\boldsymbol{Y}}))^{(1/\psi_k)}$, where $\propto$ means that the two sides differ by a normalising constant which does not depend on Ξ. The target distribution of the Kth chain is our target posterior distribution. Let t be the current iteration number and $\Xi^{k,t}$ be the current samples from the kth chain. The parallel tempering algorithm performs the

following steps in the $t+1$th iteration.

1. For each of the chains, sample $\Xi^{k,t+1}$ given $\Xi^{k,t}$ using a Metropolis-Hastings within Gibbs sampler, which will be further discussed below.

2. Randomly sample a pair of adjacent chains, k and $k+1$, and use a Metropolis-Hastings update to decide whether to swap the statuses of $\Xi^{k,t+1}$ and $\Xi^{k+1,t+1}$. That is, a Bernoulli random variable with success probability

$$\min\left\{1, \frac{f_k(\Xi^{k+1,t+1}|\tilde{\boldsymbol{Y}})f_{k+1}(\Xi^{k,t+1}|\tilde{\boldsymbol{Y}})}{f_k(\Xi^{k,t+1}|\tilde{\boldsymbol{Y}})f_{k+1}(\Xi^{k+1,t+1}|\tilde{\boldsymbol{Y}})}\right\}$$

   is generated to decide whether to swap or not. If the Bernoulli random variable takes value 1, then we swap the statuses of $\Xi^{k,t+1}$ and $\Xi^{k+1,t+1}$ and otherwise, we reject the swap and keep their statuses unchanged.

For simplicity, the MCMC sampling within each chain is conducted by using a Metropolis-Hastings within Gibbs sampler. That is, Ξ is split into multiple blocks. Each block is sampled given all the others, using a random-walk Metropolis-Hastings sampler, for which the step size is tuned following Roberts and Rosenthal (2001) that is based on the Metropolis-Hastings acceptance rate. For the reduced model, Ξ is split into 10 blocks, including (1) θ_i, $i = 1, ..., N$, (2) ξ_i, $i = 1, ..., N$, (3) β_j, $j = 1, ..., J$, (4) η_j, $j = 1, ..., J$, (5) δ, (6) π_1, (7) σ_{11}, (8) π_2, (9) μ_1, and (10) ω_{11}. For the full model, Ξ is split into 14 blocks, including (1) θ_i, $i = 1, ..., N$, (2) τ_i, $i = 1, ..., N$, (3) ξ_i, $i = 1, ..., N$, (4) β_j, $j = 1, ..., J$, (5) α_j, $j = 1, ..., J$, (6) η_j, $j = 1, ..., J$, (7) δ, (8) γ, (9) κ, (10) π_1, (11) Σ, (12) π_2, (13) $\boldsymbol{\mu}$, and (14) Ω.

The specification of the number and levels of the temperatures also needs some tuning. A fine-tuned system tends to have faster mixing. We suggest choosing the number and levels of the temperatures by following the theoretical guidance given in Atchadé et al. (2011) that is based on the Metropolis-Hastings acceptance rate.

The R code for the MCMC algorithm can be found on `https://github.com/YanLu-stats/OD2WIRT`.

C.2. Parallel Tempering Algorithm

Let Ξ be a generic notation for the parameters and hyper-parameters to be sampled. Note that $\Xi = \{\theta_i, \xi_i, \beta_j, \eta_j, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \delta : i = 1, ..., N, j = 1, ..., J\}$ for the explanatory model, where the hyper-parameters include $\boldsymbol{\nu}_1 = (\boldsymbol{a}, \boldsymbol{\lambda}, \boldsymbol{c}, \sigma^2)$ and $\boldsymbol{\nu}_2 = (\boldsymbol{b}, \boldsymbol{\phi}, \boldsymbol{d}, \omega^2)$. The goal is to sample Ξ from the target posterior distribution $f(\Xi|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$.

The algorithm involves sampling K MCMC chains with tempered target distributions. More specifically, let $0 < \psi_1 < \psi_2 < \ldots < \psi_K = 1$ be a pre-specified sequence of temperature levels. Then the kth chain has a target distribution $f_k(\Xi|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) \propto (f(\Xi|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}))^{(1/\psi_k)}$, where $\propto$ means that the two sides differ by a normalising constant which does not depend on Ξ. The target distribution of the Kth chain is our target posterior distribution. Let t be the current iteration number and $\Xi^{k,t}$ be the current samples from the kth chain. The parallel tempering algorithm performs the following steps in the $t+1$th iteration.

1. For each of the chains, sample $\Xi^{k,t+1}$ given $\Xi^{k,t}$ using a Metropolis-Hastings within Gibbs sampler, which will be further discussed below.

2. Randomly sample a pair of adjacent chains, k and $k+1$, and use a Metropolis-Hastings update to decide whether to swap the statuses of $\Xi^{k,t+1}$ and $\Xi^{k+1,t+1}$. That is, a Bernoulli random variable with success probability

$$\min\left\{1, \frac{f_k(\Xi^{k+1,t+1}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) f_{k+1}(\Xi^{k,t+1}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})}{f_k(\Xi^{k,t+1}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) f_{k+1}(\Xi^{k+1,t+1}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})}\right\}$$

   is generated to decide whether to swap or not. If the Bernoulli random variable takes value 1, then we swap the statuses of $\Xi^{k,t+1}$ and $\Xi^{k+1,t+1}$ and otherwise, we reject the swap and keep their statuses unchanged.

For simplicity, the MCMC sampling within each chain is conducted by using a Metropolis-Hastings within Gibbs sampler. For the explanatory model, Ξ is split into 10 blocks, including (1) θ_i, $i = 1, ..., N$, (2) ξ_i, $i = 1, ..., N$, (3) β_j, $j = 1, ..., J$,

(4) η_j, $j = 1, ..., J$, (5) δ, (6) a_0, (7) b_0, (8) $\boldsymbol{a}$, (9) $\boldsymbol{b}$, (10) $\boldsymbol{\lambda}$, (11) $\boldsymbol{\phi}$, (12) $\boldsymbol{c}$, (13) σ^2, (14) $\boldsymbol{d}$, and (15) ω^2. Each block is sampled given all the others, using a random-walk Metropolis-Hastings sampler, for which the step size is tuned following Roberts and Rosenthal (2001) that is based on the Metropolis-Hastings acceptance rate.

The R code for the MCMC algorithm can be found on `https://github.com/YanLu-stats/OD2WEIRT`..

D. Additional Simulation Study to Chapter 3

We provide an additional simulation study under settings similar to Study I in Section 3.6.1, but with smaller values of J to mimic educational tests with relatively smaller item sizes.

D.1. Settings

We consider two settings which are referred to as settings (D.S1) $N = 2,000$, $J = 50$, and (D.S2) $N = 4,000$, $J = 100$, respectively. The rest of the simulation setting is exactly the same as that of Study I in Section 3.6. For each simulation setting, 100 independent data sets are simulated from the full model. We apply both the reduced model for item responses and the full model for item responses and response times to these data sets.

D.2. Results

The analysis is conducted using the parallel tempering MCMC algorithm described above. For each data set, we run 10,000 iterations, with the first 3,000 iterations as the burn-in. The results are based on the posterior samples from the last 7,000 iterations.

The results are given in Tables *D.1* through *D.4*. The results are similar to those from Study I. We first examine the classification results. For each model and each

simulated data set, we classify the test takers based on the posterior means of ξ_i and evaluate the classification based on the AUC value of the corresponding ROC curve. The AUC values are shown in Table *D.1*. It can be observed that the classification is slightly more accurate under setting D.S2, due to the increased sample and item sizes. Moreover, the AUC values given by the full model tend to be slightly larger than those from the reduced model, thanks to the additional information from response times.

We further evaluate the proposed compound decision rules. For each data set, we control local FDR and local FNR at levels 1%, 5%, and 10% for test takers and items, respectively. We evaluate each decision rule by examining the resulting FDP and FNP. The results are given in Tables *D.2* and *D.3* for the classifications of test takers and items, respectively. According to these tables, the FDP is well-controlled for test takers and so is the FNP for items.

Finally, we show the results on the estimation of the global parameters, as these parameters have substantive interpretations in cheating detection. Specifically, we focus on the posterior mean estimator, for which bias and variance are estimated based on the results from 100 independent replications. These results are presented in Table *D.4*. The bias, in general, tends to be close to zero for all the global parameters from both models and both settings. In addition, the estimation tends to be more accurate under setting D.S2, due to the increased sample and item sizes.

	Test taker				Item			
	D.S1		D.S2		D.S1		D.S2	
AUC	Reduced	Full	Reduced	Full	Reduced	Full	Reduced	Full
25%	0.934	0.936	0.956	0.962	0.937	0.944	0.954	0.961
50%	0.955	0.954	0.960	0.971	0.964	0.963	0.972	0.976
75%	0.969	0.967	0.971	0.978	0.973	0.975	0.978	0.981

D.1 *Overall classification performance based on the posterior means of ξ_i and η_j. For each model, each setting, and each target (test taker/item), we show the 25%, 50%, and 75% quantiles of the AUCs of the corresponding ROC curves from 100 independent data sets.*

	D.S1						D.S2					
	Reduced			Full			Reduced			Full		
FDP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.008	0.032	0.089	0.008	0.035	0.085	0.007	0.028	0.072	0.008	0.029	0.069
50%	0.012	0.047	0.092	0.011	0.046	0.091	0.011	0.043	0.078	0.009	0.037	0.073
75%	0.013	0.051	0.098	0.013	0.053	0.099	0.013	0.047	0.084	0.012	0.044	0.087

D.2 *Local FDR control for test takers. For each model, each setting, and each local FDR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FDPs of the corresponding classifications from 100 independent data sets.*

	S1						S2					
	Reduced			Full			Reduced			Full		
FNP	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
25%	0.009	0.033	0.063	0.007	0.032	0.061	0.007	0.029	0.059	0.006	0.031	0.062
50%	0.012	0.037	0.068	0.009	0.036	0.067	0.009	0.038	0.067	0.007	0.036	0.065
75%	0.013	0.046	0.071	0.010	0.043	0.069	0.012	0.045	0.072	0.012	0.041	0.071

D.3 *Local FNR control for items. For each model, each setting, and each local FNR target (1%/5%/10%), we show the 25%, 50%, and 75% quantiles of the FNPs of the corresponding classifications from 100 independent data sets.*

D.S1	Reduced model						D.S2	Reduced model					
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ		π_1	π_2	σ_{11}	μ_1	ω_{11}	δ
Bias	0.11	-0.02	-0.05	-0.11	0.17	0.22	Bias	0.05	0.15	-0.04	-0.13	0.20	-0.07
Variance	0.16	0.09	0.32	0.43	0.17	0.30	Variance	0.21	0.11	0.39	0.25	0.22	0.22

D.S1	Full model											
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ
Bias	-0.11	-0.08	0.07	0.24	-0.08	0.11	-0.12	-0.04	-0.12	0.14	-0.08	-0.16
Variance	0.17	0.19	0.32	0.32	0.21	0.30	0.11	0.17	0.15	0.23	0.00	0.37
D.S2	Full model											
	π_1	π_2	σ_{11}	μ_1	ω_{11}	δ	σ_{22}	σ_{12}	μ_2	ω_{22}	ω_{12}	κ
Bias	-0.04	0.06	0.08	-0.15	-0.17	-0.15	-0.04	-0.11	-0.15	0.11	-0.04	-0.11
Variance	0.21	0.18	0.23	0.29	0.31	0.35	0.07	0.19	0.12	0.12	0.18	0.46

D.4 *Accuracy of the posterior mean estimator of the global parameters. The bias and variance for the posterior mean estimator are calculated based on the 100 replications.*