



Universiteit
Leiden
The Netherlands

Distance models for analysis of multivariate binary data

Worku, H.M.

Citation

Worku, H. M. (2018, December 20). *Distance models for analysis of multivariate binary data*. Retrieved from <https://hdl.handle.net/1887/67140>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/67140>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



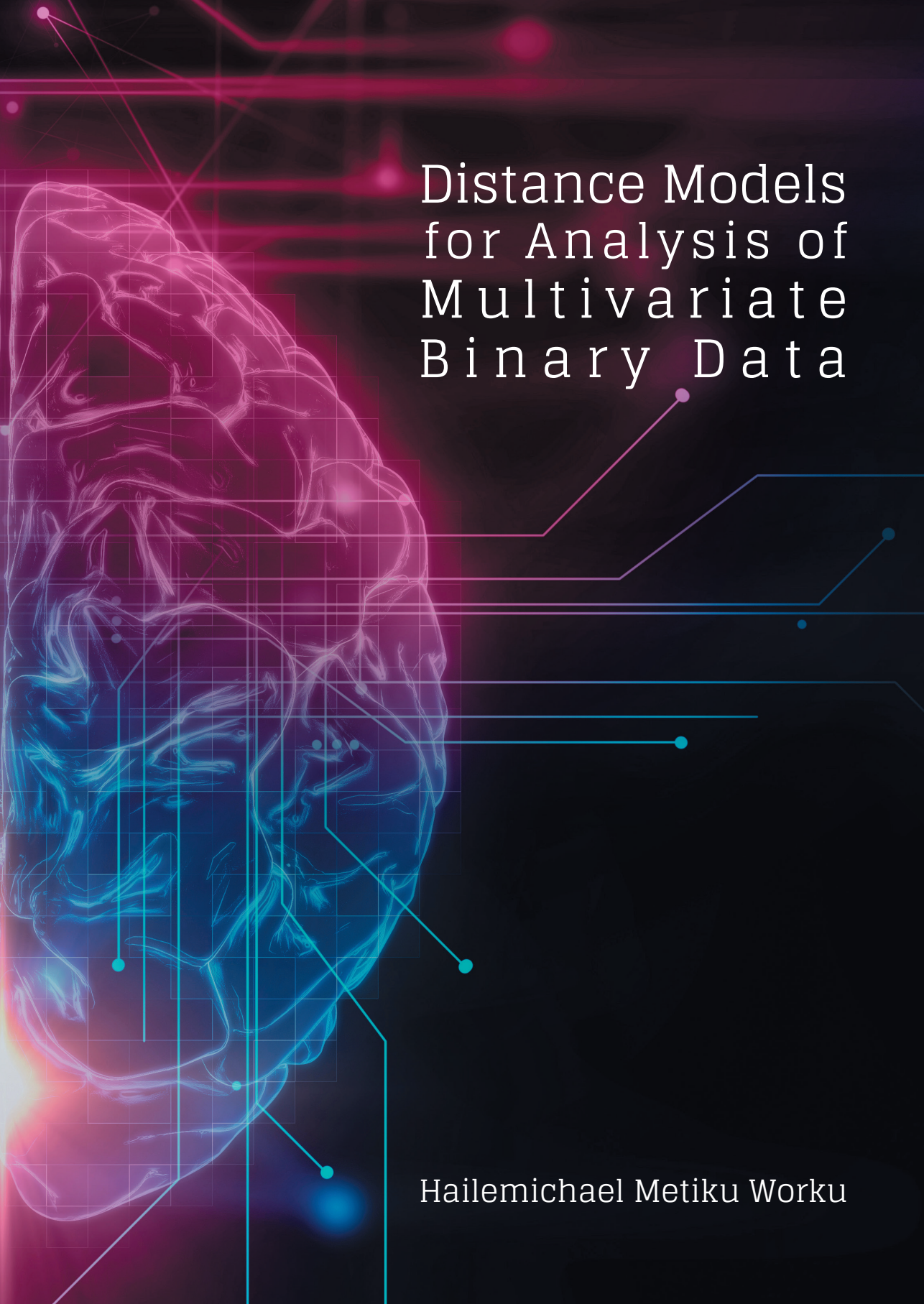
The following handle holds various files of this Leiden University dissertation:

<http://hdl.handle.net/1887/67140>

Author: Worku, H.M.

Title: Distance models for analysis of multivariate binary data

Issue Date: 2018-12-20

The background is a dark, textured surface. On the left side, there is a glowing, translucent blue and purple brain-like structure. Overlaid on this and the rest of the background are various glowing lines in shades of blue and purple, some straight and some curved, resembling circuitry or data paths. There are also small, glowing dots at the ends of some lines.

Distance Models for Analysis of Multivariate Binary Data

Hailemichael Metiku Worku

Distance Models for Analysis of Multivariate Binary Data

Copyright © 2018 by Hailemichael Metiku Worku.

Printed by Ridderprint BV

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronically, mechanically, by photocopy, by recoding, or otherwise, without prior written permission from the author.

ISBN 978-94-6375-221-3

Distance Models for Analysis of Multivariate Binary Data

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus, prof. mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 20 december 2018
klokke 15.00 uur

door Hailemichael Metiku Worku
geboren op 11 sep 1984 te Arbaminch, Ethiopië

Promotors:

prof. dr. M. de Rooij

prof. dr. W. J. Heiser

Promotiecommissie:

prof. dr. C.J.F. ter Braak (Wageningen University & Research)

prof. dr. P. Spinhoven (Leiden University FSW)

prof. dr. S. le Cessie (Leiden University Medical Center)

Dr. W. Bergsma (London School of Economics)

Acknowledgement:

The research described in this thesis was funded by the The Netherlands Organization for Scientific Research (NWO) with grant number 400-09-384.

Contents

Research Articles xi

List of Figures xii

List of Tables xv

1 Introduction 1

1.1 Categorical Response Data 1

1.1.1 Binary Response Data 3

1.1.2 Multicategory Response Data 3

1.2 Explanatory variables 3

1.3 Logistic Regression Model 3

1.3.1 Binary Logistic Regression 4

1.3.2 Multinomial Logistic Regression 5

1.3.3 Parameter Estimation in Logistic Regression Models 6

1.4 Distance Models 7

1.4.1 Multidimensional Scaling 7

1.4.2 Multidimensional Unfolding 8

1.4.3 IPDA Model 10

1.4.4 IPC Model 11

1.5 Multivariate Binary Data 14

1.5.1	Bivariate Binary Data	17
1.6	Models for Multivariate Binary Data	19
1.6.1	Marginal Models	19
1.6.2	Latent Variable Modeling	20
1.7	Outline of the Thesis	24
2	Effects of a Small Number of Dichotomous Indicators in Latent Variable Modeling: A Simulation Study	27
2.1	Introduction	28
2.2	Issues with Factor Models for Multivariate Data	30
2.2.1	Indeterminacy of Factor Scores	30
2.2.2	Improper Solutions	30
2.2.3	Previous Studies	31
2.3	Monte Carlo Simulation Study	32
2.3.1	The Research Problem	32
2.3.2	Experimental Plan	33
2.3.3	Simulation	36
2.3.4	Estimation	37
2.3.5	Replication	38
2.3.6	Analysis of Output	38
2.4	Results	39
2.4.1	Experiment-I: Confirmatory Factor Analysis	39
2.4.2	Experiment-II: The MIMIC Model	48
2.5	Conclusion and Discussion	52
3	Properties of Ideal Point Classification Models for Bivariate Binary Data	55
3.1	Introduction	56
3.2	Background	60

3.2.1	The Ideal Point Classification Model	60
3.2.2	The 2-step Approach of McCullagh and Nelder (1989)	62
3.3	Study-1: IPC Model as a Marginal Model	64
3.3.1	The 2-dimensional IPC Model	65
3.3.2	The 3-dimensional IPC Model	67
3.3.3	Discussion	69
3.3.4	Simulation Study	70
3.3.5	Summary of Study-1	74
3.4	Study-2: The Bivariate IPC Model	74
3.4.1	Simulation Study Results	77
3.5	Application	78
3.5.1	The IPC Models	80
3.5.2	The BIPC Model	82
3.6	Conclusion and Discussion	84
4	A Multivariate Logistic Distance Model for the Analysis of Multiple Binary Responses	89
4.1	Introduction	91
4.2	Multivariate Logistic Regression in a Distance Framework	94
4.2.1	Logistic Regression as a Distance Model	94
4.2.2	Multivariate Extension of the Distance Model	96
4.2.3	Parameter Estimation	99
4.2.4	The Relationship of the MLD Model to a Marginal Logistic Regression model	100
4.2.5	Model Selection	102
4.2.6	Biplot for the Multivariate Logistic Distance Model	103
4.3	Application: The NESDA Data	105
4.4	Conclusion and Discussion	114

5	mldm: An R Package for Analyzing Multivariate Binary Data	119
5.1	Introduction	120
5.2	The Multivariate Logistic Distance Model	120
5.2.1	Parameter Estimation	121
5.3	The NESDA Data	123
5.4	The mldm Package	124
5.4.1	Accessing the NESDA Data	124
5.4.2	Model Specification and Fitting	126
5.4.3	The Biplot for MLD Model	134
5.4.4	Model Selection using QIC	135
5.5	Conclusion and Discussion	140
6	Conclusions and Discussions	143
	Appendices	148
	Bibliography	167
	Samenvatting	181
	Acknowledgments	185
	Curriculum vitae	187

Motto

⁷Though your beginnings were modest, your latter days will be full of
prosperity. (Job 8:7)

Research Articles

As presented below, the chapters of this dissertation are based on published (or to be submitted) articles.

- Chapter 3: Worku, H. M. & De Rooij, M. (2017). Properties of Ideal Point Classification Models for Bivariate Binary Data. *Psychometrika*, **82** (2), 308-328.
- Chapter 4: Worku, H. M. & De Rooij, M. (2018). A Multivariate Logistic Distance Model for the Analysis of Multiple Binary Responses. *Journal of Classification*, **35**, 1-23. <https://doi.org/10.1007/s00357-018-9251-4>
- Chapter 5: Worku, H. M. & De Rooij, M. **mldm**: An R package for Analyzing Multivariate Binary Data. The package can be retrieved from <https://github.com/workuhm1/mldm-package-github>.

List of Figures

1.1 MDS Model: A two-dimensional configuration of dissimilarity data with five objects (i.e., A, B, C, D and E). 8

1.2 MDU Model: A two-dimensional configuration of preference data with four subjects (i.e., s1, s2, s3 and s4) and five objects (i.e., A, B, C, D and E). 9

1.3 A path diagram of a CFA with six indicator variables represented by a square, and two latent variables represented by a circle. 21

1.4 A path diagram for a MIMIC model with two external variables that are represented by a square. 24

2.1 A path diagram of a factor model with six indicator variables represented by a square, and two latent variables represented by a circle. 34

2.2 A path diagram for a MIMIC model with two external variables that are represented by a square. 34

2.3 Interaction plot for Nonconvergence rate: The first three panels (from left to right) show interaction plot between the type of indicators and the number of indicators, the factor structure, and the sample size, respectively. The last panel is for the interaction between the number of indicators and the sample size. 42

2.4	Interaction plot for Heywood rate: The first four panels (from left to right) show the interaction between the type of indicators and the number of indicators, the factor structure, the correlation between underlying latents, and the sample size, respectively. The last two panels are for the interaction between the number of indicators and the factor structure and the sample size.	45
2.5	Interaction plot for Quality of Recovering Factors: The first panel shows two main effects for the type and number of indicators. The second panel shows the interaction between the type of indicators and the factor structure.	48
3.1	Biplot of the final BIPC model fitted on the NESDA data. The predictors neuroticism, represented by N; extraversion, by E; and education, by EDU. The bivariate binary responses are dysthymia, represented by DYST; and generalized anxiety disorder, by GAD. The class coordinates that correspond to the multinomial response variable, denoted by G, are also displayed. . .	84
4.1	Biplot of the final “distress-fear” model fitted on the NESDA data, where the first dimension is represented by three disorders (MDD, GAD and DYST) and the second dimension by two disorders (SP and PD). The plot is based on restrictions applied on the class points.	109
4.2	Representation of the binary response variables in the Euclidean space. . .	110
4.3	Variable axes representation of the predictor variables (i.e., N: Neuroticism, E: Extraversion, C: Conscientiousness, and EDU: EDUcation) in the Euclidean space.	111
5.1	Reading the NESDA data available in the mlm package.	124
5.2	Excerpt of the NESDA data that shows records belonging to the first two subjects.	125

5.3	Specification of an indicator matrix for the depression-anxiety model fitted on the NESDA data.	126
5.4	A two-dimensional representation of model formula for depression-anxiety model fitted on the NESDA data.	127
5.5	Application of the <code>mldm.fit</code> function for fitting the depression-anxiety model on the NESDA data.	128
5.6	Summary of the depression-anxiety model fitted on the NESDA data. . .	131
5.7	Application of the Clustered Bootstrap method with the MLD model. . .	132
5.8	Summary of the depression-anxiety model fitted on the NESDA data using the Clustered Bootstrap method.	134
5.9	Application of the <code>biplot()</code> function available in the mldm package. . .	135
5.10	The biplot for depression-anxiety model fitted on the NESDA data. . . .	135
5.11	Specification of an indicator matrix for candidate models with respect to dimensionality in the model.	137
5.12	Specification of model formula for a unidimensional MLD model.	137
5.13	Model selection in MLD model for dimensionality.	138
5.14	Model formula structure of the candidate MLD models.	139
5.15	Model selection in MLD model for explanatory variables.	140
C.1	The distribution of estimated factor scores obtained from the final 2-factor (fear-distress) model. The top panel representing the distribution of scores from the first factor (F1) before and after the inclusion of external variables, respectively; and, the bottom panel for those scores from the second factor (F2) before and after the inclusion of the external variables, respectively.	163

List of Tables

1.1	The structure of multivariate data in long format.	16
1.2	Cross-classification of measurements of a bivariate binary data observed on the i -th subject.	17
2.1	Classes of Latent Variable Models.	28
2.2	The design variables with their corresponding values (or ranges) that are considered in the Monte Carlo simulation study. BLR stands for Binary indicator variables with Low success Rates; and BMR for Binary indicator variables with Moderate success Rates.	35
2.3	Percentage of nonconvergence in CFA under different experimental settings. Each cell result is based on $R = 100$ simulated replications.	40
2.4	Percentage of <i>Heywood</i> cases in CFA under different experimental settings. Each cell result is based on $R = 100$ simulated replications.	44
2.5	Quality of Recovering the True Factor scores: Average correlation between the true and estimated factor scores of CFA, i.e., $\text{Corr}(\theta_1, \hat{\theta}_1) = \hat{\rho}_1$, under different experimental settings. Each cell represents the results of $R = 100$ simulated replications, except those models that were not identified due to improper solutions.	47

2.6	Observed type-I error rates for the relationship between X_3 and the first factor, γ_{31} . The values in bold represent 95% confidence interval excluding the nominal level of significance ($\alpha = 0.05$). The number of replications per cell differs because of improper solutions. Dashed lines indicate no valid results were obtained for that cell.	49
2.7	Observed power for the relationship between X_5 and the first factor, $\gamma_{51} = -0.30$. The number of replications per cell differ because of improper solutions. Dashed lines indicate no valid results were obtained for that cell.	51
3.1	Cross-classification of bivariate binary data observed from i -th subject. . .	58
3.2	Summarized results of the simulation study for studying the performance of the IPC model for analysing bivariate binary data. IPC(2D-FIXED) corresponds to the 2-dimensional IPC model with <i>fixed</i> class points, i.e., $\phi_1 = \phi_2 = 0$; IPC(3D) to the 3-dimensional IPC model; and IPC(2D-FREE) to the 2-dimensional IPC model with <i>free</i> class points.	73
3.3	Summarized results of the simulation study for studying the performance of the BIPC model for analysing bivariate binary data.	78
3.4	Parameter estimates with corresponding standard errors (between the parenthesis) obtained from the IPC and BIPC models fitted on the NESDA data. IPC(2D-IND) corresponds to the 2-dimensional IPC model with <i>fixed</i> class coordinates; IPC(2D-FREE) to the 2-dimensional IPC model with <i>free</i> class coordinates; and IPC(3D) to the 3-dimensional IPC model.	79
4.1	The structure of multivariate data in long format.	97
4.2	Results of fitting different MLD models to NESDA data. In the first block, dimensionality of the MLD model is assessed, and followed by variable selection in the second block.	106

4.3 Summarized results of the final “distress-fear” MLD model fitted on NESDA data. Restriction was applied on the class points, and thus it is a restricted MLD model. The reported standard errors are based on both sandwich and clustered bootstrap methods. The number of bootstraps, $B = 1000$. 108

4.4 Regression weights of the final unrestricted “distress-fear” MLD model fitted on NESDA data. The number of bootstraps used to obtain standard errors equals 1000. 113

A.1 Parameter estimates of the 2-way interaction logistic regression model fitted on the nonconvergence data. For simplicity, we denote the design variables as, a: type of indicators; b: number of indicators; c: factor structure; d: correlation between underlying latent variables; and, e: sample size. 149

A.2 Parameter estimates of the 2-way interaction logistic regression model fitted on the *Heywood* data. For simplicity, we denote the design variables as, a: type of indicators; b: number of indicators; c: factor structure; d: correlation between underlying latent variables; and, e: sample size. . . . 152

A.3 Effect size of the 2-way interaction ANOVA model fitted on the average correlations reported in Table 2.5. The design variables are denoted by letters, i.e., a: type of indicators; b: number of indicators; c: factor structure; d: correlation between latent variables; and e: sample size. . . 154

A.4 Observed power for the relationship between X_7 and the second factor, $\gamma_{72} = 0.10$. The number of replications per cell differ because of improper solutions. Dashed lines indicate no valid results were obtained for that cell. 155

A.5 Observed power for the relationship between X_4 and the second factor, $\gamma_{42} = 0.95$. The number of replications per cell differ because of improper solutions. Dashed lines indicate no valid results were obtained for that cell. 156

C.1	Fit statistics for the factor models fitted on the NESDA data.	160
C.2	Parameter estimates with the corresponding standard errors (S.E.) presented in parenthesis for the final 2-factor (fear-distress) model.	161

Chapter 1

Introduction

1.1 Categorical Response Data

In statistical analysis, we often explore and analyze a single variable or many variables depending on the research question at hand. A variable, sometimes referred to as a random variable, is a statistical quantity which can be measured or observed. The following are examples of a variable: age, gender, survival of a patient (i.e., survived or not survived), mental status (i.e., normal, mild, moderate, severe), marital status (single, married, divorced, widowed), temperature and humidity, carbon emission, etc.

As described by Agresti (2002, Chap. 1), a variable can be classified in different ways: (1) response (sometimes referred to as dependent or outcome) variable versus explanatory (sometimes referred to as independent or predictor) variable; (2) continuous variable versus discrete variable; (3) quantitative variable versus qualitative variable; and, (4) nominal variable versus ordinal variable. Except for the first classification, the criteria for the other classifications are based on the type of values or measurements a variable could take. Gender, for example, is a nominal variable because it takes a value which is either male or female. Gender is also a qualitative variable. Mental status, on the other hand, could be defined either as qualitative or quantitative depending on the research.

In the above example, mental status is defined as an ordinal qualitative variable since there is a natural ordering between values for severity of mental status. Both survival of a patient and marital status, in the above example, are nominal qualitative variables. Qualitative variables are sometimes referred to as categorical variables. Age, like mental status, could be defined either as a discrete quantitative variable (e.g., Age (in years) = 23, 24, 43, etc) or as a continuous quantitative variable (e.g., Age (in hours) = 1.5, 3.5, 8.0, etc) or as an ordinal qualitative variable (e.g., Age = young, middle, elderly). The other variables in the above example (i.e., temperature, humidity and carbon emission) are defined most of the time as continuous quantitative variables.

In regression analysis or Analysis of Variance (ANOVA), for example, we study the relationship between a response variable and one or more explanatory variable(s). The aim of such analysis is to understand the amount of change on a response variable when an explanatory variable changes by some amount (usually a unit change). For example, a researcher might be interested in the relationship between mental status and age. The hypothesis of her research could be that severity of mental status of a subject might be affected by age. In this case, the response variable is mental status and the explanatory variable is age. Another example where a response variable is continuous, is the relationship between level of temperature in a given area (or country) and the amount of carbon emission. In this case, the response variable is temperature and it is a continuous variable. Carbon emission is the explanatory variable since it has the potential to explain level of atmospheric temperature.

In this thesis, the focus is on categorical response variables (where the response variable takes discrete values, e.g., yes / no, cured / not cured, etc) and the relationship between one or more explanatory variable(s) and these response variables.

1.1.1 Binary Response Data

A binary response variable is a categorical variable whose values are binary (i.e., yes or no; 1 or 0; survived or not survived; passed or failed). In many areas of research binary response variables are collected. A clinical psychologist might be interested depression, $\text{depression}=1$ if a given subject in the study has a depression, otherwise $\text{depression}=0$ representing absence of depression. A cardiologist might be interested to predict the chance of a patient to survive after performing heart surgery (i.e., $\text{survival} = 1$ if a patient survived; $\text{survival} = 0$ otherwise).

1.1.2 Multicategory Response Data

A multicategory response variable is a categorical variable with more than two possible values. Mental and marital status are examples of multicategory response variable.

1.2 Explanatory variables

An explanatory variable is expected to influence the response variable of interest. A possible set of explanatory variables for mental status could be age, residence (i.e., rural or urban), life style (e.g., smoking status, physical exercise, etc), personality traits (e.g., neuroticism, extroversion), etc. In this dissertation the explanatory variables might be continuous or categorical.

1.3 Logistic Regression Model

Logistic Regression (LR) model is a statistical model used for analyzing categorical response data. LR model is a member of the family of Generalized Linear Models (GLMs) (Agresti, 2007, chap. 3). The GLM is a general framework that extends ordinary linear regression model for continuous response variable to other types of variables (e.g., cat-

egorical response variables, i.e., both binary and multicategorical variables). Our main focus in this thesis will be GLM for categorical response data.

A GLM has three parts: (1) a random component; (2) a systematic component; and, (3) a link function. The random component represents the distribution of the response variable. The systematic component represents a linear combination of the explanatory variables. The link function is the part which does the linking between the response and the explanatory variables. Below is the mathematical representation of GLM:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p, \quad (1.1)$$

where $\mu = E(Y)$ is the random component and it is the expected value of the distribution of response variable Y from the exponential family. The right-hand side of Eq. (1.1) represents the systematic part of GLM including the intercept (i.e., β_0) and the regression coefficients (i.e., $\beta_1, \beta_2, \dots, \beta_p$ corresponding to the p explanatory variables denoted by x). The link function is $g(\cdot)$ and it connects the random part (i.e., μ) to the systematic part (i.e., $\beta_0 + \beta^T \mathbf{x}$, where $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$).

1.3.1 Binary Logistic Regression

Binary logistic regression, sometimes referred to as simple logistic regression, is a GLM for binary response data (Agresti, 2007, chap. 4). Let y_i denote the observed value of a binary dependent variable Y for subject i , where $i = 1, 2, \dots, N$. Binary logistic regression models the probability of a “success” category conditional on the value of explanatory variables $\mathbf{x}_i$, $\Pr(y_i = 1 | \mathbf{x}_i) = \pi(\mathbf{x}_i)$, i.e.,

$$\pi(\mathbf{x}_i) = \frac{\exp(\beta_0 + \beta^T \mathbf{x}_i)}{1 + \exp(\beta_0 + \beta^T \mathbf{x}_i)}, \quad (1.2)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$.

The log-odds representation of the same binary LR model (1.2) is,

$$\text{logit}[\pi(\mathbf{x}_i)] = \beta_0 + \beta^\top \mathbf{x}_i, \quad (1.3)$$

where $\text{logit}[\pi(\mathbf{x}_i)] = \log [\pi(\mathbf{x}_i)/(1 - \pi(\mathbf{x}_i))]$. This representation of binary LR is similar to the Generalized Linear model presented in Eq. (1.1) where the link function is now the “logit” function with $\mu = \pi(\mathbf{x}_i) = \Pr(y_i = 1|\mathbf{x}_i)$.

1.3.2 Multinomial Logistic Regression

Multinomial LR model is a GLM for multicategory response data (Agresti, 2007, chap. 6). Let $G_i = k$ denote the observed value of a multicategory dependent variable G for subject i , where $i = 1, 2, \dots, N$.

The Multinomial Baseline-Category Logit (MBCL) model is a natural extension of binary logistic regression model to the case of a nominal categorical variable. The probability of the k -th category in MBCL model (i.e., $\Pr(G_i = k|\mathbf{x}_i) = \pi_k(\mathbf{x}_i)$) is defined as,

$$\pi_k(\mathbf{x}_i) = \frac{\exp(\beta_{0k} + \beta_k^\top \mathbf{x}_i)}{\sum_c \exp(\beta_{0c} + \beta_c^\top \mathbf{x}_i)}. \quad (1.4)$$

The log-odds representation of the MBCL model (1.4) becomes,

$$\text{logit}[\pi_k(\mathbf{x}_i)] = \beta_{0k} + \beta_k^\top \mathbf{x}_{ik}, \quad (1.5)$$

where $\text{logit}[\pi_k(\mathbf{x}_i)] = \log [\pi_k(\mathbf{x}_i)/\pi_b(\mathbf{x}_i)]$. The index b refers to the reference (or baseline) category against which other categories are compared with. Thus, there are $(C - 1)$ number of “logit” models in MBCL for a multicategory response variable, G , with C the number of categories.

Suppose a researcher would like to study people's preference for environment (or

location) to spend their weekend. The possible values of the response variable G could be: *stay at home*, *meet friends at their place*, *meet friends at a city center*, *travel to somewhere* (e.g., park, beach, museum, other cities), and *go to the gym*. Let $G_i = 0, 1, 2, 3, 4$ be the numerical representation of the possible values and to be used in the MBCL model, respectively. Suppose the main aim of the investigation is to estimate the probability of preference of people to spend the weekend out of their home. That is, the probability of going to the gym, the park, the beach, museum, and other cities. In this case, the reference/baseline category will be staying at home (i.e., $G_i = 0$).

1.3.3 Parameter Estimation in Logistic Regression Models

In logistic regression, parameters of the model (i.e., the intercept and the regression coefficients) are unknown and thus estimated from sample data. Maximum likelihood optimization is a standard method used for estimating the parameters of LR models.

The likelihood function is the probability of the sample data, expressed as a function of model parameters (Agresti, 2002, pp. 6). The likelihood function for a binary LR model assuming a binomial distribution is defined as (Agresti, 2002),

$$L(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^N \frac{n_i!}{y_i!(n_i - y_i)} \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{n_i - y_i}, \quad (1.6)$$

where n_i represents the number of trials and y_i represents the number of successes, and $\boldsymbol{\beta}$ is a concatenation of the intercept and the regression coefficients of the binary LR model. The maximum likelihood estimation technique optimizes the likelihood function (Eq. (1.6)). Similarly, the likelihood function of MBCL model is defined as (Agresti, 2002),

$$L(\mathbf{G}|\boldsymbol{\beta}) = \prod_{i=1}^N \left[\frac{n_i!}{\prod_c G_{ic}!} \prod_c \pi_c(\mathbf{x}_i)^{G_{ic}} \right]. \quad (1.7)$$

1.4 Distance Models

Multidimensional scaling (MDS) is a technique developed in the behavioral and social sciences for studying the structure of objects or people (Davison, 1983, pp. 1). MDS uses proximity between pairs of objects as an input for analysis.

The proximity data is either similarity or dissimilarity of objects. In similarity data, the higher value for the proximity measure represents more alike pairs of objects whereas in dissimilarity data, the higher value for proximity measure represents less alike pairs of objects. An example of the latter type of proximities would be flight times.

Other examples of proximity measures are the correlation coefficient and joint probabilities (Davison, 1983, pp. 1). We will show later in this thesis that it is possible to express logistic regression models (i.e., Eq. (1.2) and (1.4)) in terms of distance models. In that case, probability is a similarity measure. That is, the smaller the relative distance between a subject (or person) point and a category point, the larger the probability that the subject chooses that category.

1.4.1 Multidimensional Scaling

In MDS, the proximities are represented in terms of distances between points in a low dimensional space (Kruskal & Wish, 1978; Davison, 1983; Borg & Groenen, 2005). The Euclidean distance model for dissimilarity measures is defined as (Davison, 1983, pp. 3),

$$\delta_{tu} = \left[\sum_{m=1}^M (z_{tm} - z_{um})^2 \right]^{1/2}, \quad (1.8)$$

where z_{tm} is the coordinate of object t on dimension m ($m = 1, 2, \dots, M$). An example of MDS solution is shown in Figure 1.1 which is a two-dimensional configuration of five objects: A, B, C, D and E. Suppose we would like to know: (1) how dissimilar A and D are, and (2) how dissimilar A and C are. This question can be answered easily by imputing

object coordinates in Eq 1.8. That is, $\delta_{AD} = [(z_{A1} - z_{D1})^2 + (z_{A2} - z_{D2})^2]^{1/2} = [(6 - 3)^2 + (7 - 6)^2]^{1/2} = 3.16$. Similarly, $\delta_{AC} = [(z_{A1} - z_{C1})^2 + (z_{A2} - z_{C2})^2]^{1/2} = [(6 - 7)^2 + (7 - 3)^2]^{1/2} = 4.1$. Thus, object A is more similar to D than to object C. The MDS problem is the reverse of this calculation: it is to find the coordinates of the points given the proximities.

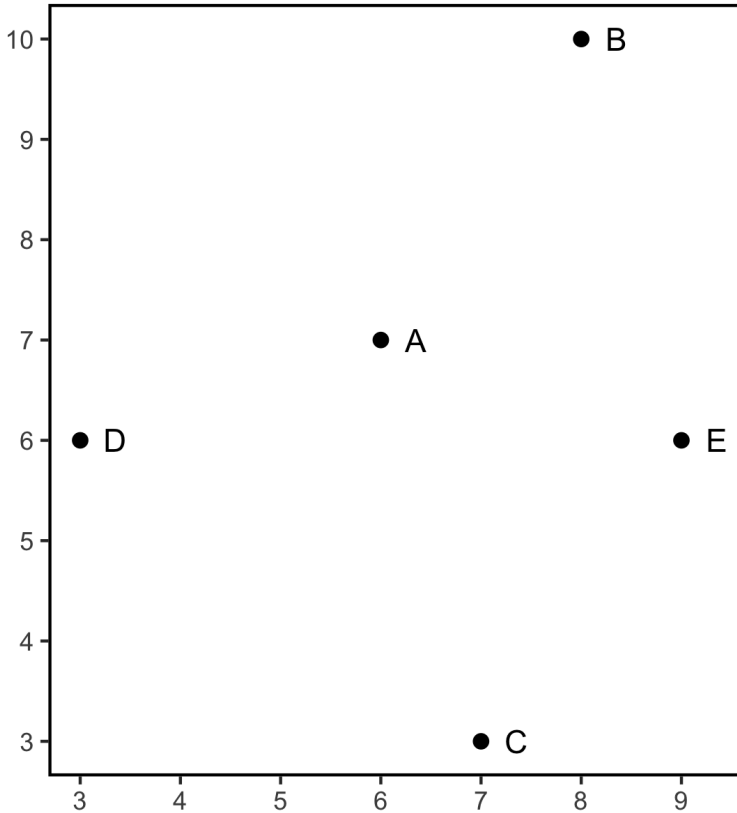


Figure 1.1: MDS Model: A two-dimensional configuration of dissimilarity data with five objects (i.e., A, B, C, D and E).

1.4.2 Multidimensional Unfolding

Coombs (1964) proposed a distance model for preference data, sometimes referred to as multidimensional unfolding (MDU) model. Preference data refers to proximity data

between a subject (usually a person) and an object (usually a product). For example, preference of students about study courses, preference of customers about set of product designs, preference of instructors about teaching methodology, etc. In this case, subjects are asked to rank their preference for a set of objects or stimuli.

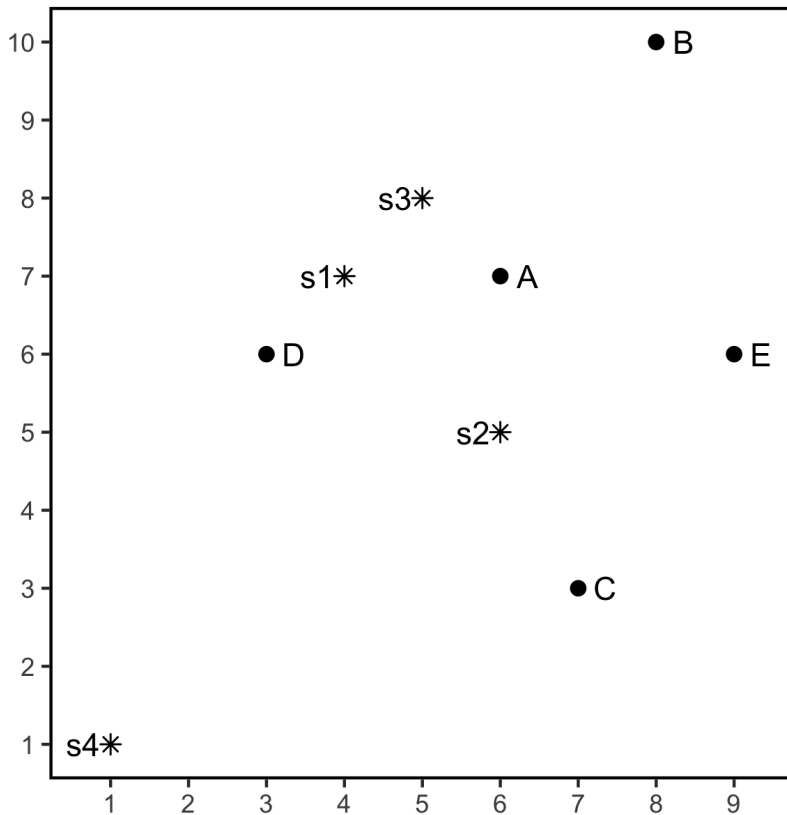


Figure 1.2: MDU Model: A two-dimensional configuration of preference data with four subjects (i.e., s1, s2, s3 and s4) and five objects (i.e., A, B, C, D and E).

The objective of MDU is to find distances in Euclidean space between subjects and objects that approximate a set of proximities as well as possible (Heiser, 1981, 1987; De Leeuw, 2005). An example of MDU is shown in Figure 1.2 which is the same configuration as Figure 1.1 with respect to the objects and with additional points for the subjects.

The position of the subjects are sometimes referred to as an *ideal* points of subjects.

The closer an object or stimulus to the ideal, the more it will be preferred (Davison, 1983, pp. 7). Suppose we would like to know which object (A or C) in Figure 1.2 most preferred by the fourth subject. This question can be answered by working out Eq 1.8. That is, $\delta_{S4,A} = [(z_{S4,1} - z_{A1})^2 + (z_{S4,2} - z_{A2})^2]^{1/2} = [(1 - 6)^2 + (1 - 7)^2]^{1/2} = 7.81$. Similarly, $\delta_{S4,C} = [(z_{S4,1} - z_{C1})^2 + (z_{S4,2} - z_{C2})^2]^{1/2} = [(1 - 7)^2 + (1 - 3)^2]^{1/2} = 6.3$. Thus, this subject prefers object C since the object is closer to its ideal position. Analogous to MDS, the unfolding problem is the reverse of this calculation: it is to find the coordinates of the object points and ideal points given the proximities between object and subjects.

1.4.3 IPDA Model

Takane, Bozdogan, and Shibayama (1987) proposed Ideal Point Discriminant Analysis (IPDA). The IPDA model is a multidimensional unfolding technique used for classification of subjects. The input data of IPDA model are not preference data but classification data, i.e., a given subject would choose one and only one object from a set of categories. The probability for the k -th category in the IPDA model is defined as (Takane, Bozdogan, & Shibayama, 1987),

$$\pi_k(\mathbf{x}_i) = \frac{m_k \exp(-\delta_{ik}^2)}{\sum_c m_c \exp(-\delta_{ic}^2)}, \quad (1.9)$$

where m_k is a bias parameter for category k which can be interpreted as a prior probability of the class, and δ_{ik}^2 is the squared Euclidean distance in an M -dimensional space between an ideal point for subject i with coordinates η_{im} and a class point for category k with coordinates γ_{km} (Takane et al., 1987), i.e.,

$$\delta_{ik}^2 = \sum_{m=1}^M (\eta_{im} - \gamma_{km})^2. \quad (1.10)$$

The ideal points are assumed to be a linear combination of the explanatory variables:

$$\boldsymbol{\eta}_i = \boldsymbol{\beta}_0 + \mathbf{x}_i \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a $(p \times M)$ matrix with regression weights and, $\boldsymbol{\beta}_0$ an M dimensional vector with intercepts. The parameters of this model are the regression weights and the class points. The class points, denoted as $\boldsymbol{\gamma}$, is a matrix of dimension $(C \times M)$.

The MBCL model, i.e., Eq. (1.4) and (1.5), is equivalent to the IPDA model in maximum dimensionality, i.e., $M = (C - 1)$ where C is number of categories or objects.

1.4.4 IPC Model

De Rooij (2009a) proposed the Ideal Point Classification (IPC) model. The IPC model is a probabilistic multidimensional unfolding model and closely related to the IPDA model.

As noted by Takane et al (1998), the interpretation of IPDA model is hampered by the bias parameters. De Rooij (2009a) showed that the bias parameters can be ignored without loss of information, except when (1) the response variable has many categories and a low-dimensional distance model is used; and (2) the response variable has a category that dominates the other categories. The probability for the k -th category in the IPC model is defined as (De Rooij, 2009a),

$$\pi_k(\mathbf{x}_i) = \frac{\exp(-0.5 * \delta_{ik}^2)}{\sum_c \exp(-0.5 * \delta_{ic}^2)}. \quad (1.11)$$

By looking at Eq. (1.9) and Eq. (1.11), it can be seen that IPC model is equivalent to the IPDA model without the bias parameters. The log-odds representation of the IPC model is,

$$\text{logit}[\pi_k(\mathbf{x}_i)] = 0.5 * \delta_{ib}^2 - 0.5 * \delta_{ik}^2, \quad (1.12)$$

where δ_{ib}^2 is the squared Euclidean distance between the b -th baseline category and the ideal point for subject i .

IPC Model for Binary Data

De Rooij (2009a) showed that logistic regression for a binary response variable, i.e., Eq. (1.2) and (1.3), can be expressed as an *unidimensional* IPC model. That is, a distance model in a joint space with points representing the two categories of the response variable and points representing the subjects.

The *unidimensional* IPC model of the binary response variable which is a simplification of Eq. (1.11) becomes,

$$\pi(\mathbf{x}_i) = \frac{\exp(-0.5 * \delta_{i1}^2)}{\exp(-0.5 * \delta_{i0}^2) + \exp(-0.5 * \delta_{i1}^2)}. \quad (1.13)$$

The class points of the *unidimensional* IPC model are given by $\gamma = \begin{bmatrix} \gamma_{01}, & \gamma_{11} \end{bmatrix}^T$, where γ_{01} is the class point of the baseline category (i.e., $Y = 0$), and γ_{11} is the class point of the “success” category (i.e., $Y = 1$). The log-odds representation of the *unidimensional* IPC model is,

$$\begin{aligned} \text{logit}[\pi(\mathbf{x}_i)] &= 0.5 * \delta_{i0}^2 - 0.5 * \delta_{i1}^2 \\ &= 0.5 * (\eta_{i1} - \gamma_{01})^2 - 0.5 * (\eta_{i1} - \gamma_{11})^2 \\ &= (\gamma_{11} - \gamma_{01}) * \eta_{i1} + 0.5 * (\gamma_{01}^2 - \gamma_{11}^2). \end{aligned} \quad (1.14)$$

With a restriction on class points for model identification (e.g., $\gamma = \begin{bmatrix} 0, & 1 \end{bmatrix}^T$), the *unidimensional* IPC model can be simplified to,

$$\text{logit}[\pi(\mathbf{x}_i)] = (\beta_0 - 0.5) + \beta^T \mathbf{x}_i. \quad (1.15)$$

Thus, the *unidimensional* IPC model is equivalent to the binary logistic regression presented in Eq. (1.2) and (1.3) and has the same regression coefficients (i.e., β) and an intercept with an offset of half (i.e., $\beta_0^{\text{IPC}} = \beta_0^{\text{LR}} + 0.5$).

IPC Model for Multicategory Data

As shown in Eq. (1.5), MBCL model is a natural extension of a simple LR model for nominal response variable. De Rooij (2009a) also showed that IPC model in a maximum dimensional space (i.e., $M = C - 1$) is equivalent to the MBCL model.

The log-odds representation of IPC model for a multicategory response variable is given in Eq. 1.12. By setting constraints on the class points, the IPC model can be identified uniquely. Suppose we have a multicategory response variable G with four categories such as $c = 0, 1, 2, 3$. For model identification, the class points in a maximum dimensional space ($M = 3$) can be represented as follows,

$$\gamma = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (1.16)$$

That is, the first category (probably the baseline) is positioned on the origin (i.e., $\gamma_{1m} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$), the second category is on the x -axis (i.e., $\gamma_{2m} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$), the third category is on the y -axis (i.e., $\gamma_{3m} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$), and the fourth category is on the z -axis (i.e., $\gamma_{4m} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$). With this class points configuration, it is possible to show that the IPC model is equivalent to the MBCL model. For demonstration purpose, let us see the derivation of the log-odds representation of the second category (i.e., $c = 1$)

against the baseline (i.e., $c = 0$). That is,

$$\begin{aligned}
 \text{logit}[\pi_1(\mathbf{x}_i)] &= 0.5 * \delta_{i0}^2 - 0.5 * \delta_{i1}^2 \\
 &= 0.5 * \sum_{m=1}^3 (\eta_{im} - \gamma_{0m})^2 - 0.5 * \sum_{m=1}^3 (\eta_{im} - \gamma_{1m})^2 \\
 &= \sum_{m=1}^3 (\gamma_{1m} - \gamma_{0m}) * \eta_{im} + 0.5 * \sum_{m=1}^3 (\gamma_{0m}^2 - \gamma_{1m}^2) \quad (1.17) \\
 &= \eta_{i1} - 0.5 \\
 &= (\beta_{01} - 0.5) + \beta_1^T \mathbf{x}_i.
 \end{aligned}$$

Similarly, the log-odds for the third category: $\text{logit}[\pi_2(\mathbf{x}_i)] = \eta_{i2} - 0.5 = (\beta_{02} - 0.5) + \beta_2^T \mathbf{x}_i$, and the log-odds for the fourth category: $\text{logit}[\pi_3(\mathbf{x}_i)] = \eta_{i3} - 0.5 = (\beta_{03} - 0.5) + \beta_3^T \mathbf{x}_i$. Thus, $\beta_p^{\text{IPC}} = \beta_p^{\text{MBCL}}$ for regression coefficients with dimension $(p \times M)$, and $\beta_0^{\text{IPC}} = \beta_0^{\text{MBCL}} - 0.5$ for intercepts with dimension $(1 \times M)$.

1.5 Multivariate Binary Data

In the previous sections, we considered only a single binary or multicategory response variable. However, it is not uncommon to see multiple binary/multicategory response variables in a given study. In medical science, for example, researchers are often interested not only on the efficacy of a newly developed drug, but also on the side effect of the drug. The explanatory variables in such a drug study setting could be the type of treatment (i.e., placebo, current drug, and newly developed drug), age, gender, etc. In this hypothetical study, there are two binary responses: efficacy (i.e., whether the subject is cured or not), and side effect (i.e., whether the drug has a side effect or not).

Multivariate binary data with multiple binary response variables and one or more explanatory variables, are often collected in empirical sciences such as psychology, criminology, epidemiology, life sciences and medicine. In the British coalminers study, for

example, researchers investigated impact of exposure to smoking and pneumoconiosis on two respiratory diseases, breathlessness (1 = yes; 0 = no) and wheeze (1 = yes; 0 = no), of coalminers in Britain (Ashford, Morgan, Rae, & Sowden, 1970; McCullagh & Nelder, 1989; Palmgren, 1989).

Another example of multivariate binary data is the Netherlands Study of Depression and Anxiety (NESDA). In NESDA, data were collected to investigate the interplay between personality traits and co-morbidity of depressive and anxiety disorders (Penninx et al., 2008; Spinhoven, De Rooij, Heiser, Penninx, & Smit, 2009). Co-morbidity is a presence of two or more mental disorders. In the area of mental disorders clinical psychologists and epidemiologists are interested in co-morbidity and how co-morbidity is related to risk factors such as personality traits and background variables (Krueger, 1999; Beesdo-Baum et al., 2009; Spinhoven, Penelo, De Rooij, Penninx, & Ormel, 2013). The NESDA data will be a leading example throughout this dissertation. We thank the NESDA consortium for providing the data.

Another study in which multivariate binary data arises is the Indonesian Children's Study (ICS: Sommer, Katz, & Tarwotjo, 1984; Liang, Zeger, & Qaqish, 1992) where over three-thousand children were medically examined to investigate whether they had respiratory infection, diarrhoeal infection, and xerophthalmia. The aim of the ICS study was to investigate whether vitamin A deficiency places children at increased risk of respiratory and diarrhoeal infections.

Suppose $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{iJ})^T$ denotes the multivariate responses observed on the i -th subject, which is a $(J \times 1)$ -dimensional vector of all responses. The y_{ij} represents a binary measurement of the j -th response variable observed on the i -th subject. In Table 1.1, we display the typical structure of such multivariate data in long format. The first column (Subject) contains subjects' identification number. The second column has binary measurements of the multivariate response variable. For demonstration purpose, we assume a total of five binary response variables that are measured for each subject.

The other columns in Table 1.1 have measurements for explanatory variables $X_1, X_2, \dots, X_p$.

Table 1.1: The structure of multivariate data in long format.

Subject	Response	Explanatory variables			
		X_1	X_2	X_p	
1	y_{11}	x_{11}	x_{12}	...	x_{1p}
1	y_{12}	x_{11}	x_{12}	...	x_{1p}
1	y_{13}	x_{11}	x_{12}	...	x_{1p}
1	y_{14}	x_{11}	x_{12}	...	x_{1p}
1	y_{15}	x_{11}	x_{12}	...	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	x_{i1}	x_{i2}	...	x_{ip}
i	y_{i2}	x_{i1}	x_{i2}	...	x_{ip}
i	y_{i3}	x_{i1}	x_{i2}	...	x_{ip}
i	y_{i4}	x_{i1}	x_{i2}	...	x_{ip}
i	y_{i5}	x_{i1}	x_{i2}	...	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	x_{n1}	x_{n2}	...	x_{np}
n	y_{n2}	x_{n1}	x_{n2}	...	x_{np}
n	y_{n3}	x_{n1}	x_{n2}	...	x_{np}
n	y_{n4}	x_{n1}	x_{n2}	...	x_{np}
n	y_{n5}	x_{n1}	x_{n2}	...	x_{np}

1.5.1 Bivariate Binary Data

Two cross classified binary variables observed on the i -th subject is displayed in Table 1.2. The rows represent measurements of the first binary response variable (y_{i1}), and the columns represent measurements of the second response variable (y_{i2}). In this Table, both marginal probabilities (shown in the margins, i.e., $\pi_{i1.}$, $\pi_{i0.}$, $\pi_{i.1}$, and $\pi_{i.0}$) and the joint probabilities (shown in the four cells, i.e., $\pi_{i,11}$, $\pi_{i,10}$, $\pi_{i,01}$, and $\pi_{i,00}$) are presented. The sum of probabilities either for the margins by row/column or for the individual cells always equals one.

Empirical researchers working with bivariate binary data are often interested in one of the following three parameters (Ashford et al., 1970; MacLean, Sofuoglu, & Rosenheck, 2018; Bhuyan, Islam, & Rahman, 2018): (1) the marginal probabilities; (2) the association between the two binary responses; or (3) the joint (or multinomial) probabilities.

Table 1.2: Cross-classification of measurements of a bivariate binary data observed on the i -th subject.

		y_{i2}		
		1	0	
y_{i1}	1	$\pi_{i,11}$	$\pi_{i,10}$	$\pi_{i1.}$
	0	$\pi_{i,01}$	$\pi_{i,00}$	$\pi_{i0.}$
		$\pi_{i.1}$	$\pi_{i.0}$	1.00

Joint Probabilities

The joint probability is an important quantity of bivariate binary data. In the Coalminers study, for example, let y_{i1} and y_{i2} denotes the measurements of breathlessness and wheeze of the coalminers, respectively. Then, the joint probability $\pi_{i,10}$ represents the probability of getting breathlessness, but no wheeze. Similarly, the joint probability $\pi_{i,01}$ represents the probability of getting wheeze, but no breathlessness. The other joint probabilities represents risk of getting both respiratory diseases ($\pi_{i,11}$), and the risk of getting none of

the diseases ($\pi_{i,00}$).

Bivariate binary data are special case of a multcategory response variable with four categories. Therefore, we can use a single index to represent the joint probabilities, i.e., $\pi_{ik}(\mathbf{x}_i) = \Pr(G_i = k|\mathbf{x}_i)$. For the joint probabilities in Table 1.2, this means: $\pi_{i1} = \pi_{i,00}$, $\pi_{i2} = \pi_{i,10}$, $\pi_{i3} = \pi_{i,01}$, and $\pi_{i4} = \pi_{i,11}$. Because of this relationship, logistic regression models for a multcategory response data such as the MBCL model (Eq. 1.4 and 1.5) and the IPC model (Eq. 1.11 and 1.12), can be used to analyze the joint probabilities of bivariate binary data.

Marginal Probabilities

The marginal probability of a bivariate binary data models a single response variable without controlling for measurements of the second response variable. Two separate simple logistic regression models (Eq. 1.2 and 1.3) can be used for this purpose, one for each response variable. In the Coalminers study, the marginal model can be used to answer a question about probability of breathlessness (wheeze) of coalminers due to exposure.

Association

The third quantity of interest is the association between the binary response variables. The association gives us information about the relationship of the two binary response variables. That is, it tells us whether the probability of occurrence of the second response variable increase/decrease when the probability of occurrence of the first response variable increases, and vice versa.

The most common measures of association structure for bivariate binary data are the odds (OR) ratio and the relative risk (RR). In this thesis, we use the OR as measure of association. The OR can also be modeled to investigate the impact of explanatory variables on the association structure (Lipsitz, Laird, & Harrington, 1990; Bahadur, 1961).

That is,

$$\log(\tau_i) = \beta_0 + \beta^\top \mathbf{x}_i, \quad (1.18)$$

where τ_i denotes the OR and is defined as $\tau_i = (\pi_{i4} \times \pi_{i1}) / (\pi_{i2} \times \pi_{i3})$.

1.6 Models for Multivariate Binary Data

The most common statistical modeling approach for analyzing multivariate binary responses in the presence of explanatory variables, are (1) marginal models (Agresti, 2002, Chap 11), and (2) latent variable models (Agresti, 2002, Chap 12). Marginal models are sometimes referred to as *population-averaged* models. Latent variable models are sometimes referred to as *random-effects* or *subject-specific* models.

1.6.1 Marginal Models

The availability of the multivariate normal distribution for multivariate interval responses, makes application of maximum likelihood-based statistical models relatively easy. However, for binary responses, there is no general parsimonious parameterization of the multivariate binary distribution, and therefore estimation becomes difficult (Agresti, 2002; Cox, 1972). Liang and Zeger (1986) proposed Generalized Estimating Equations (GEE) for marginal modelling of correlated categorical data. GEE is a quasi-likelihood (QL) estimation method that does not require specification of a particular multivariate distribution. It is widely used as a standard approach for fitting marginal models on multivariate data (Ziegler, Kastner, & Blettner, 1998; Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008; Ziegler, 2011).

1.6.2 Latent Variable Modeling

Latent variable models are a general class of models that are used for analyzing multivariate data (Bartholomew & Knott, 1999; Skrondal & Rabe-Hesketh, 2004). In Latent Variable (LV) models the multivariate response variables are treated as dependent variables, and one or more unobserved variables, referred to as latent variables, are treated as independent variables. The response variables are sometimes called indicators because they are used as an indirect measure of the latent variables.

The main application of LV models are: (1) for reducing the dimensionality of the multivariate data (to explain the variation of observed variables in few dimensions), (2) as measurement model (for representing a concept or construct that cannot be directly measured, e.g., depression, quality of life, political attitude, mathematical ability, intelligence, etc), and (3) for assigning scores on the latent scale which correspond to subjects' profile (Bartholomew, Steele, Moustaki, & Galbraith, 2002; Bollen, 2002; Rizopoulos, 2006). Tomarken and Waller (2005) provided a detailed literature review on Structural Equation Modeling (SEM) focusing on its strengths, limitations, and misconceptions.

Confirmatory Factor Analysis of Multivariate Data

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ij})$ be a j -dimensional vector of interval indicator variables observed on the i -th subject. The Confirmatory Factor Analysis (CFA) is based on the assumption that $\mathbf{y}_i$ can be attributed to q common factors, denoted by $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iq})$, and j unique factors (or measurement errors), denoted by $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{ij})$, with $j > q$

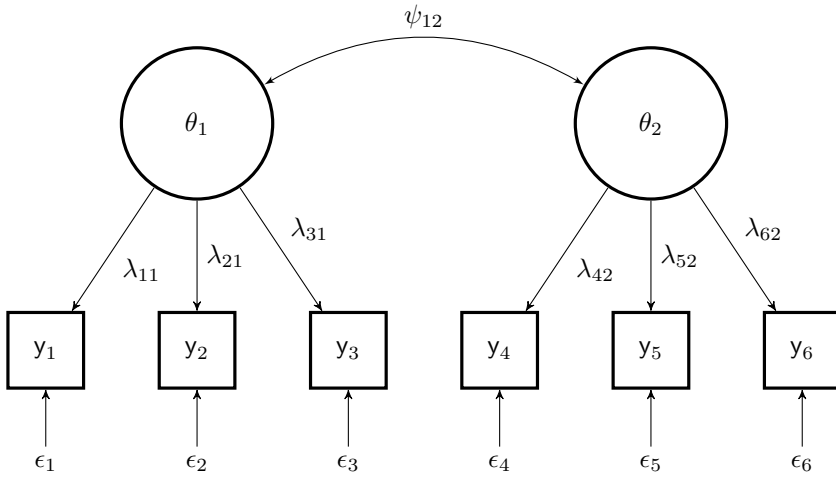


Figure 1.3: A path diagram of a CFA with six indicator variables represented by a square, and two latent variables represented by a circle.

(Thurstone, 1947; Jöreskog & Sörbom, 1981). The CFA is defined as,

$$y_{i1} = \lambda_{11}\theta_{i1} + \dots + \lambda_{1q}\theta_{iq} + \epsilon_{i1}$$

$$y_{i2} = \lambda_{21}\theta_{i1} + \dots + \lambda_{2q}\theta_{iq} + \epsilon_{i2}$$

$$\vdots$$

$$y_{ij} = \lambda_{p1}\theta_{i1} + \dots + \lambda_{jq}\theta_{iq} + \epsilon_{ij}$$

or, in matrix form

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \quad (1.19)$$

where $\mathbf{\Lambda}$ is the matrix of factor loadings. Let $\boldsymbol{\Psi}$ be the covariance matrix of common factors, and let $\boldsymbol{\Phi}$ be the covariance matrix of the unique factors. In Figure 2.1 an example of a path diagram is displayed which corresponds to a measurement model with six indicators ($j = 6$) and two underlying latent variables ($q = 2$).

In CFA, the common and unique latent variables follow multivariate normal distribu-

tions, i.e., $\theta \sim N_q(\mathbf{0}, \Psi)$ and $\epsilon \sim N_j(\mathbf{0}, \Phi)$, where Φ is a diagonal matrix. Given the model, the expected covariance matrix of the indicator variables becomes

$$\Sigma = \Lambda \Psi \Lambda^T + \Phi. \quad (1.20)$$

CFA for Multivariate Dichotomous Data

CFA was originally developed for modeling interval indicator variables. The covariance or correlation matrix of the observed variables was used as a primary object of analysis. The same method was later proposed for handling categorical (or dichotomous) indicator variables (Christoffersson, 1975; B. Muthen, 1978).

Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{iJ})$ be a J -dimensional vector of dichotomous indicator variables observed on the i -th subject. CFA of dichotomous variables assumes an underlying latent variable for each indicator variable, which is denoted by $\mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, \dots, y_{ij}^*, \dots, y_{iJ}^*)$. Thus, the variable y_{ij} equals one if its underlying latent variable y_{ij}^* is above a certain threshold value τ_j , otherwise it equals zero. Therefore, the measurement model for $\mathbf{y}_i$ is given by

$$\mathbf{y}_i^* = \Lambda \theta_i + \epsilon_i, \quad y_{ij} = \begin{cases} 1, & \text{if } y_{ij}^* \geq \tau_j, \\ 0, & \text{if } y_{ij}^* < \tau_j. \end{cases} \quad (1.21)$$

The formula for the covariance matrix remains the same, i.e., $\mathbf{V}(\mathbf{y}^*) = \Sigma$, but the elements in Φ matrix are not free parameters anymore, rather

$$\Phi = \mathbf{I} - \text{diag}(\Lambda \Psi \Lambda^T), \quad (1.22)$$

yielding $\text{diag}(\Sigma) = \mathbf{I}$. Therefore, the model has three sets of free parameters: τ , Λ , and Ψ (Christoffersson, 1975; B. Muthen, 1978).

Multivariate Regression with Latent Variables: The MIMIC Model

The measurement model is often not an ultimate step since researchers are interested in group differences and/or measurement invariance on the latent variables (Stapleton, 1978; Kenneth, 1989; T. Brown, 2006). This can be done by including external variables into CFA, and the new model becomes the Multiple Indicators Multiple Causes (MIMIC) model (Jöreskog & Goldberger, 1975; B. Muthen, 1983, 1984).

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ be the external variables observed on the i -th subject. Figure 2.2 shows the path diagram for a MIMIC model with two external variables connected to the two common latent variables. The MIMIC model extends the CFA model presented in (1.19) with relationships between the latent variables and the external variables, i.e.,

$$\begin{aligned} \mathbf{y}_i &= \Lambda \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\theta}_i &= \mathbf{\Gamma}^T \mathbf{x}_i + \boldsymbol{\zeta}_i, \end{aligned} \tag{1.23}$$

where $\mathbf{\Gamma}$ gives the regression coefficients, and $\boldsymbol{\zeta}$ the structural disturbances. It is assumed that the disturbances and the measurement errors are uncorrelated to each other and to $\mathbf{x}$, but not necessarily among themselves. The covariance matrix of the latent variables now becomes

$$\boldsymbol{\Psi} = \mathbf{\Gamma}^T \boldsymbol{\Sigma}_x \mathbf{\Gamma} + \boldsymbol{\Sigma}_\zeta,$$

where $\boldsymbol{\Sigma}_x$ is a covariance matrix for the external variables, and $\boldsymbol{\Sigma}_\zeta$ for the disturbances. For estimation and identification of the MIMIC model, we refer to Muthén (1983, 1984).

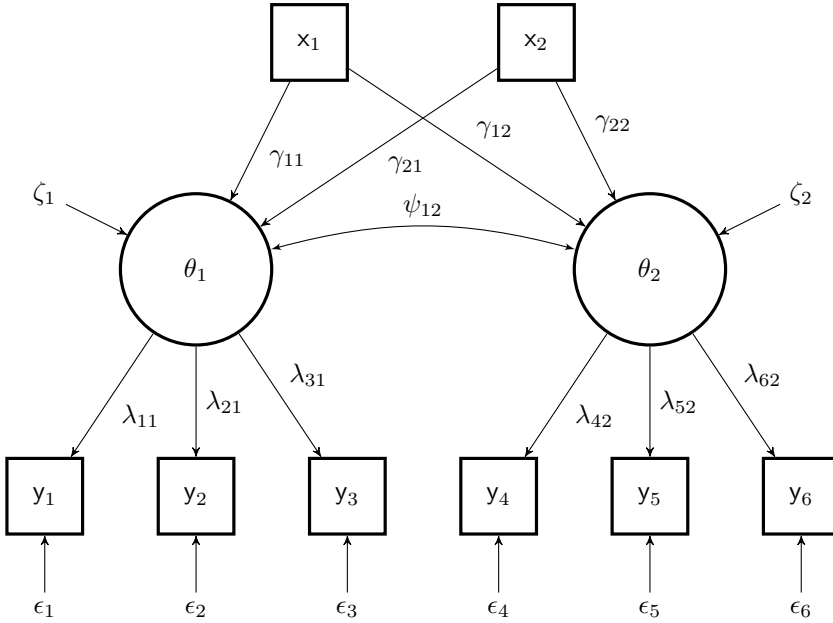


Figure 1.4: A path diagram for a MIMIC model with two external variables that are represented by a square.

1.7 Outline of the Thesis

Latent variable models are often used for analyzing multivariate binary data with and without the presence of explanatory variables. In Chapter 2 we investigate the performance of such models using a simulation study. We show the impact of the number of indicator variables, sample size, and type of indicator variables, on the performance of latent variable models.

In Chapter 3 we study properties of the IPC model for analyzing bivariate binary data. The main aim of this chapter is to investigate the potential of the IPC model in recovering three parameters of bivariate binary data: the marginal probabilities, joint probabilities, and association structure. A simulation study is used to evaluate the performance of the model. As the IPC model is not able to fully recover the three parameters, a Bivariate IPC (BIPC) model is proposed. The BIPC model is an adjusted form of the IPC model

to fully recover parameters of interest for bivariate binary data.

However, it is not straight forward to extend the BIPC model for the analysis of multivariate binary data. This is due to the fact that both the pairwise and higher-order association structure parameters must be specified in the likelihood function, and thus the computation becomes cumbersome. This issue will be addressed in Chapter 4 by developing a Multivariate Logistic Distance (MLD) model which is a new model for analyzing multivariate binary data. The MLD model unifies two domains of statistical methods, i.e., Multidimensional Scaling (MDS: Kruskal & Wish, 1978; Borg & Groenen, 2005) and Generalized Linear Model (GLM: McCullagh & Nelder, 1989; Agresti, 2002). As a form of regularization, the MLD model allows for dimension reduction and therefore less parameters are estimated compared to existing marginal models for multivariate binary data. Moreover, the model enhances interpretation by using a biplot (Gabriel, 1971; Gower & Hand, 1996; Gower, Lubbe, & Le Roux, 2011) based on a distance interpretation.

For this newly proposed distance model we developed an R package called **mldm**. Using an empirical dataset, usage of the package is demonstrated in Chapter 5. The package handles both the clustered bootstrap method and the sandwich estimators for obtaining standard errors of model parameters. It also provides a biplot function for the graphical representation of the fitted model. In Chapter 6 we conclude the thesis with a recommendation for future research.

Chapter 2

Effects of a Small Number of Dichotomous Indicators in Latent Variable Modeling: A Simulation Study

Abstract

Structural equation models were originally proposed for the analysis of continuous or interval indicator variables. Recently, factor analysis and structural equation models have been applied for data with dichotomous indicators and with only a few indicators per latent variable, i.e. 2 or 3. We investigated the performance of Confirmatory Factor Analysis (CFA) and the Multiple Indicators Multiple Causes (MIMIC) model for dichotomous indicators in comparison with interval indicators in a Monte Carlo simulation study.

The performance of both CFA and the MIMIC model was studied in terms of the quality of recovering the true factor scores and the incidence of improper solutions, more specifically non-convergence and *Heywood* cases. Furthermore, in the case of the MIMIC model, the focus was on the type-I error rate and power.

We showed that both CFA and the MIMIC model performed poorly with a small number of dichotomous indicators, i.e., (1) improper solutions occurred much more frequently; (2) the true factor scores are poorly recovered; (3) the type-I error rates are too conservative mostly and inflated sometimes; and (4) the observed power was weak.

2.1 Introduction

Latent variable models are a general class of models that are used for analyzing multi-variate data (Bartholomew & Knott, 1999; Skrondal & Rabe-Hesketh, 2004). In Latent Variable (LV) models the multivariate observed variables (manifest variables) are treated as dependent variable, and one or more unobserved variables (latent variables) are treated as independent variables. The observed variables are also known as indicators because they are used as an indirect measure of the latent variables.

The Latent variables can be interval or categorical. As displayed in Table 2.1, there are four classes of LV models based on the cross-classification of whether the observed variable and/or latent variable is interval and/or categorical (Bartholomew & Knott, 1999). Our main focus in this paper will be on Confirmatory Factor Analysis (CFA), which is a special case of Structural Equation Modeling (SEM: Thurstone, 1947; Jöreskog & Sörbom, 1981; Christoffersson, 1975; B. Muthen, 1978; Bock & Lieberman, 1970; Mislevy, 1986). Tomarken and Waller (2005) provided a detailed literature review on Structural Equation Modeling (SEM) focusing on its strengths, limitations, and misconceptions.

Table 2.1: Classes of Latent Variable Models.

Observed variable	Latent variable	
	Interval	Categorical
Interval	Factor Analysis/ Structural Equation Modeling	Latent Profile Analysis/ Mixture Modeling
Categorical	Item Response Theory/ Latent Trait Analysis	Latent Class Analysis

Traditionally SEM focuses on the analysis of continuous (or interval) indicator variables. Many studies have been performed to investigate the performance of structural

equation models (Boomsma, 1983, 1985; J. C. Anderson & Gerbing, 1984; Acito & Anderson, 1986). Recently, in clinical psychological research structural equation models have been proposed for the analysis of comorbidity of depressive and anxiety disorders (Krueger, 1999; Beesdo-Baum et al., 2009). A typical characteristic of these models is that the indicators are dichotomous, i.e. the indicators indicate whether someone has or does not have a particular disorder, and that there are only a few indicators per latent variable, i.e. 2 or 3. We believe the application of structural equation models in such a scenario (i.e., dichotomous indicators with a few number of variables per factor) is not adequate enough to obtain a valid result about the research question that we would like to answer. This is because with two indicator variables, there are only four patterns (i.e., (0, 0), (0, 1), (1, 0), and (1, 1)) with four corresponding observed proportions. Similarly, for three indicators, there will be eight patterns. Therefore, there is only limited information and it is hard to satisfy the normality assumption of the underlying latent variables in the structural equation model. However, we did not find large scale simulation studies that address our concerns. The aim of the current paper is to fill this gap. Therefore, we conducted a simulation study to investigate the performance of SEM for the analysis of a small number of dichotomous indicator variables per factor.

In our simulation study, we conducted two types of experiments. In the first experiment, the performance of Confirmatory Factor Analysis (CFA) as a measurement model is studied. The outcome variables of interest for this experiment are the incidence of nonconvergence, occurrence of *Heywood* cases, and the quality of recovering the true factor scores in CFA. In the second experiment, we study the performance of the Multiple Indicators Multiple Causes (MIMIC) model (Stapleton, 1978; Kenneth, 1989; T. Brown, 2006). In this case, the outcome variables of interest are the type-I error rate and the power of the statistical test for the regression coefficients in the MIMIC model. In both experiments we study the impact of five design variables on the outcome variables. The design variables are type of indicator variables (i.e., interval or categorical), number of

indicators, strength of factor structure, correlation between factor scores, and sample size.

The outline of this paper is as follows. In Section 2.2 we discuss issues with factor analysis and results found in the literature. The design and analysis of the simulation study is presented in Section 2.3. In Section 2.4, the results of the simulation studies are discussed. We conclude with a discussion of the results and some remarks for future research in Section 2.5.

2.2 Issues with Factor Models for Multivariate Data

2.2.1 Indeterminacy of Factor Scores

The indeterminacy of factor scores refers to a situation where the same indicator variables may produce different factor scores with the same model fit, and thus no unique solution does exist for the factor scores (Acito & Anderson, 1986; Guttman, 1955; Heermann, 1964, 1966; Schonemann, 1971; Schonemann & Wong, 1972; Green, 1976; Elffers, Bethlehem, & Gill, 1978). Some argue that the reason for the indeterminacy of factor scores is due to the presence of too many parameters compared to the number of equations in the model (Grice, 2001).

2.2.2 Improper Solutions

Factor analysis of multivariate data can sometimes produce improper solutions (Rindskopf, 1984; Boomsma, 1983, 1985; Kenneth, 1989; Chen, Bollen, Paxton, Curran, & Kirby, 2001). The most common improper solutions are nonconvergence and *Heywood* cases (Kenneth, 1989, pp. 282). *Heywood* cases occur when the estimated variances of a model become negative.

2.2.3 Previous Studies

Indeterminacy of factor scores in CFA has been studied by Acito and Anderson (1986). The impact of the number of indicators, factoring method, factor structure (i.e., the magnitude of factor loadings), number of factors, and sample size on indeterminacy of factor scores was investigated. Acito and Anderson found that both the factor structure and the factoring method have large effects on the indeterminacy of factor scores. A limitation of their study was that only interval indicators were considered. In the current study we also consider dichotomous indicators.

Improper solutions, i.e., nonconvergence and *Heywood* cases, in CFA has been studied using a Monte Carlo simulation by Anderson and Gerbing (1984) and by Boomsma (1985). Anderson and Gerbing studied the impact of sampling error and model characteristics on the incidence of improper solutions. Improper solutions occurred more frequently for smaller sample sizes and for models with fewer indicators for each factor (J. C. Anderson & Gerbing, 1984). Boomsma studied the impact of the number of indicators, correlation between factors, and factor structure. All of the design variables had a large effect on the incidence of improper solutions (Boomsma, 1985). Like the study by Acito and Anderson (1986), however, both studies considered only interval indicators. In this paper, we extend their study on improper solutions by including both interval and dichotomous indicator variables.

Marsh, Hau, Balla and Grayson (1998) performed a simulation study and studied extensively the impact of the number of indicators and sample size in a CFA on the occurrence of improper and nonconverged solutions, accuracy of parameter estimates, and goodness-of-fit indexes. Their main aim was to provide data driven evidence (contrary to rules of thumbs) for the number of indicators per factor and sample size to fit confirmatory factor models. They concluded that it is always good to have more indicators per factor and a larger sample size whenever possible. Similar studies were conducted by Ding,

Velicer, and Harlow (1995), Kenny and McCaouch (2003) and Marsh, Balla, and McDonald (1988), although the main focus of these studies was on measures of fit for factor models.

In general, all these studies investigated the impact of variables of interest on statistical properties of CFA when the indicator variables are interval. Categorical (or dichotomous) indicator variables were not considered. Furthermore, emphasis was given for factor models and the MIMIC model was not studied in a similar fashion. Our present simulation study fills these gaps since both issues are addressed.

2.3 Monte Carlo Simulation Study

We followed the six-step approach of Monte Carlo simulation design in structural equation modelling (Paxton, Curran, Bollen, Kirby, & Fen, 2001; Skrondal, 2000; Boomsma, 2013).

2.3.1 The Research Problem

Our main objective is to investigate the performance of SEM models, specifically Confirmatory Factor Analysis (CFA) and the MIMIC model, with only a few dichotomous indicators assumed per factor. We investigate the quality of recovering the true factor scores and the incidence of improper solutions (nonconvergence and *Heywood* cases). We study the impact of five design variables on the outcome variables. To have a benchmark to compare the performance of latent variable models for analyzing dichotomous indicator variables, we also consider interval indicator variables. We used Mplus statistical software package (L. Muthen & Muthen, 1998-2012) with the default estimation procedures to analyze our simulated datasets, because that is how most applied researchers analyze their data.

The design variables are type of indicator variables (i.e., interval or categorical), number of indicators, strength of factor structure, strength of correlation between latent variables, and sample size. We conducted two experiments, where in the first experiment

we study the performance of CFA and in the second experiment the performance of the MIMIC model.

2.3.2 Experimental Plan

The simulated data were generated from a 2-factor model whose path diagram is shown in Figure 2.1 for the CFA and in Figure 2.2 for the MIMIC model. In the path diagrams we have six observed variables y_j ($j = 1, 2, \dots, 6$), two latent variables θ_q ($q = 1, 2$), and unique factors indicated by ϵ_j . The model parameters in CFA are the factor loadings λ_{jq} and the covariance between the latent variables ψ_{12} .

In the case of the MIMIC model (Figure 2.2) explanatory variables ($x_k, k = 1, 2$) are added to the path diagrams. In addition to parameters of the factor model, the MIMIC model also has regression weights (γ_{kq}).

An equal number of indicator variables per factor was assumed in the data generation process. The variances of the factors were restricted to unity for identifiability of CFA. Table 2.2 shows the design variables considered in our Monte Carlo simulation and in the last column their corresponding values (or ranges) are given. The first design variable is type of indicators which is either dichotomous or interval. Two possible situations were considered for dichotomous indicator variables. The first case assumes a low success rate, i.e., between 5% – 15%. This case is denoted by BLR in Table 2.2 which stands for Binary indicators with Low success Rates. The other situation has moderate success rate (between 40% – 50%), and is denoted by BMR which stands for Binary indicators with Moderate success Rates.

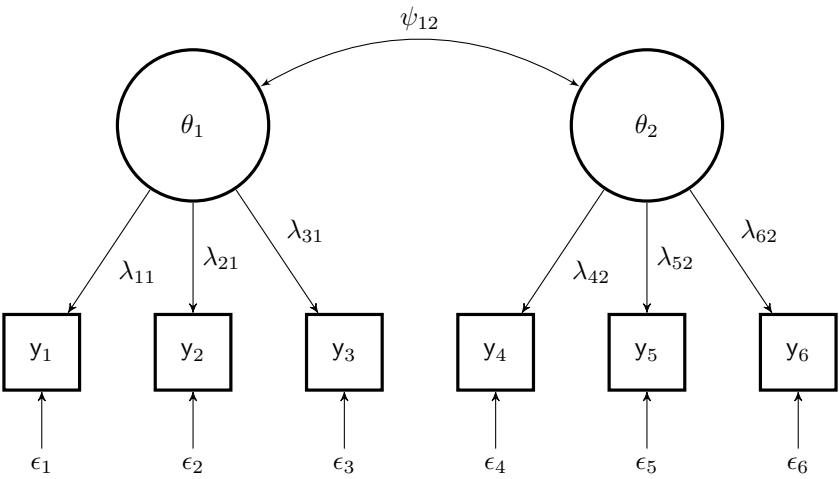


Figure 2.1: A path diagram of a factor model with six indicator variables represented by a square, and two latent variables represented by a circle.

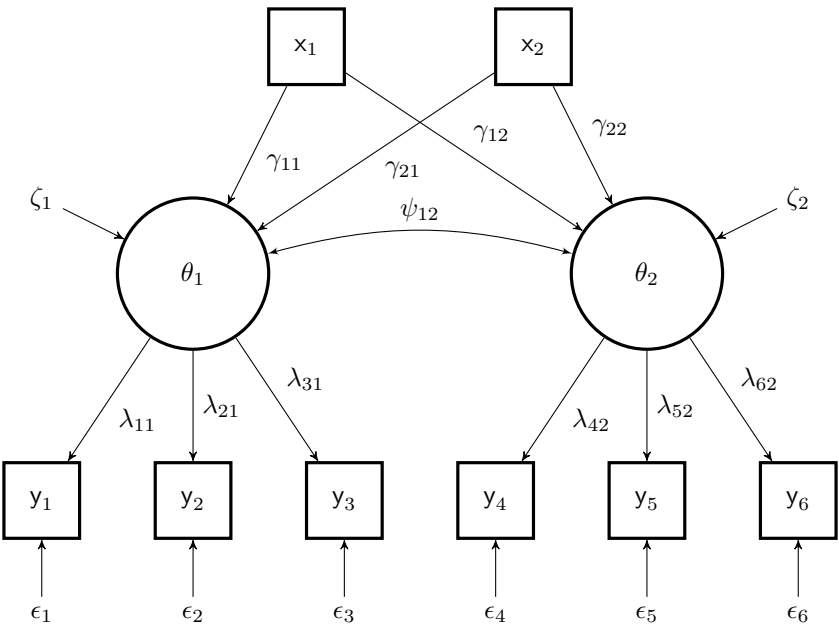


Figure 2.2: A path diagram for a MIMIC model with two external variables that are represented by a square.

Table 2.2: The design variables with their corresponding values (or ranges) that are considered in the Monte Carlo simulation study. BLR stands for Binary indicator variables with Low success Rates; and BMR for Binary indicator variables with Moderate success Rates.

Variable	Parameter	Level	Value/Range
Type of Indicators	—	BLR	5% – 15%
		BMR	40% – 50%
		Interval	—
Number of Indicators	J	Few	6
		Medium	10
		Large	16
Factor structure	λ_{jq}	Weak	(0.316, 0.447)
		Moderate	(0.316, 0.632)
		Strong	(0.632, 0.775)
Correlation between Factors	ψ_{12}	Independence	0.0
		Moderate	0.4
		Strong	0.8
Sample size	N	Very Small	50
		Small	100
		Big	300
		Very Big	3,000

For the number of indicators, Anderson and Gerbing (1984) suggested at least 3 indicators per factor in CFA. Kenny and McCoach (2003) varied the number of indicators from four to twenty-five to assess the impact on measures of fit. The number of indicators in our simulation study was varied from 3 to 8 per factor, which is equivalent to $J = 6$ to $J = 16$ indicators in total. For the factor structure, we used the ranges proposed by Acito and Anderson (1986). Both factor structure and variances for the measurement errors can be derived from the factor loadings, i.e., $\psi_{11} = \sum_{j=1}^J \lambda_{j1}^2$ and $\psi_{22} = \sum_{j=1}^J \lambda_{j2}^2$,

and $\phi_j = 1 - \lambda_j^2$, respectively. In our simulation study, following Acito and Anderson (1986), we set factor loading values to: (0.316, 0.447) for weak structure, (0.316, 0.632) for moderate structure, and (0.632, 0.775) for strong structure.

For the sample size, Boomsma (1985) recommended a sample size of at least $N = 50$ and Anderson and Gerbing (1984) suggested a sample size of at least $N = 150$. Boomsma and Hoogland (2001) showed that a sample size below $N = 200$ is vulnerable for the occurrence of improper solutions. In our Monte Carlo simulation the sample size was varied from $N = 50$ to $N = 3,000$. Three possible situations for the correlation between the latent variables were considered: $\psi_{12} = 0.0$ (independence), $\psi_{12} = 0.4$ (moderate association), and $\psi_{12} = 0.8$ (strong association).

2.3.3 Simulation

The simulated data is generated following the MIMIC model. In the simulated MIMIC model eight explanatory variables were considered. The true values that are used in the simulation study are based on the fitted MIMIC model on the NESDA data (Penninx et al., 2008). The first explanatory variable was generated from a Binomial distribution and the others from a Standard Normal distribution, i.e., $x_1 \sim \text{Bin}(0.67)$ and $x_k \sim N(0, 1)$ for $k = 2, \dots, 8$. The regression coefficients used in the simulation are the following, for x_1 : $\gamma_{11} = \gamma_{12} = 0.00$; x_2 : $\gamma_{21} = -0.10$, $\gamma_{22} = -0.20$; x_3 : $\gamma_{31} = \gamma_{32} = 0.00$; x_4 : $\gamma_{41} = 1.00$, $\gamma_{42} = 0.95$; x_5 : $\gamma_{51} = -0.30$, $\gamma_{52} = -0.25$; x_6 : $\gamma_{61} = \gamma_{62} = 0.00$; x_7 : $\gamma_{71} = 0.00$, $\gamma_{72} = 0.10$; and x_8 : $\gamma_{81} = \gamma_{82} = 0.00$.

Factor structures were then generated from a bivariate normal distribution $\boldsymbol{\theta} \sim N_2(\boldsymbol{\mu}, \Psi)$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \gamma_1^T \mathbf{x} + \zeta_1 \\ \gamma_2^T \mathbf{x} + \zeta_2 \end{bmatrix},$$

and

$$\Psi = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix} = \begin{bmatrix} \gamma_1^T \Sigma_x \gamma_1 + \text{Var}(\zeta_1) & \psi_{12} \\ \psi_{12} & \gamma_2^T \Sigma_x \gamma_2 + \text{Var}(\zeta_2) \end{bmatrix},$$

where γ_1 is a vector of regression coefficients for the first factor, and similarly γ_2 for the second factor.

2.3.4 Estimation

For each simulated data set a 2-factor model was fitted with and without explanatory variables, which corresponds to the CFA and the MIMIC model, respectively. The analysis was done using the Mplus statistical software version 7 (L. Muthen & Muthen, 1998-2012). A Maximum Likelihood Robust (MLR) estimator was employed for interval indicators whereas a Weighted Least Square estimator with Mean and Variance adjusted (WLSMV) was used for dichotomous indicator variables. The WLSMV is the default estimator in Mplus. We used the package called MplusAutomation (Hallquist, 2012) to help us call and run Mplus from the R environment.

The analysis procedure in our Monte Carlo simulation can be summarized as follows,

1. Fit a 2-factor CFA (or MIMIC model) on the simulated data.
2. Check if the fitted model is estimated without any problem due to improper solutions. Otherwise, identify the problem and record as nonconvergence and/or *Heywood*.
3. Estimate the factor scores, i.e., $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$.
4. If the fitted model is estimated correctly, calculate the correlation between true and estimated factor scores, i.e., $\rho_q = \text{Corr}(\theta_q, \hat{\theta}_q)$ where $q = 1, 2$. In the case of the MIMIC model, in addition to the correlation between factor scores, calculate:

- (a) the type-I error rate the regression coefficients.
- (b) the power of the regression coefficients.

5. Repeat Step 1 to 4 for each simulated data set.

2.3.5 Replication

In our Monte Carlo simulation we use a full factorial $3 \times 3 \times 3 \times 3 \times 4$ design, with in total 324 In each cell we use $R = 100$ replications.

2.3.6 Analysis of Output

For the first experiment, the variables of interest are the incidence of improper solutions and the quality of recovering the true factor scores in CFA .

The incidence of improper solutions was analyzed using a logistic regression model (Agresti, 2002). When a low rate of improper solutions is found in the data, Firth logistic regression (FLR: Firth, 1993) was used because it yields finite parameter estimates in the presence of complete or quasi-complete separation (Heinze & Schemper, 2002). SAS version 9.4 was used to fit the logistic regression models (SAS Institute Inc., 2013). The Odds Ratio (OR) was used as an effect size measure for evaluating the practical significance of the design variables and their interactions. We used the guidelines suggested by Ferguson (2009), i.e., an odds ratio of about 2 (or 0.50) indicates a small effect, about 3 (or 0.33) a medium effect, and about 4 (or 0.25) a large effect. For interpretation of simulation results, we focus on large effects for type of indicators and number of indicators, and their interactions with the other design variables.

Analysis of Variance (ANOVA) was used for analyzing the correlation data, i.e., ρ_1 and ρ_2 , to assess the impact of design variables on the quality of recovering the true factor scores. Because ρ_1 and ρ_2 are very similar we focus on ρ_1 . A Fisher's transformation is used to obtain an unbounded dependent variable, i.e., $z_1 = 0.5 \times \ln[(1 + \rho_1)/(1 - \rho_1)]$.

We used SPSS version 21 to fit the ANOVA model (IBM SPSS, 2012). The partial eta squared, denoted by η^2 , will be used as a measure of effect size for the ANOVA model. According to Cohen (1988), a value of $\eta^2 = 0.01$ indicates a small effect, $\eta^2 = 0.059$ a medium effect, and $\eta^2 = 0.138$ a large effect. For interpretation of simulation results, we focus on large effects for type of indicators and number of indicators, and their interactions with the other design variables.

In the second experiment we are further interested in the type-I error rate and the power for the regression weights of the MIMIC model. These measures were obtained by first calculating the proportion of cases in which an effect becomes statistically significant. For the effects equal to zero, the calculated proportion represents the type-I error rate; otherwise, the proportion corresponds to the power of the test. In the case of type-I error rate, a 95% confidence interval of the proportion using the Wilson interval was calculated (L. D. Brown, Cai, & DasGupta, 2001).

2.4 Results

2.4.1 Experiment-I: Confirmatory Factor Analysis

Nonconvergence in CFA

About 18.9% of the analyses in our simulation study did not converge. We applied logistic regression on the nonconvergence outcome variable (1: not converged; 0: converged) to investigate the impact of the design variables. The observed proportions of nonconvergence cross classified by design variables are presented in Table 2.3. A two-way interaction logistic model was fitted to the nonconvergence data.

The results of the 2-way interaction model are displayed in the Appendix (Table A.1); our focus here will be on the effects of two of the design variables, i.e., type of indicators and number of indicators, and their interaction with the other design variables. The type

Table 2.3: Percentage of nonconvergence in CFA under different experimental settings. Each cell result is based on $R = 100$ simulated replications.

Type of Indicators		Factor structure	Correlation between factors	Sample Size															
				Number of Indicators															
				50				100				300				3000			
				6	10	16	6	10	16	6	10	16	6	10	16				
BLR	Weak	Independence Moderate Strong	73.0 78.0 72.0	52.0 70.0 73.0	54.0 51.0 76.0	81.0 81.0 87.0	69.0 71.0 65.0	43.0 41 48.0	79.0 76.0 64.0	58.0 33.0 36.0	7.0 10.0 23.0	13.0 2.0 15.0	0.0 0.0 2.0	0.0 0.0 0.0					
			BMR	Weak	Independence Moderate Strong	79.0 85.0 79.0	51.0 60.0 56.0	27.0 31.0 46.0	75.0 68.0 66.0	41.0 47.0 33.0	7.0 6.0 27.0	48.0 31.0 29.0	8.0 2.0 17.0	0.0 0.0 6.0	0.0 0.0 4.0	0.0 0.0 0.0			
						Interval	Weak	Independence Moderate Strong	36.0 27.0 23.0	25.0 22.0 15.0	7.0 5.0 2.0	16.0 11.0 6.0	3.0 4.0 0.0	0.0 0.0 0.0	2.0 2.0 0.0	0.0 0.0 0.0	0.0 0.0 0.0	0.0 0.0 0.0	

of indicators has a large effect on the incidence of nonconvergence in CFA. Moreover, we found a large effect of 2-way interaction between the type of indicators and the following variables: number of indicators, factor structure, and sample size. There is also a large effect of number of indicators on the incidence of nonconvergence in CFA, and its 2-way interaction with the sample size. Figure 2.3 displays the corresponding interaction plots for the large effects. The first three panels (from left to right) show interaction plot between the type of indicators and the other design variables (i.e., the number of indicators, the factor structure, and the sample size). The last panel is for the interaction plot between the number of indicators and the sample size.

Regardless of the other design variables (i.e., number of indicators, factor structure, and sample size), we found a large effect of type of indicators on the prevalence of nonconvergence in CFA. The worst result was obtained for the binary indicators, specifically for the BLR data. For interval indicators, there was not much effect of the other design variables on the prevalence of nonconvergence in CFA.

By looking at the first and the last panel in Figure 2.3, there is a large interaction effect between the number of indicators with the type of indicators and the sample size. That is, the worst prevalence of nonconvergence due to the number of indicators was obtained when the binary indicators (i.e., BLR and BMR data) are analyzed by CFA. For the sample size, the worst prevalence of nonconvergence due to the number of indicators was found when the sample size is below $N \leq 300$.

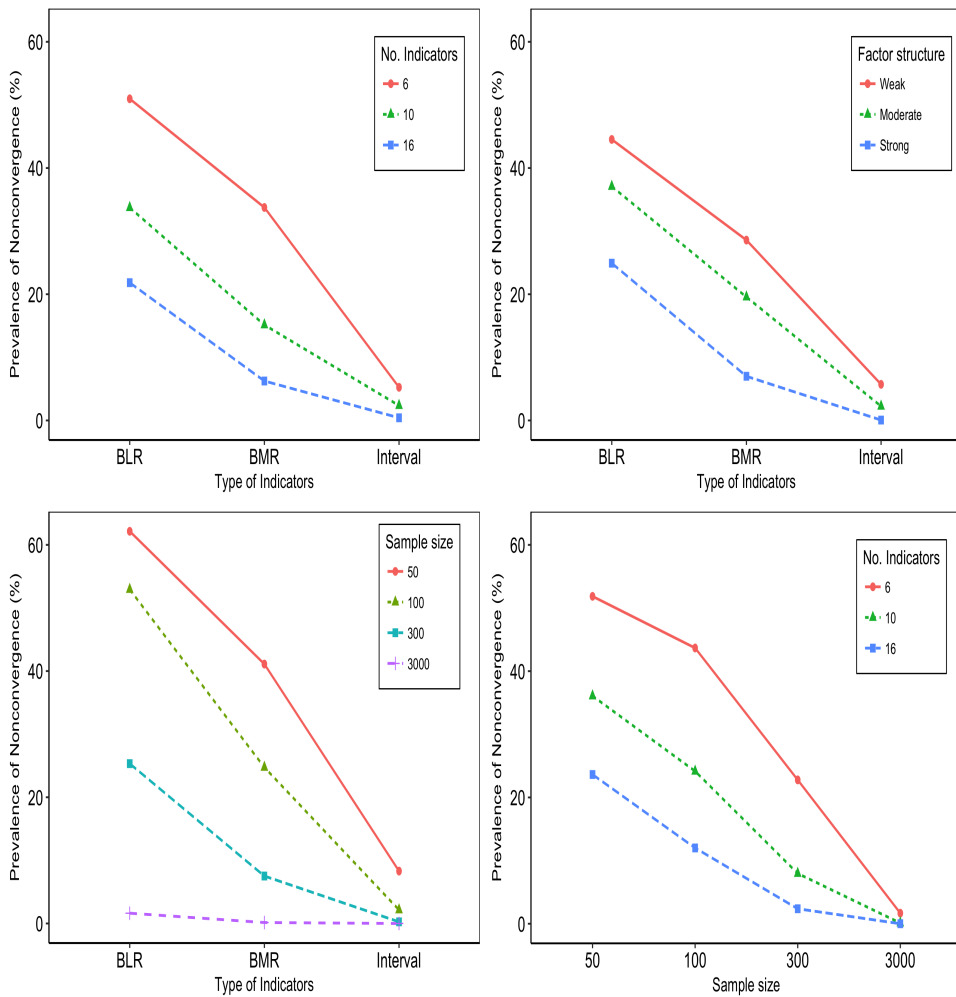


Figure 2.3: Interaction plot for Nonconvergence rate: The first three panels (from left to right) show interaction plot between the type of indicators and the number of indicators, the factor structure, and the sample size, respectively. The last panel is for the interaction between the number of indicators and the sample size.

Heywood cases in CFA

About 6.6% of the analyses in our simulation study resulted in *Heywood* cases. We applied logistic regression models on the *Heywood* outcome variable (1: yes; 0: no). The observed proportions of *Heywood* cases cross classified by the design variables are shown

in Table 2.4. A 2-way interaction logistic model was fitted on the *Heywood* data, and the results are presented in the Appendix (Table A.2).

Like for the nonconvergence analysis, our focus will be on the effects of the type of indicators and the number of indicators, and their interaction with the other design variables. The type of indicators has a large effect on the incidence of *Heywood* in CFA. Moreover, we found a large effect of 2-way interaction between the type of indicators and all the other design variables, i.e., number of indicators, factor structure, correlation between latent variables, and sample size. There is also a large effect of number of indicators on the incidence of nonconvergence in CFA, and its 2-way interaction with both the factor structure and sample size. Figure 2.4 displays the interaction plots for the large effects. The first four panels (from left to right) show interaction plot between the type of indicators with the number of indicators, the factor structure, the correlation between underlying latents, and the sample sizes. The last two panels are for the interaction plot between the number of indicators with the factor structure and sample size.

In the first three panels, it can be seen that there is no large difference in prevalence of *Heywood* cases among the type of indicators used in CFA. The fourth panel shows the interaction with the sample size, where the highest number of *Heywood* cases was found for the BLR data, except for a large and small data sets.

There is a large effect of the number of indicators on the prevalence of *Heywood* cases in CFA. The first panel in the last row shows that the worst result was found for a small number of indicators regardless of the type of indicators in CFA. Furthermore, more *Heywood* cases were found for the smallest number of indicators with weak factor structure and with small sample sizes.

Table 2.4: Percentage of *Heywood* cases in CFA under different experimental settings. Each cell result is based on $R = 100$ simulated replications.

Type of Indicators		Factor structure		Correlation between factors		Sample Size															
						Number of Indicators															
						50				100				300				3000			
						6	10	16	6	10	16	6	10	16	6	10	16	6	10	16	16
BLR	Weak	Independence	21.0	6.0	3.0	51.0	24.0	2.0	50.0	26.0	3.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			21.0	3.0	6.0	41.0	23.0	7.0	44.0	15.0	4.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			14.0	4.0	2.0	42.0	9.0	1.0	28.0	6.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Moderate	Independence	18.0	2.0	4.0	52.0	16.0	0.0	48.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			9.0	3.0	1.0	33.0	7.0	0.0	26.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			8.0	0.0	4.0	31.0	4.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Strong	Independence	10.0	4.0	7.0	12.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			6.0	3.0	4.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			5.0	2.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BMR	Weak	Independence	47.0	13.0	3.0	53.0	18.0	3.0	34.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			45.0	20.0	2.0	37.0	13.0	1.0	12.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			51.0	14.0	1.0	25.0	2.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Moderate	Independence	41.0	10.0	1.0	38.0	6.0	0.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			38.0	5.0	0.0	23.0	1.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			17.0	1.0	0.0	13.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Strong	Independence	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			3.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Interval	Weak	Independence	66.0	48.0	14.0	54.0	14.0	2.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			51.0	34.0	12.0	46.0	7.0	1.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			38.0	18.0	6.0	13.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Moderate	Independence	60.0	22.0	2.0	40.0	2.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			40.0	13.0	1.0	24.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			14.0	4.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Strong	Independence	8.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			7.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
			0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

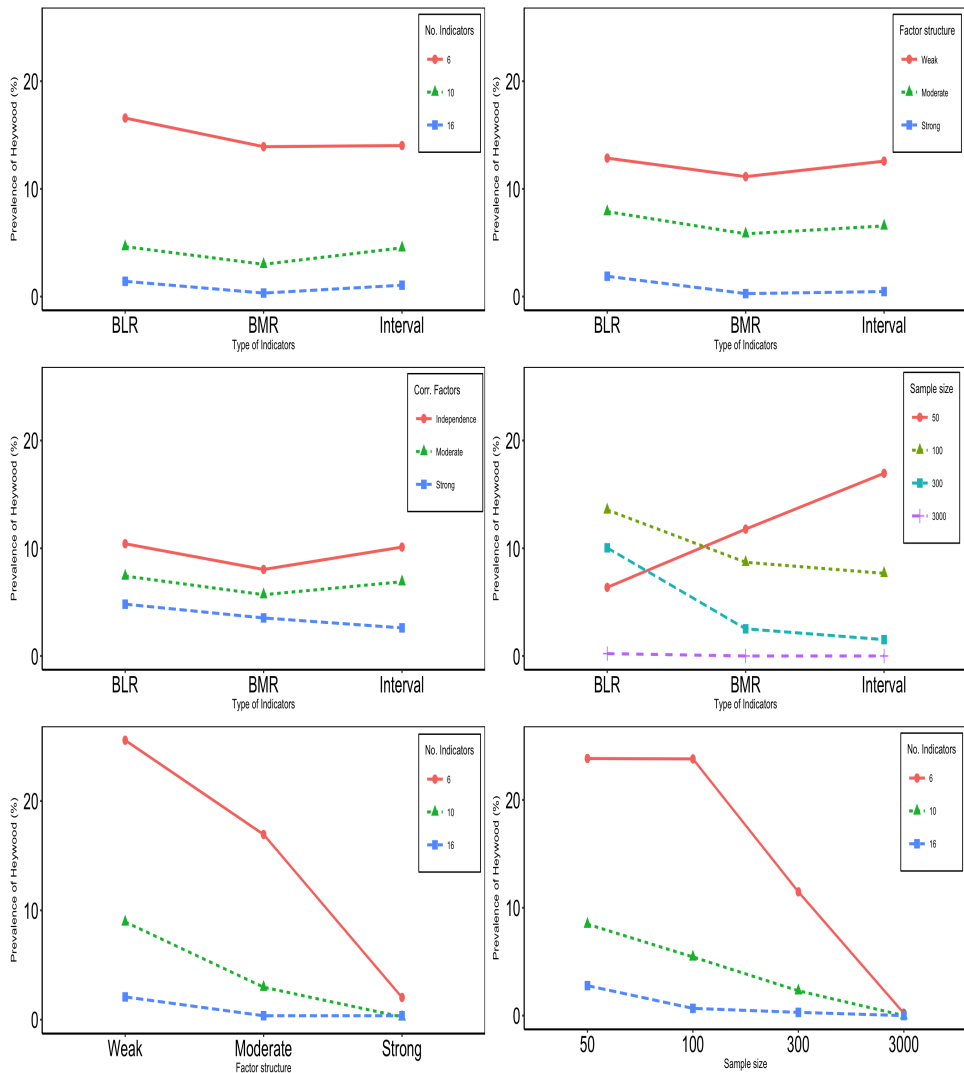


Figure 2.4: Interaction plot for Heywood rate: The first four panels (from left to right) show the interaction between the type of indicators and the number of indicators, the factor structure, the correlation between underlying latents, and the sample size, respectively. The last two panels are for the interaction between the number of indicators and the factor structure and the sample size.

Recovery of Factor Scores

The observed average correlations between the true and estimated factor scores for the first latent variable cross classified by the design variables as displayed in Table 2.5. A 2-way interaction ANOVA model was fitted on the transformed correlation data, and the results are presented in the Appendix (Table A.3). Large effects were found for the number of indicators, the type of indicators, and the interaction between type of indicators and factor structure. Figure 2.5 displays the interaction plots for the large effects. The first panel shows two main-effect plots corresponding to effects of the type of indicators and the number of indicators, respectively. The second panel shows the interaction plot between the type of indicators and the factor structure.

Both plots show that there is a large effect of type of indicators on the quality of recovering the true factor scores in CFA. The worst result was obtained for the binary indicators (i.e., BLR and BMR) regardless of the other design variables.

The first plot also shows a large effect of the number of indicators on the quality of recovering the factor scores in CFA. The worst result was found for the smallest number of indicators. Although we found a large interaction effect between the type of indicators and the factor structure, the second plot in Figure 2.5 does not clearly show this effect.

Table 2.5: Quality of Recovering the True Factor scores: Average correlation between the true and estimated factor scores of CFA, i.e., $\text{Corr}(\hat{\theta}_1, \hat{\theta}_1) = \hat{\rho}_1$, under different experimental settings. Each cell represents the results of $R = 100$ simulated replications, except those models that were not identified due to improper solutions.

		Sample Size															
		50				100				300				3000			
		Number of Indicators															
Type of Indicators	Factor structure	Correlation between factors			6	10	16	6	10	16	6	10	16	6	10	16	
BLR	Weak	Independence Moderate Strong	0.29	0.32	0.37	0.29	0.32	0.40	0.23	0.38	0.47	0.35	0.44	0.53	0.55	0.53	
			0.30	0.42	0.49	0.29	0.37	0.46	0.32	0.41	0.50	0.36	0.47	0.55	0.53		
			0.39	0.47	0.56	0.36	0.41	0.54	0.38	0.48	0.58	0.44	0.53	0.62	0.62		
	Moderate	Independence Moderate Strong	0.32	0.38	0.51	0.39	0.46	0.53	0.38	0.51	0.60	0.44	0.53	0.62	0.62	0.62	
			0.42	0.50	0.57	0.41	0.50	0.59	0.43	0.53	0.62	0.46	0.55	0.64	0.64		
			0.46	0.56	0.65	0.52	0.56	0.65	0.49	0.59	0.68	0.52	0.62	0.69	0.69		
	Strong	Independence Moderate Strong	0.48	0.61	0.70	0.54	0.65	0.72	0.57	0.66	0.73	0.59	0.66	0.73	0.73	0.73	
			0.57	0.64	0.72	0.59	0.67	0.73	0.60	0.67	0.74	0.61	0.69	0.75	0.75		
			0.61	0.70	0.76	0.64	0.69	0.77	0.65	0.72	0.78	0.66	0.73	0.78	0.78		
BMR	Weak	Independence Moderate Strong	0.36	0.46	0.50	0.43	0.50	0.57	0.43	0.54	0.65	0.47	0.58	0.67	0.67	0.67	
			0.44	0.42	0.56	0.43	0.52	0.63	0.47	0.58	0.67	0.50	0.60	0.68	0.68	0.68	
			0.49	0.51	0.67	0.52	0.60	0.70	0.53	0.64	0.73	0.57	0.66	0.74	0.74		
	Moderate	Independence Moderate Strong	0.47	0.58	0.68	0.49	0.63	0.74	0.56	0.66	0.75	0.58	0.68	0.76	0.76	0.76	
			0.53	0.62	0.70	0.58	0.65	0.75	0.57	0.68	0.76	0.60	0.69	0.77	0.77		
			0.56	0.71	0.77	0.63	0.70	0.79	0.65	0.74	0.81	0.66	0.75	0.81	0.81		
	Strong	Independence Moderate Strong	0.73	0.80	0.87	0.74	0.82	0.87	0.75	0.82	0.87	0.75	0.83	0.88	0.88	0.88	
			0.74	0.81	0.87	0.75	0.82	0.88	0.76	0.83	0.88	0.77	0.83	0.88	0.88		
			0.79	0.85	0.89	0.79	0.85	0.90	0.80	0.86	0.90	0.81	0.86	0.90	0.90		
Interval	Weak	Independence Moderate Strong	0.41	0.54	0.66	0.46	0.64	0.73	0.54	0.67	0.75	0.58	0.68	0.76	0.76	0.76	
			0.53	0.59	0.68	0.53	0.64	0.74	0.57	0.68	0.76	0.60	0.70	0.77	0.77	0.77	
			0.52	0.65	0.77	0.61	0.71	0.79	0.64	0.73	0.81	0.66	0.75	0.81	0.81	0.81	
	Moderate	Independence Moderate Strong	0.60	0.73	0.82	0.65	0.75	0.83	0.66	0.78	0.84	0.69	0.78	0.85	0.85	0.85	
			0.66	0.75	0.82	0.67	0.76	0.85	0.69	0.78	0.84	0.71	0.79	0.85	0.85	0.85	
			0.69	0.80	0.86	0.73	0.81	0.87	0.75	0.82	0.88	0.76	0.83	0.88	0.88	0.88	
	Strong	Independence Moderate Strong	0.84	0.91	0.94	0.85	0.90	0.94	0.86	0.91	0.94	0.87	0.91	0.94	0.91	0.94	0.94
			0.84	0.90	0.94	0.86	0.91	0.94	0.87	0.91	0.94	0.87	0.92	0.94	0.92	0.94	0.94
			0.87	0.92	0.95	0.88	0.92	0.95	0.89	0.93	0.95	0.89	0.93	0.95	0.89	0.93	0.95

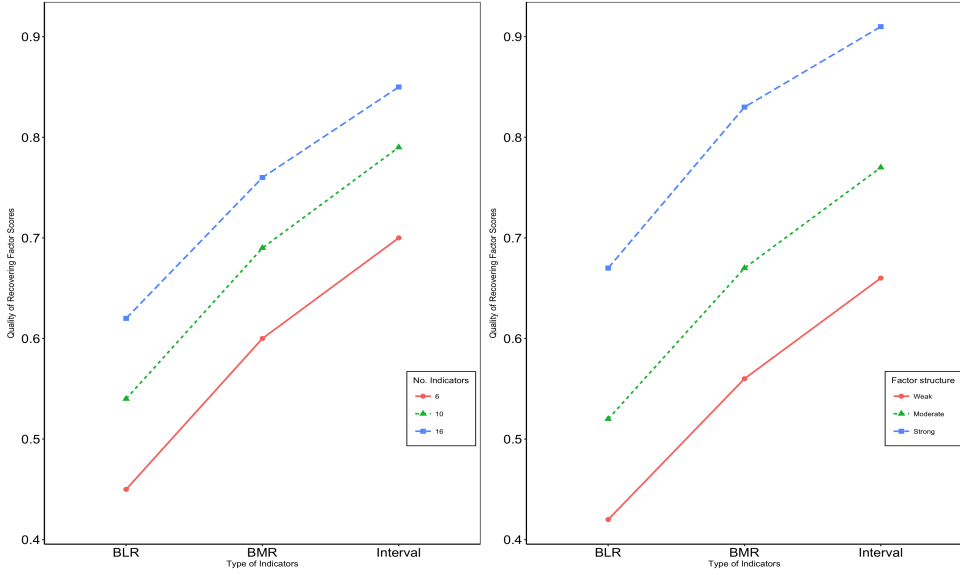


Figure 2.5: Interaction plot for Quality of Recovering Factors: The first panel shows two main effects for the type and number of indicators. The second panel shows the interaction between the type of indicators and the factor structure.

2.4.2 Experiment-II: The MIMIC Model

In the second experiment, we study the type-I error rate and the power of the regression parameters (i.e., γ) of the MIMIC model.

Type-I Error

In this section we study the impact of design variables on the Type-I error rate. Although there are a couple of parameters whose true values are set to zero in the simulation study, for demonstration purpose we chose one of the parameters, i.e., γ_{31} which indicates the relationship between X_3 and the first latent variable whose results are presented in Table 2.6.

Those values whose 95% confidence interval excluded the nominal level of significance ($\alpha = 0.05$) were made bold; there are a total of fourteen cells which resulted in such cases,

Table 2.6: Observed type-I error rates for the relationship between X_3 and the first factor, γ_{31} . The values in bold represent 95% confidence interval excluding the nominal level of significance ($\alpha = 0.05$). The number of replications per cell differs because of improper solutions. Dashed lines indicate no valid results were obtained for that cell.

		Sample Size																			
		50					100					300					3000				
		Number of Indicators																			
Type of Indicators	Factor structure	Correlation between factors																			
		6	10	16	6	10	16	6	10	16	6	10	16	6	10	16	6	10	16		
BLR	Weak	Independence	—	—	—	0.00	0.00	0.00	0.01	0.04	0.02	0.01	0.04	0.02	0.01	0.01	0.04	0.01	0.01	0.04	
		Moderate	—	—	—	0.00	0.03	0.00	0.02	0.02	0.07	0.03	0.02	0.07	0.03	0.02	0.03	0.02	0.03	0.03	
		Strong	—	—	—	0.00	0.00	0.03	0.01	0.05	0.02	0.05	0.02	0.05	0.02	0.05	0.05	0.05	0.04	0.04	
	Moderate	Independence	—	—	—	0.00	0.00	0.00	0.00	0.09	0.06	0.10	0.03	0.03	0.03	0.03	0.06	0.02	0.02	0.02	
		Moderate	—	—	—	0.00	0.00	0.00	0.04	0.04	0.01	0.03	0.08	0.06	0.06	0.06	0.02	0.02	0.02	0.02	
		Strong	0.00	—	—	0.00	0.00	0.17	0.01	0.01	0.03	0.08	0.08	0.08	0.08	0.08	0.06	0.02	0.02	0.02	
	Strong	Independence	—	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.10	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.00	
		Moderate	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.08	0.07	0.00	0.08	0.07	0.00	0.00	0.00	0.00	
		Strong	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.11	0.03	0.08	0.03	0.08	0.03	0.08	0.08	0.02	0.02	
	BMR	Weak	Independence	0.07	0.02	0.02	0.01	0.06	0.02	0.04	0.08	0.02	0.05	0.08	0.02	0.05	0.08	0.01	0.01	0.04	0.04
Moderate			0.03	0.00	0.07	0.02	0.05	0.03	0.02	0.05	0.05	0.08	0.06	0.05	0.08	0.06	0.05	0.05	0.05	0.05	
Strong			0.00	0.02	0.07	0.06	0.05	0.06	0.03	0.06	0.03	0.09	0.03	0.09	0.03	0.09	0.05	0.03	0.03	0.03	
Moderate		Independence	0.05	0.10	0.00	0.03	0.06	0.03	0.03	0.05	0.04	0.08	0.02	0.05	0.08	0.02	0.02	0.02	0.08	0.08	
		Moderate	0.00	0.03	0.08	0.05	0.03	0.05	0.03	0.04	0.11	0.05	0.02	0.04	0.11	0.05	0.02	0.02	0.02	0.02	
		Strong	0.03	0.04	0.05	0.04	0.04	0.07	0.02	0.05	0.05	0.08	0.03	0.05	0.05	0.08	0.03	0.03	0.03	0.03	
Strong		Independence	0.00	0.00	0.00	0.00	0.00	0.07	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
		Moderate	0.00	0.00	—	0.00	0.00	0.11	0.00	0.00	0.00	0.05	0.06	0.00	0.05	0.06	0.00	0.00	0.00	0.00	
		Strong	0.00	—	0.00	0.20	0.00	0.00	0.07	0.38	0.10	0.03	0.06	0.04	0.03	0.06	0.06	0.04	0.04	0.04	
Interval		Weak	Independence	0.03	0.07	0.09	0.01	0.08	0.05	0.08	0.07	0.04	0.08	0.05	0.03	0.05	0.03	0.05	0.03	0.03	0.03
	Moderate		0.07	0.11	0.11	0.07	0.08	0.00	0.02	0.04	0.04	0.07	0.03	0.04	0.07	0.07	0.06	0.06	0.06	0.06	
	Strong		0.04	0.05	0.06	0.09	0.07	0.11	0.03	0.04	0.05	0.03	0.02	0.03	0.03	0.06	0.05	0.05	0.05	0.05	
	Moderate	Independence	0.08	0.09	0.12	0.11	0.10	0.05	0.05	0.06	0.05	0.04	0.06	0.05	0.04	0.06	0.01	0.01	0.01	0.01	
		Moderate	0.10	0.09	0.09	0.10	0.05	0.05	0.01	0.06	0.07	0.03	0.04	0.07	0.03	0.04	0.04	0.04	0.04	0.04	
		Strong	0.04	0.10	0.08	0.03	0.04	0.05	0.06	0.03	0.03	0.02	0.03	0.02	0.03	0.05	0.05	0.08	0.08	0.08	
	Strong	Independence	0.10	0.09	0.07	0.08	0.06	0.08	0.03	0.06	0.11	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	
		Moderate	0.08	0.09	0.07	0.04	0.06	0.05	0.03	0.06	0.05	0.06	0.05	0.06	0.05	0.06	0.04	0.04	0.04	0.04	
		Strong	0.09	0.14	0.07	0.07	0.08	0.09	0.07	0.08	0.06	0.06	0.06	0.06	0.06	0.06	0.04	0.04	0.04	0.04	
		Strong	0.09	0.14	0.07	0.07	0.08	0.09	0.07	0.08	0.06	0.06	0.06	0.06	0.06	0.06	0.04	0.04	0.04	0.04	

i.e., one for BLR data, one for BMR data, and the rest for interval data. However, for BLR and BMR data with $N = 50$ no results were found due to the presence of improper solutions and are thus represented by dashed lines in Table 2.6. Therefore, this resulted in a wider confidence interval for the binary indicators (i.e., the BLR and the BMR data) and because of that the 5% level of significance level is included.

Although the confidence intervals obtained for both BLR and BMR data do include the level of significance, most of the point estimates are either zero or very high. Therefore, the observed type-I error rates were too conservative (i.e., values very close to zero) for BLR, and inflated (i.e., above the 5% nominal level of significance) for BMR indicators. In the case of interval indicators, most of the confidence intervals included the nominal level of significance, and their results seems stable around $\alpha = 0.05$.

The number of indicators had an impact on the recovery of type-I error rate, particularly for dichotomous indicator variables with low success rates (BLR). That is, among the 108 observed type-I error rates obtained from BLR only thirteen included the 5% nominal level of significance in their 95% confidence interval for $J = 6$, while sixteen for $J = 10$ and nineteen for $J = 16$ recovered the required type-I error rate. In the case of both BMR and interval indicators, we found no significant difference on the recovery of type-I error rate among the type of indicators.

Power

In this section we focus on the observed power of three of the parameters in the MIMIC model, i.e., $\gamma_{51} = -0.30$ representing a moderate effect, $\gamma_{72} = 0.10$ for a small effect, and $\gamma_{42} = 0.95$ for a strong effect. Their results are shown in Table 2.7 and in the Appendix (Table A.4 and A.5), respectively. Some of the cells of these tables had no values due to the presence of improper solutions and are consequently represented by dashed lines. A threshold value of 0.80 will be used as a criterion to represent adequate power.

It is evident from Table 2.7 that most of the observed power values obtained from

Table 2.7: Observed power for the relationship between X_5 and the first factor, $\gamma_{51} = -0.30$. The number of replications per cell differ because of improper solutions. Dashed lines indicate no valid results were obtained for that cell.

		Sample Size																			
		50					100					300					3000				
		Number of Indicators																			
Type of Indicators	Factor structure	Correlation between factors																			
		6	10	16	6	10	16	6	10	16	6	10	16	6	10	16					
BLR	Weak	—	—	—	0.12	0.06	0.03	0.27	0.31	0.43	0.99	1.00	1.00	1.00	1.00	1.00					
		—	—	—	0.00	0.00	0.00	0.29	0.34	0.56	1.00	1.00	1.00	1.00	1.00	1.00					
		—	—	—	0.09	0.03	0.03	0.22	0.35	0.42	1.00	1.00	1.00	1.00	1.00	1.00					
	Moderate	—	—	—	0.00	0.00	0.00	0.21	0.34	0.43	1.00	1.00	1.00	1.00	1.00	1.00					
		—	—	—	0.07	0.00	0.00	0.17	0.41	0.38	1.00	1.00	1.00	1.00	1.00	1.00					
		0.00	—	—	0.05	0.06	0.08	0.30	0.38	0.42	1.00	1.00	1.00	1.00	1.00	1.00					
BMR	Strong	—	0.00	0.00	0.00	0.00	0.00	0.06	0.47	0.51	1.00	1.00	1.00	1.00	1.00	1.00					
		0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.53	0.38	1.00	1.00	1.00	1.00	1.00	1.00					
		—	0.00	—	0.25	0.00	0.00	0.20	0.40	0.26	1.00	1.00	1.00	1.00	1.00	1.00					
	Weak	0.10	0.08	0.11	0.11	0.23	0.31	0.53	0.64	0.69	1.00	1.00	1.00	1.00	1.00	1.00					
		0.06	0.13	0.07	0.21	0.29	0.23	0.52	0.65	0.71	1.00	1.00	1.00	1.00	1.00	1.00					
		0.06	0.11	0.12	0.16	0.26	0.31	0.60	0.59	0.65	1.00	1.00	1.00	1.00	1.00	1.00					
Interval	Moderate	0.09	0.03	0.00	0.26	0.21	0.26	0.53	0.73	0.72	1.00	1.00	1.00	1.00	1.00	1.00					
		0.16	0.08	0.25	0.23	0.33	0.37	0.52	0.68	0.70	1.00	1.00	1.00	1.00	1.00	1.00					
		0.15	0.11	0.08	0.21	0.30	0.18	0.53	0.67	0.72	1.00	1.00	1.00	1.00	1.00	1.00					
	Strong	0.00	0.00	0.33	0.14	0.50	0.13	0.67	0.80	0.35	1.00	1.00	1.00	1.00	1.00	1.00					
		0.00	0.00	—	0.18	0.33	0.11	0.64	0.71	0.73	1.00	1.00	1.00	1.00	1.00	1.00					
		0.00	—	0.00	0.60	0.50	0.50	0.71	0.50	0.70	1.00	1.00	1.00	1.00	1.00	1.00					
Interval	Weak	0.12	0.21	0.22	0.21	0.28	0.23	0.58	0.69	0.79	1.00	1.00	1.00	1.00	1.00	1.00					
		0.17	0.23	0.15	0.26	0.27	0.25	0.69	0.72	0.85	1.00	1.00	1.00	1.00	1.00	1.00					
		0.16	0.24	0.23	0.21	0.26	0.31	0.59	0.67	0.71	1.00	1.00	1.00	1.00	1.00	1.00					
	Moderate	0.13	0.28	0.21	0.25	0.26	0.38	0.65	0.75	0.86	1.00	1.00	1.00	1.00	1.00	1.00					
		0.21	0.22	0.17	0.34	0.43	0.44	0.64	0.73	0.82	1.00	1.00	1.00	1.00	1.00	1.00					
		0.22	0.24	0.27	0.32	0.42	0.32	0.75	0.72	0.86	1.00	1.00	1.00	1.00	1.00	1.00					
Strong	0.19	0.34	0.26	0.35	0.35	0.42	0.81	0.84	0.84	1.00	1.00	1.00	1.00	1.00	1.00						
	0.30	0.25	0.26	0.33	0.35	0.38	0.80	0.86	0.78	1.00	1.00	1.00	1.00	1.00	1.00						
	0.17	0.29	0.34	0.39	0.30	0.48	0.86	0.77	0.83	1.00	1.00	1.00	1.00	1.00	1.00						

dichotomous indicators are very low, except when the sample size is very large. Let us elaborate on this point using the results under a sample size of $N = 300$ from Table 2.7. Out of the twenty-seven observed values for BLR (i.e., Binary indicators with Low success Rates) none of them passed the threshold value of 0.80 and the maximum value achieved was only 0.56. For the case of BMR (i.e., Binary indicators with Moderate success Rates), only one out of twenty-seven had an observed power of exactly 0.80. In the case of interval indicators, however, eleven out of twenty-seven satisfied the criteria and the maximum value achieved was an observed power of 0.86.

The minimum power was 0.00 for both BLR and BMR while it was 0.12 for interval indicators. With a small effect size of $\gamma_{72} = 0.10$ as shown in the Appendix (Table A.4), among the 216 cells for dichotomous indicators (i.e., both BLR and BMR) only three of the observed values are larger than 0.8 while in the case of interval indicators nineteen out of 108 cells satisfied this criterion. These results show that the MIMIC model performed poorly for analyzing dichotomous indicators, particularly for dichotomous indicators with low success rates.

2.5 Conclusion and Discussion

Structural equation models are originally proposed for analysis of continuous (or interval) indicator variables. Recently, factor analysis and structural equation models have been applied for data with dichotomous indicators and with only a few indicators per latent variable, i.e., 2 or 3 (Krueger, 1999; Beesdo-Baum et al., 2009). Using a Monte Carlo simulation study, we showed that latent variable models applied on such type of data performed poorly with higher incidence of improper solutions, poor quality of recovering the true factor scores, too conservative or inflated type-I error rates, and weak power.

About 18.9% out of all the analyses in the CFA did not achieve convergence, and about 6.6% were *Heywood* cases (i.e., out-of-bound problem). It is shown that the type

of indicators and the number of indicators in CFA plays a major role on the occurrence of the nonconvergence in CFA. That is, high prevalence of nonconvergence was obtained for binary indicators with a few indicators per latent variable. For the occurrence of *Heywood* cases in CFA, the number of indicators also played a major role. The quality of recovering the true factor scores in CFA was poor in the case of binary indicators, and it became worse with less indicators per latent variable.

We evaluated the performance of the MIMIC model using the type-I error rate and the power of test for the regression weights. Most of the confidence intervals of the type-I error distribution obtained from the dichotomous indicators, did not include the 5% nominal level of significance. The type-I error rates were mostly conservative, although few of them were inflated. For interval indicators, however, most of the results included the nominal level of significance within their 95% confidence interval. The power of the test with dichotomous indicators was poor compared to the interval indicators.

It is important to note that we used an advantageous design for our Monte Carlo simulation study. The latent variables were generated from a bivariate normal distribution. Moreover, the population model was correctly specified. In empirical studies it is likely that assumptions are only partially valid. Moreover, the fitted model could be misspecified; for example, an important indicator variable may not have been included in the analysis. Under such conditions we would expect even more improper solutions and factor scores that are further off than what we found in our current study.

Latent trait or Item Response Theory (IRT) model has also been proposed for analyzing dichotomous indicator variables (Lord & Novick, 1968). It was shown by Mislevy (1986) and Takane and De Leeuw (1987), that CFA and IRT models are formally equivalent and thus yielding similar results. Knol and Berger (1991) conducted a simulation study to compare the performance of these models and found that the common factor analysis on the matrix of tetrachoric correlations performed similar to IRT models for multidimensional data. Furthermore, Glöckner-Rist and Hoijtink (2003) recommended a joint application

of the models. We conclude our discussion with a recommendation for those researchers who do confirmatory factor analysis on data with a small number of dichotomous indicator variables. It is shown in our Monte Carlo simulation that the method performed poorly for this type of data and therefore must be used carefully. An alternative statistical model which requires less assumptions might be more appropriate, for example the multivariate logistic distance model (Worku & De Rooij, 2018).

Chapter 3

Properties of Ideal Point Classification Models for Bivariate Binary Data

Abstract

The Ideal Point Classification (IPC) model was originally proposed for analysing multinomial data in the presence of predictors. In this paper, we studied properties of the IPC model and extended it for analysing bivariate binary responses with a specific focus on three parameters: (1) the marginal probabilities; (2) the association structure between the two binary responses; and (3) the joint probabilities. We found that the IPC model with a specific class point configuration, represents either the marginal probabilities or the association structure. However, the IPC model is not able to represent both parameters at the same time. We then derived a new parameterization of the model, the Bivariate IPC (BIPC) model, which is able to represent both the marginal probabilities and the association structure. Like the standard IPC model, the results of the BIPC model can be displayed in a biplot, from which the effects of predictors on the binary responses and on their association can be read. We will illustrate our findings with a psychological example relating personality traits to depression and anxiety disorders.

This chapter was published as Worku, H. M. & De Rooij, M. (2017). Properties of Ideal Point Classification Models for Bivariate Binary Data. *Psychometrika*, 82 (2), 308-328. To address remarks of the PhD committee, this chapter is slightly modified.

3.1 Introduction

Multiple binary outcome data are often collected in epidemiology, psychology, medicine, and other life and behavioral sciences. For example, in the Netherlands Study of Depression and Anxiety (NESDA) data were collected on depression and anxiety disorders, and how these disorders are influenced by personality traits and background variables (Penninx et al., 2008; Spinhoven et al., 2009). In this paper, we focus on bivariate binary data in which two dichotomous response variables are observed for each subject in a study. Another example with bivariate binary data is the British coalminers study (Ashford et al., 1970), which investigated data on breathlessness (1 = difficult; 0 = Normal) and wheeze (1 = difficult; 0 = Normal) of coalminers in Britain, to study the impact of exposure on these respiratory indicators (Ashford et al., 1970; McCullagh & Nelder, 1989; Palmgren, 1989).

Let us denote the bivariate binary responses observed from the i -th subject by Y_{i1} and Y_{i2} . The p dimensional vector $\mathbf{x}_i$ represents the explanatory variables without including an intercept, where $i = 1, 2, \dots, N$. The cross-classified binary responses are displayed in Table 3.1 in which the corresponding probabilities are also presented, i.e., the probabilities within the four cells represent the joint probabilities; and those at the margins represent the marginal probabilities. Empirical researchers working with bivariate binary data are often interested in the following parameters: (1) the marginal probabilities; (2) the association between the two binary responses; and (3) the joint (or multinomial) probabilities.

In marginal modelling, the main focus is on the analysis of the marginal probabilities separately in which the association structure between the binary responses could be a direct interest or treated as a nuisance parameter (Agresti, 2002, pp. 455; Molenberghs & Verbeke, 2005, pp. 55). In the margins of Table 3.1, the marginal probabilities are denoted by $\pi_{i1\cdot} = \Pr(Y_{i1} = 1)$ and $\pi_{i\cdot l} = \Pr(Y_{i2} = l)$, where $l = 0, 1$. Bahadur (1961) proposed a marginal model based on the full likelihood for analysing bivariate binary data. The

joint distribution was characterized by the two marginal distributions and the correlation between the two binary responses. Lipsitz, Laird and Harrington (1990) followed the idea of Bahadur (1961) and showed that other measures of association can also be used (e.g., the odds ratio or relative risk). For a 2×2 contingency table, the odds ratio is calculated as $\tau_i = (\pi_{i,11} \times \pi_{i,00}) / (\pi_{i,10} \times \pi_{i,01})$ where $\pi_{i,11} = \Pr(Y_{i1} = 1, Y_{i2} = 1)$; $\pi_{i,00} = \Pr(Y_{i1} = 0, Y_{i2} = 0)$; $\pi_{i,10} = \Pr(Y_{i1} = 1, Y_{i2} = 0)$; and $\pi_{i,01} = \Pr(Y_{i1} = 0, Y_{i2} = 1)$.

Marginal model parameters can be fitted directly or by imposing restrictions on the joint distribution (Molenberghs & Verbeke, 2005, pp. 49). Aitchison and Silvey (1958, 1960) originally proposed constraints on parameters in maximum likelihood function. Their approach was later applied to categorical data by Lang and Agresti (1994), and other researchers (Lang, 1996; Bergsma, 1997; Bergsma & Rudas, 2002; Vermunt, Rodrigo, & Ato-Garcia, 2001). McCullagh and Nelder (1989) introduced a multivariate logistic transformation which can be used to relate the joint distribution to the marginal probabilities and the association structure. Their approach is widely used for marginal modelling of multivariate categorical responses (Glonek & McCullagh, 1995; Molenberghs & Lesaffre, 1994, 1999).

In recent years, the marginal modelling strategy has shifted from fitting and testing linear constraints on parameters to inequality constraints for addressing certain scientific questions (Colombi & Forcina, 2001; Bartolucci, Forcina, & Dardanoni, 2001; Bartolucci, Colombi, & Forcina, 2007). For ordinal responses, for example, it may be interesting to know whether the univariate distributions are stochastically ordered in some way, i.e., whether pairs of responses are positively correlated, or whether the degree of positive dependence changes with certain predictor variables (Colombi & Forcina, 2001).

The main drawback of a full likelihood-based marginal modelling approach is that it is computationally intensive and prone to model misspecification, especially when the number of response variables increases (Agresti, 2002, pp. 465; Molenberghs & Verbeke, 2005, pp. 151). Liang and Zeger (1986) proposed an extension of quasi-likelihood

Table 3.1: Cross-classification of bivariate binary data observed from i -th subject.

		Y_{i2}		
		1	0	
Y_{i1}	1	$\pi_{i,11}$	$\pi_{i,10}$	$\pi_{i1.}$
	0	$\pi_{i,01}$	$\pi_{i,00}$	$\pi_{i0.}$
		$\pi_{i.1}$	$\pi_{i.0}$	1.00

method, called Generalized Estimating Equations (GEE or GEE1), that does not require full specification of the response distribution. In GEE1 the association structure is treated as a nuisance parameter. Second-order GEE, called GEE2, (Liang et al., 1992) and Alternating Logistic Regression (ALR: Carey, Zeger, & Diggle, 1993) are commonly used for modelling both the marginal probabilities and the association structure.

The third parameter of interest are the joint probabilities. The joint probabilities as displayed in Table 3.1 (i.e., $\pi_{i,00}$; $\pi_{i,10}$; $\pi_{i,01}$; and $\pi_{i,11}$) correspond to a multinomial response variable, denoted by G_i , with four categories ($g = 4$). For simplicity, we use a single index to refer to the joint probabilities, i.e., $\pi_{ij} = \Pr(G_i = j)$. For example, the four cells in Table 3.1 can be represented as: $\pi_{i1} = \pi_{i,00}$; $\pi_{i2} = \pi_{i,10}$; $\pi_{i3} = \pi_{i,01}$; and $\pi_{i4} = \pi_{i,11}$. In the NESDA study, for example, a multinomial response variable can be defined from the two binary outcome variables. That is, $G_i = 1$ if the subject has no depression or anxiety; $G_i = 2$ if (s)he has an anxiety disorder, but no depression disorder; $G_i = 3$ if the subject has depression disorder, but no anxiety disorder; and $G_i = 4$ if there is co-morbidity. Statistical models such as the Multinomial Baseline-Category Logit (MBCL: Agresti, 2002, pp. 267) or Ideal Point Classification (IPC: De Rooij, 2009a), can be used to analyse multinomial response variables in the presence of predictors.

De Rooij (2009a) proposed the IPC model for analysing a multinomial response variable in the presence of predictors. The IPC model is a probabilistic multidimensional unfolding model and closely related to Ideal Point Discriminant Analysis (IPDA) as proposed by Takane, Bozdogan, and Shibayama (1987). Both IPDA and IPC models are classification methods based on multidimensional unfolding (MDU) (Heiser, 1981, 1987; De Leeuw,

2005). The objective of MDU is to find distances in Euclidean space between subjects and objects that approximate a set of proximities as good as possible. In IPC and IPDA models, the proximity is given by an indicator matrix that corresponds to the multinomial response.

De Rooij (2009a) showed that the IPC model in maximum dimensionality is equivalent to the MBCL model, i.e., if the dimensionality of the Euclidean space equals the number of categories of the response variable minus one. The MBCL is a natural extension of binary logistic regression to the case of nominal categorical variables. Both the IPC and the MBCL models use the joint probabilities to define their likelihood function. Unlike in the MBCL model, dimension reduction is possible in the IPC models. Thus, less model parameters are estimated in the reduced space. Furthermore, the results of the IPC model can be displayed using a biplot (Gower & Hand, 1996; Gower et al., 2011) which enhance interpretation of the model.

In this paper, our main aim is to study properties of the IPC model for bivariate binary data, specifically about the representation of the marginal probabilities and of the association structure. We will show that the IPC model either represents the marginal models or the association structure well. Next, we study a new parametrization of the IPC model, namely the Bivariate IPC (BIPC) model, in which both the marginal probabilities and the association structure are represented. This new model builds forward on the work of Bahadur (1961) and Lipsitz, Laird and Harrington (1990). Compared to this existing methodology for jointly modelling the marginal and association structure, our method has the advantage of dimension reduction and a graphical representation of the model using a biplot.

The paper is organized as follows. Section 2 presents the theoretical background. Section 3 studies properties of the IPC models both mathematically and with a simulation study. Section 4 proposes the BIPC model. Section 5 shows an example application and then we conclude in Section 6 with a discussion.

3.2 Background

3.2.1 The Ideal Point Classification Model

In the IPC model (De Rooij, 2009a) the conditional joint probabilities, i.e., $\pi_j(\mathbf{x}_i) = \Pr(G_i = j | \mathbf{x}_i)$, are modelled using a distance between two points in an Euclidean space of dimensionality M : one point representing subject i with coordinates $\boldsymbol{\eta}_i = [\eta_{i1}, \dots, \eta_{iM}]^T$, and the other representing class j with coordinates $\boldsymbol{\gamma}_j = [\gamma_{j1}, \dots, \gamma_{jM}]^T$. The smaller the relative distance between the two points, the larger the probability that the subject belongs to that class. The IPC model is defined as (De Rooij, 2009a),

$$\pi_j(\mathbf{x}_i) = \frac{\exp(-0.5 \times \delta_{ij})}{\sum_h \exp(-0.5 \times \delta_{ih})}, \quad (3.1)$$

where δ_{ij} is a squared Euclidean distance between the two points and is defined as

$$\delta_{ij} = \sum_{m=1}^M (\eta_{im} - \gamma_{jm})^2. \quad (3.2)$$

The coordinates of the subject points are assumed to be a linear combination of the predictor variables $\mathbf{x}_i$ and an intercept, i.e., $\boldsymbol{\eta}_i = \boldsymbol{\beta}_0 + \mathbf{x}_i \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $(p \times M)$ matrix with regression weights and, $\boldsymbol{\beta}_0$ an M dimensional intercept. The parameters of this model are the regression weights and the class points.

Parameter estimates in the IPC model can be obtained by maximizing a multinomial log-likelihood function

$$\sum_{i=1}^N \left[\log \left(\prod_j \pi_j(\mathbf{x}_i)^{f_{ij}} \right) \right], \quad (3.3)$$

where $f_{ij} = 1$ if subject i is in category j , zero otherwise.

The IPC model has translation, rotational freedom, and multinomial indeterminacy (i.e., the class probability remains the same if a constant is added to each subject's squared

distance). The total number of restrictions needed is $\max[M(M-1)/2, M(M+1) - (g-1)]$, and thus the total number of free parameters becomes $\text{npar} = (p+g)M - \max[M(M-1)/2, M(M+1) - (g-1)]$ (De Rooij, 2009a). Depending on dimensionality of the fitted model, γ -parameters are set at fixed values to identify the model. For a multinomial response variable with $g = 4$ categories, for example, the maximum dimensionality of the IPC model is $M = 3 (= g - 1)$ and the total number of parameters in that case will be $\text{npar} = 3 \times (p + 1)$ that corresponds to the regression parameters only since the class points can be set to fixed values that span the three-dimensional space. The class point coordinates can be specified, for example, as

$$\gamma = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}. \quad (3.4)$$

The rows in (3.4) correspond to the response categories and the columns to the dimensions. In this case, the IPC model is equivalent to the MBCL model. The advantage of the IPC model over the MBCL model is that it provides the possibility of dimension reduction. For the multinomial response with $g = 4$, a 2-dimensional IPC model can be fitted with a total number of parameters $\text{npar} = 2 \times (p + 1) + 3$, where the first part ($2 \times (p + 1)$) represents the number of regression coefficients and the second part ($+3$) the free class coordinates. From the eight class coordinates five need to be fixed for identification. This can be accomplished, for example, by defining

$$\gamma = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \gamma_{32} \\ \gamma_{41} & \gamma_{42} \end{bmatrix}, \quad (3.5)$$

where γ_{32} , γ_{41} , and γ_{42} are the free class coordinates, i.e., these can be estimated from the data.

3.2.2 The 2-step Approach of McCullagh and Nelder (1989)

We revisit a 2-step approach often used for constructing multivariate regression models using joint probabilities of multivariate (or bivariate) binary data, as proposed by McCullagh and Nelder (1989). We later apply this approach in the distance framework to study the properties of IPC models.

In the first step, a linear transformation is applied on the joint probabilities to obtain the marginal probabilities, i.e.,

$$\Lambda_i = \mathbf{L}\pi_i, \quad (3.6)$$

where $\mathbf{L}$ is a matrix of zeros and ones and $\pi_i = [\pi_{i4} \ \pi_{i3} \ \pi_{i2} \ \pi_{i1}]^T$. In the case of bivariate binary data, for example, the row margin is given by

$$\begin{aligned} \Lambda_{i1} &= \mathbf{L}_1 \pi_i \\ &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \times [\pi_{i4} \ \pi_{i3} \ \pi_{i2} \ \pi_{i1}]^T \\ &= \begin{bmatrix} \pi_{i4} + \pi_{i2} \\ \pi_{i3} + \pi_{i1} \end{bmatrix} = \begin{bmatrix} \pi_{i1\bullet} \\ \pi_{i0\bullet} \end{bmatrix}. \end{aligned} \quad (3.7)$$

Similarly, the column margin is given by

$$\begin{aligned}
 \mathbf{\Lambda}_{i2} &= \mathbf{L}_2 \boldsymbol{\pi}_i \\
 &= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \times [\pi_{i4} \ \pi_{i3} \ \pi_{i2} \ \pi_{i1}]^T \\
 &= \begin{bmatrix} \pi_{i4} + \pi_{i3} \\ \pi_{i2} + \pi_{i1} \end{bmatrix} = \begin{bmatrix} \pi_{i \cdot 1} \\ \pi_{i \cdot 0} \end{bmatrix}.
 \end{aligned} \tag{3.8}$$

In the second step, logarithmic contrasts of interest are formulated, i.e.,

$$\boldsymbol{\Psi}_i = \mathbf{C}^T \log[\mathbf{\Lambda}_i], \tag{3.9}$$

for an appropriately chosen contrast matrix $\mathbf{C}^T$. For the bivariate binary data, the contrast matrices can be chosen to be $\mathbf{C}^T = \begin{bmatrix} 1 & -1 \end{bmatrix}$. Thus,

$$\begin{aligned}
 \psi_{i1} &= \begin{bmatrix} 1 & -1 \end{bmatrix} \log[\mathbf{\Lambda}_{i1}] \\
 &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \log(\pi_{i1 \cdot}) & \log(\pi_{i0 \cdot}) \end{bmatrix}^T \\
 &= \log(\pi_{i1 \cdot}) - \log(\pi_{i0 \cdot}) \\
 &= \log(\pi_{i1 \cdot} / \pi_{i0 \cdot}) \\
 &= \text{logit}(\pi_{i1 \cdot}).
 \end{aligned} \tag{3.10}$$

Similarly, $\psi_{i2} = \log(\pi_{i \cdot 1} / \pi_{i \cdot 0}) = \text{logit}(\pi_{i \cdot 1})$. In the presence of predictors these logits can be linked to the systematic part as used in Generalized Linear Models (Agresti, 2002); that is,

$$\begin{aligned}
 \text{logit}(\pi_{i1 \cdot}) &= \beta_{01} + \boldsymbol{\beta}_1^T \mathbf{x}_i, \\
 \text{logit}(\pi_{i \cdot 1}) &= \beta_{02} + \boldsymbol{\beta}_2^T \mathbf{x}_i.
 \end{aligned} \tag{3.11}$$

The above derivations (equation 3.6 - 3.11) can be summarized as follows.

$$\begin{aligned}\mathbf{C}^T \log(\mathbf{L}_1 \boldsymbol{\pi}_i) &= \beta_{01} + \boldsymbol{\beta}_1^T \mathbf{x}_i, \\ \mathbf{C}^T \log(\mathbf{L}_2 \boldsymbol{\pi}_i) &= \beta_{02} + \boldsymbol{\beta}_2^T \mathbf{x}_i.\end{aligned}\tag{3.12}$$

To obtain the association structure for bivariate binary data, the joint probabilities can also be transformed linearly. In this case $\mathbf{C}^T = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}$ and $\mathbf{L} = \mathbf{I}$ such that,

$$\begin{aligned}\mathbf{C}^T \log(\mathbf{L} \boldsymbol{\pi}_i) &= \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix} \log[\mathbf{I} \boldsymbol{\pi}_i] \\ &= \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \log(\pi_{i4}) & \log(\pi_{i3}) & \log(\pi_{i2}) & \log(\pi_{i1}) \end{bmatrix}^T \\ &= \log(\pi_{i4}) - \log(\pi_{i3}) - \log(\pi_{i2}) + \log(\pi_{i1}) \\ &= \log \left[\frac{\pi_{i4} \times \pi_{i1}}{\pi_{i3} \times \pi_{i2}} \right] \\ &= \log(\tau_i).\end{aligned}\tag{3.13}$$

This odds ratio can be linked to predictors as

$$\log(\tau_i) = \beta_{03} + \boldsymbol{\beta}_3^T \mathbf{x}_i.\tag{3.14}$$

3.3 Study-1: IPC Model as a Marginal Model

In this section, our aim is in how the IPC model represents both the marginal probabilities and the association structure for bivariate binary data. We use the 2-step approach of McCullagh and Nelder (1989) within the distance framework to transform the joint probabilities into the marginal probabilities and the association structure.

3.3.1 The 2-dimensional IPC Model

In this section, we show the representation of both the marginal probabilities and the association structure by a 2-dimensional IPC model. The class point matrix introduced in equation (3.5) will be used here with an additional restriction imposed on one of the free class points. That is, $\gamma_{32} = 1$ so that the first dimension pertains to a logistic regression of the first response and the second dimension to a logistic regression of the second response (i.e., no further scaling is required).

Representation of the Marginal Probabilities

Let us first show how the marginal probabilities of the two binary responses are represented by the 2-dimensional IPC model. The joint probability as defined by the IPC model in equation (3.1) will be used to define the marginal probabilities, that is,

$$\begin{aligned}
 \log \left[\frac{\pi_{i1\bullet}}{\pi_{i0\bullet}} \right] &= \log \left[\frac{\pi_{i4} + \pi_{i2}}{\pi_{i3} + \pi_{i1}} \right] \\
 &= \log \left[\frac{\frac{\exp(-0.5\delta_{i4})}{\sum_h \exp(-0.5\delta_{ih})} + \frac{\exp(-0.5\delta_{i2})}{\sum_h \exp(-0.5\delta_{ih})}}{\frac{\exp(-0.5\delta_{i3})}{\sum_h \exp(-0.5\delta_{ih})} + \frac{\exp(-0.5\delta_{i1})}{\sum_h \exp(-0.5\delta_{ih})}} \right] \\
 &= \log \left[\frac{\exp(-0.5\delta_{i4}) + \exp(-0.5\delta_{i2})}{\exp(-0.5\delta_{i3}) + \exp(-0.5\delta_{i1})} \right].
 \end{aligned} \tag{3.15}$$

Let us write out the Euclidean distances δ_{ij} as defined in equation (3.2). The marginal model (3.15) becomes,

$$\log \left[\frac{\pi_{i1\bullet}}{\pi_{i0\bullet}} \right] = \log \left[\frac{\exp[\gamma_{41}(\eta_{i1} - 0.5\gamma_{41})] \times \exp[\gamma_{42}(\eta_{i2} - 0.5\gamma_{42})] + \exp[\eta_{i1} - 0.5]}{\exp(\eta_{i2} - 0.5) + 1} \right]. \tag{3.16}$$

In this paper, we find it convenient to re-parametrize γ_{41} and γ_{42} in terms of two other parameters, i.e., $\gamma_{41} = 1 + \phi_1$ and $\gamma_{42} = 1 + \phi_2$. The ϕ -parameters represent the deviation of the last category from (1, 1). By setting $\phi_1 = \phi_2 = 0$, the above result (16) can be simplified to:

$$\begin{aligned}
 \text{logit}[\pi_{i1\bullet}] &= \log \left[\frac{[\exp(\eta_{i1} - 0.5) \times \exp(\eta_{i2} - 0.5)] + \exp(\eta_{i1} - 0.5)}{\exp(\eta_{i2} - 0.5) + 1} \right] \\
 &= \log \left[\frac{\exp(\eta_{i1} - 0.5) \times [\exp(\eta_{i2} - 0.5) + 1]}{\exp(\eta_{i2} - 0.5) + 1} \right] \\
 &= \eta_{i1} - 0.5 \\
 &= (\beta_{01} - 0.5) + \beta_1^T \mathbf{x}_i \\
 &= \beta_{01}^* + \beta_1^T \mathbf{x}_i.
 \end{aligned} \tag{3.17}$$

Similarly,

$$\begin{aligned}
 \log \left[\frac{\pi_{i\bullet 1}}{\pi_{i\bullet 0}} \right] &= \log \left[\frac{\pi_{i4} + \pi_{i3}}{\pi_{i2} + \pi_{i1}} \right] \\
 &= \log \left[\frac{\frac{\exp(-0.5\delta_{i4})}{\sum_h \exp(-0.5\delta_{ih})} + \frac{\exp(-0.5\delta_{i3})}{\sum_h \exp(-0.5\delta_{ih})}}{\frac{\exp(-0.5\delta_{i2})}{\sum_h \exp(-0.5\delta_{ih})} + \frac{\exp(-0.5\delta_{i1})}{\sum_h \exp(-0.5\delta_{ih})}} \right] \\
 &= \log \left[\frac{\exp[\gamma_{41}(\eta_{i1} - 0.5\gamma_{41})] \times \exp[\gamma_{42}(\eta_{i2} - 0.5\gamma_{42})] + \exp[\eta_{i2} - 0.5]}{\exp[\eta_{i1} - 0.5] + 1} \right].
 \end{aligned} \tag{3.18}$$

By setting $\phi_1 = \phi_2 = 0$ a straightforward marginal model is obtained, $\text{logit}[\pi_{i\bullet 1}] = (\beta_{02} - 0.5) + \beta_2^T \mathbf{x}_i = \beta_{02}^* + \beta_2^T \mathbf{x}_i$; and, thus we call this the fixed class case. Without the constraints on the ϕ -parameters, the marginal models in (3.16) and (3.18) can not

be simplified further.

Representation of the Association

The odds ratio is defined in terms of the joint probabilities as shown in (3.13). Let us rewrite the probabilities in terms of the IPC model as in equation (3.1); that is,

$$\begin{aligned}
 \log(\tau_i) &= \log \left[\frac{\pi_{i4} \times \pi_{i1}}{\pi_{i2} \times \pi_{i3}} \right] \\
 &= \log \left[\frac{\frac{\exp(-0.5\delta_{i4})}{\sum_h \exp(-0.5\delta_{ih})} \times \frac{\exp(-0.5\delta_{i1})}{\sum_h \exp(-0.5\delta_{ih})}}{\frac{\exp(-0.5\delta_{i2})}{\sum_h \exp(-0.5\delta_{ih})} \times \frac{\exp(-0.5\delta_{i3})}{\sum_h \exp(-0.5\delta_{ih})}} \right] \\
 &= \log \left[\frac{\exp(-0.5\delta_{i4}) \times \exp(-0.5\delta_{i1})}{\exp(-0.5\delta_{i2}) \times \exp(-0.5\delta_{i3})} \right] \\
 &= 0.5 \times [\delta_{i2} + \delta_{i3} - \delta_{i4} - \delta_{i1}]. \tag{3.19}
 \end{aligned}$$

This result implies that the differences between pairs of squared Euclidean distances correspond to the log-odds ratio. The distances can be written out and the association model becomes,

$$\log(\tau_i) = \phi_1 \times (\eta_{i1} - 1) + \phi_2 \times (\eta_{i2} - 1) - 0.5 * (\phi_1^2 + \phi_2^2). \tag{3.20}$$

In the case of $\phi_1 = \phi_2 = 0$, $\log(\tau_i) = 0$ which is equal to $\tau_i = 1$. An odds ratio of unity indicates no association between the two binary responses, i.e., independence.

3.3.2 The 3-dimensional IPC Model

We now show the representation of the marginal probabilities and the association structure in a 3-dimensional IPC model. The class point introduced in equation (3.4) will be used

in the next derivations of the 3-dimensional IPC model.

Representation of the Marginal Probabilities

We follow the same derivation as before, but now the joint probabilities are defined in the 3-dimensional Euclidean space. For the marginal probabilities, we have

$$\begin{aligned}
 \log \left[\frac{\pi_{i1\bullet}}{\pi_{i0\bullet}} \right] &= \log \left[\frac{\pi_{i4} + \pi_{i2}}{\pi_{i3} + \pi_{i1}} \right] \\
 &= \log \left[\frac{\frac{\exp(-0.5 \times \delta_{i4})}{\sum_h \exp(-0.5 \times \delta_{ih})} + \frac{\exp(-0.5 \times \delta_{i2})}{\sum_h \exp(-0.5 \times \delta_{ih})}}{\frac{\exp(-0.5 \times \delta_{i3})}{\sum_h \exp(-0.5 \times \delta_{ih})} + \frac{\exp(-0.5 \times \delta_{i1})}{\sum_h \exp(-0.5 \times \delta_{ih})}} \right] \\
 &= \log \left[\frac{\exp[\eta_{i1} + \eta_{i2} + \eta_{i3} - (3/2)] + \exp[\eta_{i1} - 0.5]}{\exp[\eta_{i2} - 0.5] + 1} \right]. \quad (3.21)
 \end{aligned}$$

Similarly,

$$\begin{aligned}
 \log \left[\frac{\pi_{i\bullet 1}}{\pi_{i\bullet 0}} \right] &= \log \left[\frac{\pi_{i4} + \pi_{i3}}{\pi_{i2} + \pi_{i1}} \right] \\
 &= \log \left[\frac{\frac{\exp(-0.5 \times \delta_{i4})}{\sum_h \exp(-0.5 \times \delta_{ih})} + \frac{\exp(-0.5 \times \delta_{i3})}{\sum_h \exp(-0.5 \times \delta_{ih})}}{\frac{\exp(-0.5 \times \delta_{i2})}{\sum_h \exp(-0.5 \times \delta_{ih})} + \frac{\exp(-0.5 \times \delta_{i1})}{\sum_h \exp(-0.5 \times \delta_{ih})}} \right] \\
 &= \log \left[\frac{\exp[\eta_{i1} + \eta_{i2} + \eta_{i3} - (3/2)] + \exp[\eta_{i2} - 0.5]}{\exp[\eta_{i1} - 0.5] + 1} \right]. \quad (3.22)
 \end{aligned}$$

It is not possible to simplify the above formulas further because of the parameters η_{i3} . Compared to the 2-dimensional IPC model with *fixed* class point, the marginal models

are not clearly represented in the 3-dimensional IPC models.

Representation of the Association

Using the formula derived in equation (3.19), but with the distances defined in three dimensions, the association model becomes

$$\begin{aligned}
 \log[\tau_i] &= 0.5 \times [\delta_{i2} + \delta_{i3} - \delta_{i4} - \delta_{i1}] \\
 &= 0.5 \times \left\{ \left[\sum_{m=1}^3 (\eta_{im}^2 - 2\eta_{i1} + 1) \right] + \left[\sum_{m=1}^3 (\eta_{im}^2 - 2\eta_{i2} + 1) \right] \right. \\
 &\quad \left. - \left[\sum_{m=1}^3 (\eta_{im}^2 - 2\eta_{i1} - 2\eta_{i2} - 2\eta_{i3} + 3) \right] - \left[\sum_{m=1}^3 \eta_{im}^2 \right] \right\} \\
 &= \eta_{i3} - 0.5.
 \end{aligned} \tag{3.23}$$

This result proves that the 3-dimensional IPC model represents the association structure where the third dimension uniquely pertains to the association model.

3.3.3 Discussion

We studied both 2- and 3-dimensional IPC models in terms of marginal probabilities and association structure of bivariate binary data in the presence of predictors. We showed that both models with a specific class point specification are able to recover either the marginal probabilities or the association structure. That is, the 2-dimensional IPC model with *fixed* class point, i.e., $\phi_1 = \phi_2 = 0$, is equivalent to the marginal model with an *independence* association structure. In the case of a 3-dimensional model, the association structure is represented by the third dimension.

Based on the results of Section 3.1.1 and 3.1.2, we showed that a 2-dimensional IPC model with *fixed* class points, i.e., $\gamma_{41} = \gamma_{42} = 1$, represents a marginal model with an *independence* association structure. Each of the dimensions in the IPC model is related to one of the two binary responses. As shown in equation (3.20), the 2-dimensional IPC

model with free ϕ -parameters represents the association structure by a mixture of the marginal parameters and the ϕ -parameters.

According to the analytical results shown in equations (3.16) and (3.18), the marginal models can not be further simplified unless $\phi_1 = \phi_2 = 0$. When $\phi_1 \neq 0$ and $\phi_2 \neq 0$, neither the marginal model nor the association structure is well represented. At this stage, however, we do not know whether the IPC model is capable of recovering the models for the marginal probabilities and the association structure; therefore, we conducted a simulation study.

3.3.4 Simulation Study

We were able to show mathematically the performance of both the 2-dimensional IPC model with *fixed* class point, denoted by IPC(2D-FIXED), and the 3-dimensional IPC model, denoted by IPC(3D), in representing the marginal probabilities and the association structure for bivariate binary data. The analytical derivation under the 2-dimensional IPC model with *free* class points, denoted by IPC(2D-FREE), however, was cumbersome. We conducted a simulation study to fully understand to what degree the IPC(2D-FREE) model recovers the marginal models and/ or the association model.

Data-generating Model

Bivariate binary data were generated from a Bivariate Logistic Regression model (Palmgren, 1989). The data generating model for the marginal probabilities is defined as follows,

$$\begin{aligned}\text{logit}[\pi_{i1\cdot}] &= \beta_{01} + \beta_{11}X_{1i} + \beta_{21}X_{2i} + \beta_{31}X_{3i} + \beta_{41}X_{4i} + \beta_{51}X_{5i}, \\ \text{logit}[\pi_{i\cdot 1}] &= \beta_{02} + \beta_{12}X_{1i} + \beta_{22}X_{2i} + \beta_{32}X_{3i} + \beta_{42}X_{4i} + \beta_{52}X_{5i}.\end{aligned}\tag{3.24}$$

We set $(\beta_{01}, \beta_{02}) = (-2.20, -1.50)$; $(\beta_{11}, \beta_{12}) = (0.00, -0.25)$; $(\beta_{21}, \beta_{22}) = (0.20, 0.00)$; $(\beta_{31}, \beta_{32}) = (-0.15, -0.15)$; $(\beta_{41}, \beta_{42}) = (1.05, 1.15)$; and $(\beta_{51}, \beta_{52}) = (-0.45, -0.15)$.

To generate data we need a representation of the association structure, i.e., $\log[\tau_i] = \beta_{03} + \beta_{13}X_{1i} + \beta_{23}X_{2i} + \beta_{33}X_{3i} + \beta_{43}X_{4i} + \beta_{53}X_{5i}$. In the 2-dimensional IPC model, the association structure is defined in terms of the other parameters as shown in (3.20). That is, $\beta_{03}^* = \phi_1 \times \beta_{01} + \phi_2 \times \beta_{02} - 0.5 \times \phi_1^2 - 0.5 \times \phi_2^2 - \phi_1 - \phi_2$ and $\beta_{k3}^* = \phi_1 \times \beta_{k1} + \phi_2 \times \beta_{k2}$, where $k = 1, 2, \dots, 5$. Therefore, the data generating model for the association is $\log[\tau_i] = \beta_{03}^* + \beta_{13}^*X_{1i} + \beta_{23}^*X_{2i} + \beta_{33}^*X_{3i} + \beta_{43}^*X_{4i} + \beta_{53}^*X_{5i}$. We set $\phi_1 = -0.20$ and $\phi_2 = -0.45$; thus, the association parameters become $\beta_{03}^* = 1.65$ and $\beta_{k3}^* = (0.10, -0.05, 0.10, -0.70, 0.15)$.

Four of the predictors were generated from the standard normal distribution, $X_{qi} \sim N(0, 1)$ where $q = 2, \dots, 5$, and one from a binomial distribution, i.e., $X_{1i} \sim \text{BIN}(0.67)$. The VGAM package in the R software was used for generating the bivariate binary data (Yee, 2010).

Design and Analysis

A sample size of $N = 500$ was used in the simulations and each simulation was replicated $R = 1000$ times to obtain the sampling distributions of model parameters.

The performance of the proposed methods was evaluated by Bias (B), Root Mean Squared Error (RMSE), and Coverage. The bias of a parameter is defined as the difference between true value and the average of estimated values, i.e., $B(\hat{\beta}) = \bar{\hat{\beta}} - \beta$, with

$$\bar{\hat{\beta}} = \sum_{r=1}^{1000} \hat{\beta}_r / 1000,$$

and $\hat{\beta}_r$ is the estimate obtained from r -th replication. The RMSE is defined as

$$\text{RMSE} = \sqrt{\sum_{r=1}^{1000} [(\hat{\beta}_r - \beta)^2 / 1000]}.$$

Finally, the coverage is defined as the proportion of times the $100(1 - \alpha)\%$ confidence

interval (CI) includes the true β value, where α corresponds to the nominal level of significance. The CI is defined as $[\hat{\beta}_r \pm Z_{1-\alpha/2} \widehat{SE}(\hat{\beta}_r)]$ in which SE stands for the standard error of a parameter.

Simulation Study Results

The simulation results of the 2- and 3-dimensional IPC models are summarized in Table 3.2. The results for IPC(2D-FIXED) are given in columns 4-6, for IPC(3D) in columns 7-9, and for IPC(2D-FREE) in the last three columns. Because we showed analytically that the marginal models are represented well by the 2-dimensional fixed IPC model, and the association structure is represented well by the 3-dimensional IPC model, we focus here on the contrast of the 2-dimensional free model with the other two.

Compared to the IPC(2D-FIXED) results, marginal parameters under the IPC(2D-FREE) model were more biased. Specifically, two of the effects (i.e., X_2 and X_4) including the intercept, were poorly estimated. More specifically, $B(\beta_{21}) = 0.037$ is about nine times bigger compared to the IPC(2D-FIXED) result, $B(\beta_{22}) = -0.016$, $B(\beta_{41}) = 0.106$, and $B(\beta_{42}) = 0.050$ which all are about three times bigger than those obtained from the IPC(2D-FIXED). All the RMSE results for the IPC(2D-FREE) model were higher than those obtained from the IPC(2D-FIXED) model. The coverage of the marginal parameters by the IPC(2D-FREE) model, compared to the former results, seems promising. However, both the intercepts and some of the effects were not covered well (i.e., β_{01} : 85.2%; β_{02} : 91.0%; β_{21} : 92.5%; β_{41} : 92.6%; β_{52} : 91.9%). Unlike the marginal parameters, the association parameters were fairly well estimated by the IPC(2D-FREE). This is evident if we compare the results of the association parameters under the IPC(2D-FREE) and the IPC(3D) models.

Table 3.2: Summarized results of the simulation study for studying the performance of the IPC model for analysing bivariate binary data. IPC(2D-FIXED) corresponds to the 2-dimensional IPC model with fixed class points, i.e., $\phi_1 = \phi_2 = 0$; IPC(3D) to the 3-dimensional IPC model; and IPC(2D-FREE) to the 2-dimensional IPC model with free class points.

Effect	Parameter	True	IPC (2D-FIXED)				IPC (3D)				IPC (2D-FREE)*			
			Bias	RMSE	Coverage		Bias	RMSE	Coverage		Bias	RMSE	Coverage	
Intercept	β_{01}	-2.20	-0.083	0.337	96.3		-0.487	0.652	88.3		-0.461	0.617	85.2	
	β_{02}	-1.50	-0.044	0.260	94.8		-0.236	0.368	91.5		-0.236	0.362	91.0	
	β_{03}	1.65	—	—	—		-0.045	0.786	94.4		0.020	0.586	94.8	
X_1	β_{11}	0.00	0.018	0.373	94.4		0.079	0.486	95.9		0.040	0.456	94.9	
	β_{12}	-0.25	-0.008	0.287	96.0		-0.024	0.335	95.7		-0.020	0.323	95.2	
	β_{13}	0.10	—	—	—		-0.076	0.717	96.2		-0.031	0.411	98.9	
X_2	β_{21}	0.20	0.004	0.174	93.0		0.031	0.228	94.2		0.037	0.215	92.5	
	β_{22}	0.00	-0.006	0.144	93.9		-0.026	0.163	95.9		-0.016	0.158	94.8	
	β_{23}	-0.05	—	—	—		0.001	0.372	93.0		-0.035	0.238	95.8	
X_3	β_{31}	-0.15	-0.009	0.167	95.2		-0.011	0.206	95.5		-0.015	0.195	95.5	
	β_{32}	-0.15	-0.005	0.136	96.3		-0.004	0.160	96.1		-0.009	0.151	95.9	
	β_{33}	0.10	—	—	—		-0.027	0.341	96.6		-0.007	0.172	98.9	
X_4	β_{41}	1.05	0.033	0.198	94.7		0.065	0.255	95.1		0.106	0.270	92.6	
	β_{42}	1.15	0.019	0.178	94.6		0.034	0.207	94.2		0.050	0.201	95.4	
	β_{43}	-0.70	—	—	—		0.083	0.430	93.0		-0.023	0.308	95.6	
X_5	β_{51}	-0.45	-0.001	0.163	96.4		-0.032	0.212	96.0		-0.040	0.205	95.9	
	β_{52}	-0.15	-0.004	0.149	93.5		0.033	0.175	92.9		0.012	0.173	91.9	
	β_{53}	0.15	—	—	—		-0.034	0.352	95.8		0.035	0.240	96.9	

* $\beta_{03} = \phi_1 \times \beta_{01} + \phi_2 \times \beta_{02} - 0.5 \times \phi_1^2 - \phi_1 - \phi_2$; $\beta_{k3}^* = \phi_1 \times \beta_{k1} + \phi_2 \times \beta_{k2}$, where $k = 1, 2, \dots, 5$.

3.3.5 Summary of Study-1

De Rooij (2009a) studied IPC model for categorical data and showed its equivalence to logistic regression models. It was shown that the MBCL model is equivalent to the IPC model in maximum dimensionality. These models represent the joint probabilities.

In this Section we studied properties of the IPC model and extended it for analysing bivariate binary data, focusing on the marginal probabilities and the association structure. We showed their connection both mathematically and using a simulation study. We found that a 2-dimensional IPC model with *fixed* class point (i.e., $\phi_1 = \phi_2 = 0$) represents the marginal models with an *independence* association structure. We also found that a 3-dimensional IPC model with a specific class point configuration represents the association model in the third dimension.

We also studied the performance of a 2-dimensional IPC model with *free* class point. Since its analytical part was cumbersome, we conducted a simulation study to see if it can recover both the marginal models and the association model. This model represents the association model well, but the marginal models were misspecified. Therefore, we conclude that a given IPC model can recover either the marginal models or the association model of bivariate binary data, but not both of them at the same time.

3.4 Study-2: The Bivariate IPC Model

In the first study, we investigated properties of the standard IPC models for the representation of both the marginal probabilities and the association structure. It was concluded that a given IPC model is not able to represent both types of the models at the same time. In this section, we re-parametrize the IPC model in order to provide a better representation of both the marginal probabilities and the association structure.

Bahadur (1961) proposed a full likelihood-based marginal model for bivariate binary data by characterizing the multinomial probabilities in terms of both the marginal prob-

abilities and the correlation coefficient between the two responses (Y_{i1} and Y_{i2}). Lipsitz, Laird and Harrington (1990) followed the Bahadur (1961) approach and showed that other measures of association, such as the odds ratio and the relative risk, can also be used.

In this second study, our aim is to adopt the Lipsitz, Laird and Harrington (1990) approach into the IPC model framework for better representation of the required statistical models. As shown in equation (3.3), parameter estimation under the IPC model is based on the multinomial likelihood function. To avoid confusion with the former IPC model presented in Section 2.1, we refer the Bahadur-based IPC model as the Bivariate IPC (BIPC) model.

In the BIPC model framework, the Euclidean distance defined in equation (3.2) will be used only to define the joint probabilities which are related to the association structure. For defining the marginal models, we use another Euclidean distance definition emphasizing the marginal models. That is,

$$\begin{aligned}\pi_{i1\cdot} &= \frac{\exp(-0.5\delta_{i1\cdot})}{\exp(-0.5\delta_{i0\cdot}) + \exp(-0.5\delta_{i1\cdot})}; \\ \pi_{i\cdot 1} &= \frac{\exp(-0.5\delta_{i\cdot 1})}{\exp(-0.5\delta_{i\cdot 0}) + \exp(-0.5\delta_{i\cdot 1})},\end{aligned}\tag{3.25}$$

where $\delta_{il\cdot} = \sum_{m=1}^2 (\eta_{im} - \gamma_{l\cdot m})^2$ and $\delta_{i\cdot l} = \sum_{m=1}^2 (\eta_{im} - \gamma_{\cdot lm})^2$, $l = 0, 1$. As shown in Appendix B, the class points of the BIPC model are defined as

$$\gamma_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$

and

$$\gamma_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

where γ_1 is the class point matrix that corresponds to the first response variable and γ_2

to the second response variable.

The first step according to Bahadur (1961) is to rewrite the the association structure between the two binary responses, i.e., the odds ratio in our case, using the marginal probabilities. That is,

$$\tau_i = \frac{\pi_{i4} \times \pi_{i1}}{\pi_{i2} \times \pi_{i3}} = \frac{\pi_{i4} \times (1 - \pi_{i1\cdot} - \pi_{i\cdot 1} + \pi_{i4})}{(\pi_{i1\cdot} - \pi_{i4}) \times (\pi_{i\cdot 1} - \pi_{i4})}. \quad (3.26)$$

We showed in (3.19) that given the IPC model, the odds ratio can be defined in terms of Euclidean distances, i.e., $\tau_i = \exp[0.5 \times (\delta_{i2} + \delta_{i3} - \delta_{i1} - \delta_{i4})]$. We will use this representation as the defining characteristics of the association in the BIPC model. With free class points, i.e., $\phi_1 \neq 0$ and $\phi_2 \neq 0$, the odds ratio becomes,

$$\tau_i = \exp[\phi_1 \times (\eta_{i1} - 0.5\phi_1 - 1) + \phi_2 \times (\eta_{i2} - 0.5\phi_2 - 1)]. \quad (3.27)$$

We can then replace τ_i in (3.26) by (3.27), and solve the quadratic equation to get solutions for π_{i4} (Mardia, 1967). The valid solution for π_{i4} is,

$$\pi_{i4} = \begin{cases} \frac{w_i - \{w_i^2 - 4 \exp(a_i)[\exp(a_i) - 1]\pi_{i1\cdot}\pi_{i\cdot 1}\}^{1/2}}{2[\exp(a_i) - 1]} & \text{if } a_i \neq 0 \\ \pi_{i1\cdot} \times \pi_{i\cdot 1} & \text{if } a_i = 0, \end{cases} \quad (3.28)$$

where $w_i = 1 - [1 - \exp(a_i)][\pi_{i1\cdot} + \pi_{i\cdot 1}]$ and $a_i = \phi_1 \times (\eta_{i1} - 0.5\phi_1 - 1) + \phi_2 \times (\eta_{i2} - 0.5\phi_2 - 1)$.

The final step is to rewrite the joint probabilities in the multinomial likelihood in terms of the marginal probabilities and the association structure, i.e., $\pi_{i2} = \pi_{i1\cdot} - \pi_{i4}$

and $\pi_{i3} = \pi_{i\cdot 1} - \pi_{i4}$ in which π_{i4} will be replaced by (3.28). That is,

$$\pi_i^* = \begin{bmatrix} \pi_{i4} \\ \pi_{i\cdot 1} - \pi_{i4} \\ \pi_{i1\cdot} - \pi_{i4} \end{bmatrix}. \quad (3.29)$$

This modified likelihood will be used for estimating the parameters of the BIPC model.

3.4.1 Simulation Study Results

The simulation results of the BIPC model are summarized in Table 3.3. We compare these results against those in Table 3.2, particularly the results from IPC(2D-FIXED) and IPC(3D) models.

The bias and RMSE results for the marginal parameters under the BIPC model are very close to those under the IPC(2D-FIXED) model, which proves that the BIPC model represents the marginal models well. Almost all the coverages of the marginal parameters were satisfactory, except two of the effects, one for β_{22} equal to 92.8% and for β_{52} equal to 92.2%. Their coverage by the IPC(2D-FIXED) model was 93.9% and 93.5%, respectively.

Compared to the results presented in Table 3.2 for IPC(3D), the BIPC model produced smaller bias, except for two of the effects, i.e., $B(\beta_{43}) = -0.176$ and $B(\beta_{53}) = 0.087$. However, all the RMSEs under the BIPC model were smaller than those obtained from the IPC models. Almost all the parameters were covered well by the BIPC model, with a coverage above 95.0%. Compared to the IPC models, the BIPC model estimates are generally less biased, more accurate, and well covered parameters for both the marginal models and the association model.

We conclude that the BIPC model represents not only the marginal models, but also the association model for the analysis of bivariate binary data in the presence of predictors.

Table 3.3: Summarized results of the simulation study for studying the performance of the BIPC model for analysing bivariate binary data.

Effect	Parameter	True	Bias	RMSE	Coverage
Intercept	β_{01}	-2.20	-0.074	0.333	95.5
	β_{02}	-1.50	-0.048	0.262	94.9
	β_{03}^*	1.65	0.109	0.602	94.5
X_1	β_{11}	0.00	0.021	0.365	94.3
	β_{12}	-0.25	0.001	0.288	95.4
	β_{13}^*	0.10	-0.022	0.369	98.7
X_2	β_{21}	0.20	0.005	0.171	94.2
	β_{22}	0.00	-0.010	0.144	92.8
	β_{23}^*	-0.05	-0.044	0.230	95.1
X_3	β_{31}	-0.15	-0.007	0.162	95.2
	β_{32}	-0.15	0.004	0.137	96.1
	β_{33}^*	0.10	0.013	0.158	99.3
X_4	β_{41}	1.05	0.022	0.195	94.5
	β_{42}	1.15	0.009	0.179	93.1
	β_{43}^*	-0.70	-0.176	0.382	95.7
X_5	β_{51}	-0.45	0.007	0.162	96.5
	β_{52}	-0.15	0.001	0.149	92.2
	β_{53}^*	0.15	0.087	0.249	96.8

$$\beta_{03}^* = \phi_1 \times \beta_{01} + \phi_2 \times \beta_{02} - 0.5 \times \phi_1^2 - 0.5 \times \phi_2^2 - \phi_1 - \phi_2; \quad \beta_{k3}^* = \phi_1 \times \beta_{k1} + \phi_2 \times \beta_{k2}, \text{ where } k = 1, 2, \dots, 5.$$

3.5 Application

The NESDA data introduced earlier (Penninx et al., 2008), were analysed using the proposed distance models. The sample comprised of $N = 2,938$ subjects aged 18 to 65 years (Mean=42; S.D.=13.1). About 66.5% were female and the average number of years of education attained was 12.2 with S.D. = 3.3. The responses of interest were diagnoses of dysthymia (DYST: 1 if diseased; 0, otherwise) and generalized anxiety disorder (GAD: 1 if diseased; 0, otherwise). About 10.2% and 15.3% of the subjects in the study developed DYST and GAD, respectively.

One of the objectives of NESDA is to measure the effect of personality traits on the

risk of developing mental disorders (Spinhoven et al., 2009). We considered the Big-Five personality variables, i.e., Neuroticism (N), Extraversion (E), Openness to experience (O), Agreeableness (A), and Conscientiousness (C). We also took into account the background variables, i.e., age (AGE), years of educations attained (EDU), and gender (GEN: 1=female; 0=male). Both the personality traits and the background variables will be treated as predictors.

In the final fitted (B)IPC models, all background variables and two of the personality traits such as neuroticism and extraversion, are retained since the other traits (such as O, A and C) are not statistically significant on both dimensions.

Table 3.4: Parameter estimates with corresponding standard errors (between the parenthesis) obtained from the IPC and BIPC models fitted on the NESDA data. IPC(2D-IND) corresponds to the 2-dimensional IPC model with *fixed* class coordinates; IPC(2D-FREE) to the 2-dimensional IPC model with *free* class coordinates; and IPC(3D) to the 3-dimensional IPC model.

Effect	Parameter	Models			
		IPC(2D-FIXED)	IPC(2D-FREE) [†]	IPC(3D)	BIPC [†]
Dysthymia					
Intercept	β_{01}	−2.20(0.131)	−2.57(0.148)	−2.55(0.167)	−2.21(0.131)
Gender	β_{11}	−0.18(0.140)	−0.21(0.143)	−0.25(0.180)	−0.17(0.139)
Age	β_{21}	0.20(0.072)*	0.20(0.073)*	0.18(0.093)*	0.20(0.072)*
Education	β_{31}	−0.15(0.066)*	−0.17(0.067)*	−0.18(0.085)*	−0.15(0.065)*
Neuroticism	β_{41}	1.03(0.102)*	1.14(0.127)*	1.13(0.133)*	1.03(0.102)*
Extraversion	β_{51}	−0.46(0.085)*	−0.47(0.087)*	−0.47(0.11)*	−0.45(0.085)*
Generalized Anxiety Disorder					
Intercept	β_{02}	−1.51(0.105)	−1.69(0.118)	−1.69(0.118)	−1.51(0.103)
Gender	β_{12}	−0.26(0.119)*	−0.31(0.136)*	−0.31(0.137)*	−0.26(0.117)*
Age	β_{22}	0.06(0.060)	0.03(0.068)	0.02(0.069)	0.05(0.059)
Education	β_{32}	−0.13(0.056)*	−0.14(0.064)*	−0.14(0.065)*	−0.12(0.055)*
Neuroticism	β_{42}	1.16(0.086)*	1.22(0.098)*	1.22(0.098)*	1.14(0.085)*
Extraversion	β_{52}	−0.15(0.070)*	−0.10(0.080)	−0.10(0.081)	−0.14(0.070)*
Association					
Intercept	β_{03}	—	1.75(0.199)	2.19(0.274)	1.69(0.207)
Gender	β_{13}	—	0.23(0.116)*	0.30(0.281)	0.16(0.081)*
Age	β_{23}	—	−0.02(0.055)	0.01(0.145)	−0.06(0.043)
Education	β_{33}	—	0.10(0.051)*	0.14(0.133)	0.09(0.034)*
Neuroticism	β_{43}	—	−0.92(0.187)*	−0.89(0.211)*	−0.73(0.170)*
Extraversion	β_{53}	—	0.08(0.072)	0.07(0.169)	0.16(0.067)*

[†] $\beta_{03} = \phi_1 \times \beta_{01} + \phi_2 \times \beta_{02} - 0.5 \times \phi_1^2 - 0.5 \times \phi_2^2 - \phi_1 - \phi_2$; $\beta_{k3} = \phi_1 \times \beta_{k1} + \phi_2 \times \beta_{k2}$, where $k = 1, 2, \dots, 5$.

* statistically significant, i.e., $p < 0.05$.

3.5.1 The IPC Models

The results of 2- and 3-dimensional IPC models fitted on the NESDA data are shown in Table 3.4.

The 2-dimensional IPC Model

The 2-dimensional IPC model with *fixed* class points, which is a marginal model with an *independence* association structure, is presented in the third column of Table 3.4 and has a fit statistic of $BIC = 3,784.1$ with twelve parameters.

We found a strong positive effect of neuroticism on risk of developing both mental disorders, i.e., $\hat{\beta}_{41} = 1.03$ with DYST; and $\hat{\beta}_{42} = 1.16$ with GAD. This implies that on average neurotic (i.e., emotionally unstable) people have a higher chance of developing the mental disorders. The other personality trait with stronger effect was extraversion with a moderate negative effect, i.e., $\hat{\beta}_{51} = -0.46$ with DYST; and $\hat{\beta}_{52} = -0.15$ with GAD. Being an introvert (i.e., having lower social engagement) seems to increase the chance of developing the mental disorders.

Among the background variables, education was the only predictor with statistically significant association with both disorders, i.e., $\hat{\beta}_{31} = -0.15$ with DYST; and $\hat{\beta}_{32} = -0.13$ with GAD. That is, less educated people had a higher chance of developing the disorders. The other vulnerable groups were males (i.e., $\hat{\beta}_{12} = -0.26$ with GAD) and elders ($\hat{\beta}_{21} = 0.20$ with DYST).

The fourth column shows the results of the 2-dimensional IPC model with *free* class points; its fit statistics was $BIC = 3,723.6$ with fourteen parameters. The additional two parameters are due to the estimated class points, i.e., $\hat{\phi}_1 = -0.01$ and $\hat{\phi}_2 = -0.74$. The association parameters presented in the last row block of Table 3.4 under IPC(2D-FREE), are not free parameters because they are estimated using the other parameters including the class coordinates as shown in equation (3.20). Gender, education, and neuroticism

had significant effect on the log-odds ratio, i.e., $\hat{\beta}_{13} = 0.23$, $\hat{\beta}_{33} = 0.10$ and $\hat{\beta}_{43} = -0.92$, respectively. Neuroticism had a negative strong effect on the log-odds ratio, which implies that the association between the two disorders became weaker when the level of neuroticism for a given person increased; and the rate of change was about 0.92 for a unit change in neuroticism. In the case of education, the direction was positive which implies that the association between the disorders became stronger when a person became more educated and the rate of change was about 0.10 for a unit change in education.

The results of IPC(2D-FIXED) and IPC(2D-FREE) models are not comparable as shown mathematically in Section 3.1. This is also evident if we compare the effect of extraversion under these models, i.e., $\hat{\beta}_{52} = -0.15$ under the IPC(2D-FIXED) model which is statistically significant, but it became insignificant under the IPC(2D-FREE) model, i.e., $\hat{\beta}_{52} = -0.10$.

The 3-dimensional IPC Model

The results of the 3-dimensional IPC model are presented in the fifth column that corresponds to IPC(3D) and its fit statistic was $BIC = 3,755.4$ with eighteen parameters. The first two row blocks of parameters under the IPC(3D) model have the same interpretation as the other models for the joint probabilities. Thus, we focus on the additional parameters that are displayed in the last row block, which corresponds to the association model as shown in equation (3.23).

It is important to note that these parameters are not comparable to those under the 2-dimensional IPC model, because the latter are specified in a lower-dimensional space and thus are restricted, while the former handles the association structure using separate parameters on third dimension. Only neuroticism had a significant effect on the log-odds ratio, i.e., $\hat{\beta}_{43} = -0.89$. This implies that the association between the two disorders became weaker when the level of neuroticism for a given person increased. The rate of change was about 0.89 for a unit change in neuroticism.

3.5.2 The BIPC Model

The last column of Table 3.4 shows the results from the BIPC model which had a fit statistic $BIC = 3,735.6$ with fourteen parameters. The first two row blocks display the marginal parameters. These results are equivalent to the IPC(2D-FIXED), and thus they both have the same interpretation.

The last row block shows the parameters of the association model that are obtained using the other parameters and the estimated class points, i.e., $\hat{\phi}_1 = -0.21$ and $\hat{\phi}_2 = -0.46$. Except age, all the predictors were statistically significant in the association model. The effect of extraversion was $\hat{\beta}_{53} = 0.16$, which implies that the association between the two disorders became stronger when the level of extraversion increased. In the case of neuroticism, the effect was negative, $\hat{\beta}_{43} = -0.73$. Thus the more neurotic a person was the weaker the association between the disorders.

The results of the BIPC model can also be displayed using a biplot (Gower & Hand, 1996; Gower et al., 2011). Figure 1 displays the biplot for the final BIPC model in which only the predictors having significant effect on both dimensions are considered. The labels of the predictors are placed at the positive side of the variable axis. On the variable axes markers are placed that represent $\mu_X \pm t\sigma_X$, where μ_X is the mean of X , σ_X is the standard deviation and $t = 0, 1, 2, 3$. From the biplot it is evident that neuroticism had a strong association with both mental disorders because its variable axis is long. The second influential predictor was extraversion pointing to the reverse direction compared to neuroticism.

The axes of the biplot corresponds to the marginal models, i.e., the horizontal axis corresponds to DYST and the vertical axis to GAD. The angle between a variable axis and each axis of the biplot, can be used to evaluate the strength of their association, i.e., the smaller the angle the stronger the association between them. For example, the angle between extraversion and DYST is smaller compared to the angle between extraversion

and GAD, which indicates that the association between extraversion and dysthymia is stronger. This result is in line with the estimates shown in the last column of Table 3.4 under extraversion, i.e., $\hat{\beta}_{51} = -0.45$ with DYST and $\hat{\beta}_{52} = -0.14$ with GAD.

The effect of predictors on the association model can also be read from the biplot. We showed mathematically in Section 3.1 that the IPC(2D-FIXED) is a marginal model with an *independence* association structure. This would correspond to the spatial solution in the biplot if the last category was positioned at $(\gamma_{41}, \gamma_{42}) = (1, 1)$. In the biplot displayed in Figure 1, however, the last category was positioned at $(0.79, 0.54)$ because $\hat{\phi}_1 = -0.21$ and $\hat{\phi}_2 = -0.46$. With every unit increase of neuroticism the log odds ratio of dysthymia and GAD changes by $\beta_{43} = \phi_1\beta_{41} + \phi_2\beta_{42}$. Both β_{41} and β_{42} were positive while ϕ_1 and ϕ_2 were negative. Therefore, with an increase of neuroticism the log odds ratio goes down. Along similar lines, we can derive that the log odds ratio increases with an increase of extraversion. These derivations show explicitly that the marginal model and the association structure are intuitively coupled, i.e., the same regression coefficients are used and only the ϕ -parameters can be used to adjust sign and strength. The adjustment by ϕ_1 and ϕ_2 is the same for every predictor variable.

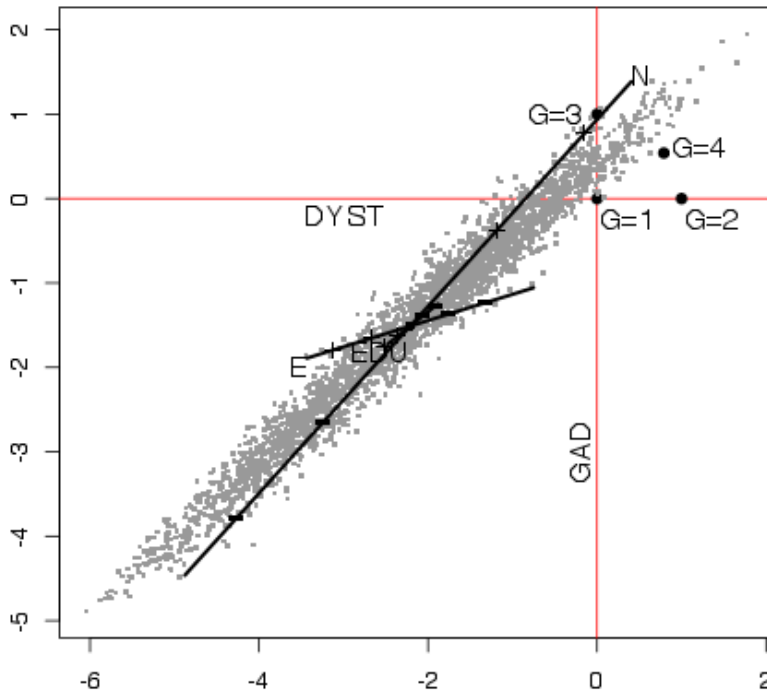


Figure 3.1: Biplot of the final BIPC model fitted on the NESDA data. The predictors neuroticism, represented by N ; extraversion, by E ; and education, by EDU . The bivariate binary responses are dysthymia, represented by $DYST$; and generalized anxiety disorder, by GAD . The class coordinates that correspond to the multinomial response variable, denoted by G , are also displayed.

3.6 Conclusion and Discussion

In this paper, we studied properties of the IPC model and extended it for analysing bivariate binary data in the presence of predictors, focusing on the marginal probabilities and the association structure. Researchers often model the marginal probability of an outcome variable without the influence of the other outcome variable. Such models are

referred as marginal models since the effect of the other outcome variable is marginalized. In addition to the marginal models, investigators are sometimes interested in modelling the association structure between the binary responses. It is expected that the two binary responses are correlated as they are measured on the same subject.

We found the following three results about the IPC model for analysing bivariate binary data. The 2-dimensional IPC model with *fixed* class point (IPC(2D-FIXED)) represents the marginal models with an *independence* association structure between the binary responses. Each dimension under the IPC(2D-FIXED) model pertains to one of the binary response variables. This result does agree with the finding by Liang and Zeger (1986) in which they showed that fitting a separate logistic regression model for each binary response variable gives consistent parameter estimates but biased standard errors. In the IPC model, however, the standard errors are not biased because estimation of model parameters are based on a multinomial likelihood function.

The 3-dimensional IPC model (IPC(3D)) represents the association structure in the third dimension. This model, however, misspecifies the models for the marginal probabilities. The compromise between the former two IPC models is a 2-dimensional IPC model with *free* class points (IPC(2D-FREE)). We showed, using simulation studies, that this latter model represents the association model as a form of restricted model. Like the IPC(3D) model, the IPC(2D-FREE) model misspecified the models for the marginal probabilities. Therefore, we conclude that the IPC model represents either the models for the marginal probabilities or the model for the association structure, but not both of them at the same time.

We therefore considered a possible extensions of the IPC model for representing both the marginal models and the association model at the same time. We modified the multinomial likelihood function following Bahadur (1961) and Lipsitz, Laird and Harrington (1990). The extended IPC model is called the Bivariate IPC (BIPC) model. Using simulation studies we showed that the BIPC model represented both the models for the

marginal probabilities and the model for the association structure well.

Unlike existing marginal models for bivariate binary data, the results of the BIPC model can be displayed graphically in a biplot which enhances the interpretation of the model. The axes in the biplot correspond to marginal models of the bivariate binary data, i.e., the horizontal axis corresponds to the first response variable and the vertical axis to the second response variable. The angle between the variable axis and each axis of the biplot is used to explain the strength of their association. In the same biplot, one can also read the relationship between a predictor variable and association structure (i.e., odds ratio). Therefore, we use both the ϕ -parameters and the marginal parameters to explain the direction and strength of their relationship. If both ϕ -parameters are found to be positive, it is an indication of a strong positive relationship between a predictor variable and the association structure. Similarly, an inverse relationship is characterized by the presence of negative estimates for both ϕ -parameters.

In this paper our focus was on application of the (B)IPC model for analysing bivariate binary data. Marginal modelling of multivariate polytomous type of responses has been an interest in social and other empirical sciences (Bergsma, 1997; Bergsma, Croon, & Hagenaars, 2009; Molenberghs & Lesaffre, 1994, 1999). The BIPC model can easily be extended for analysing bivariate polytomous responses by modifying the class coordinates to accomodate the additional response categories. At this stage, it is, however, not straight forward to extend the BIPC model for analysing multivariate binary responses. This is due to the fact that both the pairwise and higher-order association structure parameters must be specified in the likelihood function. With three binary responses (i.e., Y_1 , Y_2 , and Y_3), for example, three pairwise associations and a three-way association parameters must be specified which makes the computation cumbersome. If the interest is only on the pairwise association, the BIPC model for bivariate binary data can be extended by modifying the class point matrix.

We made the data and source codes (R / SAS) used in the simulation studies and in

the application available on the online repository system GitHub. The following link can be used to get access to the files: <https://github.com/workuhm1/BIPCM>.

Chapter 4

A Multivariate Logistic Distance Model for the Analysis of Multiple Binary Responses

Abstract

We propose a Multivariate Logistic Distance (MLD) model for the analysis of multiple binary responses in the presence of predictors. The MLD model can be used to simultaneously assess the dimensional/factorial structure of the data and to study the effect of the predictor variables on each of the response variables. To enhance interpretation, the results of the proposed model can be graphically represented in a biplot, showing predictor variable axes, the categories of the response variables and the subjects' positions. The interpretation of the biplot uses a distance rule. The MLD model belongs to the family of marginal models for multivariate responses, as opposed to latent variable models and conditionally specified models. By setting the distance between the two categories of every response variable to be equal, the MLD model becomes equivalent to a marginal model for multivariate binary data estimated using a GEE method. In that case the MLD model can be fitted using existing statistical packages with a GEE procedure, e.g., the *genmod* procedure from SAS or the *geepack* package from R. Without the equality constraint, the

This chapter was published as Worku, H. M. & De Rooij, M. (2018). A Multivariate Logistic Distance Model for the Analysis of Multiple Binary Responses. *Journal of Classification*, **35**, 1-23, <https://doi.org/10.1007/s00357-018-9251-4>. To address remarks of the PhD committee, this chapter is slightly modified.

MLD model is a general model which can be fitted by its own right. We applied the proposed model to empirical data to illustrate its advantages.

4.1 Introduction

Multivariate binary data with multiple binary response variables and one or more predictor variables, are often collected in empirical sciences such as psychology, criminology, epidemiology, life sciences and medicine. In the Netherlands Study of Depression and Anxiety (NESDA), for example, data were collected to investigate the interplay between personality traits and co-morbidity of depressive and anxiety disorders (Penninx et al., 2008; Spinhoven et al., 2009). Another study in which multivariate binary data arises is the Indonesian Children's Study (ICS: Sommer et al., 1984; Liang et al., 1992) where over three-thousand children were medically examined to investigate whether they had respiratory infection, diarrhoeal infection, and xerophthalmia. The aim of the ICS study was to investigate whether vitamin A deficiency places children at increased risk of respiratory and diarrhoeal infections.

The availability of the multivariate normal distribution for multivariate interval responses, makes application of maximum likelihood-based statistical models on such data relatively easy. However, for binary responses, no multivariate distribution is available and therefore estimation becomes more difficult. Liang and Zeger (1986) proposed Generalized Estimating Equations (GEE) for marginal modelling of correlated categorical data. GEE is a quasi-likelihood (QL) estimation method that does not require specification of a particular multivariate distribution. It is widely used as a standard approach for fitting marginal models on multivariate data (Ziegler et al., 1998; Fitzmaurice et al., 2008; Ziegler, 2011). The GEE approach, however, does not allow for a dimensional approach to analysis. Often researchers have theories how different response variables belong to one underlying dimension, factor, or latent variable.

For the dimensional approach often latent variable models are used, such as structural equation models or item response models. These models explicitly define underlying dimensions. However, these models make distributional assumptions of the latent di-

mensions or assume an underlying distribution for the dichotomous responses or both. Such assumptions are often unverifiable, i.e. we cannot check the assumptions using the data. In Appendix C we showed limitation of latent variable models regarding normality assumption of factor scores using empirical data.

In this paper we will develop a dimensional model for multivariate binary data within the marginal framework. The model will be developed within a distance framework, but we show it can also be seen as a specific marginal model. To enhance interpretation, a biplot is developed to accompany the model that visualises the result.

De Rooij (2009a) proposed the Ideal Point Classification (IPC) model for analyzing a multinomial response variable in the presence of predictors. The IPC is a probabilistic distance model based on unfolding distance function. (De Rooij & Heiser, 2005). De Rooij (2009a)) also showed that a simple logistic regression for binary response variable can be written as a unidimensional IPC model. Worku and De Rooij (2017, Chapter 3) extended the IPC model to the analysis of two binary response variables, i.e., the bivariate binary data setting, and showed that a new parameterization of the IPC model recovered both the marginal probabilities and the association structure of bivariate binary data well. However, this parameterization cannot be easily extended to handling multivariate binary data because all the possible pairwise and higher order association terms must be specified in the likelihood function, which makes the model complex and therefore hard to estimate.

Therefore, in this paper we propose a Multivariate Logistic Distance (MLD) model for analyzing multivariate binary data that extends marginal models for multivariate data. The MLD model unifies two domains of statistical methods, i.e., Multidimensional Scaling (MDS: Kruskal & Wish, 1978; Borg & Groenen, 2005) and Generalized Linear Model (GLM: McCullagh & Nelder, 1989; Agresti, 2013). As a form of regularization, the MLD model allows for dimension reduction and therefore less parameters are estimated compared to the existing marginal models for multivariate data. Moreover, the model enhances interpretation by using a biplot (Gabriel, 1971; Gower & Hand, 1996; Gower et

al., 2011) based on a distance interpretation.

Unlike existing marginal models for multivariate data, the MLD model can be used for assessing the factorial/dimensional structure of multivariate data. In the area of mental disorders (with the NESDA data as example) clinical psychologists and epidemiologists are often interested in co-morbidity and how co-morbidity is related to risk factors such as personality traits (Krueger, 1999; Beesdo-Baum et al., 2009; Spinhoven et al., 2013). Three candidate theories about the co-morbidity of mental disorders have been proposed, i.e., (1) a 2-dimensional structure with one dimension representing distress and the other one fear (d/f); (2) a different 2-dimensional structure with one dimension representing depression and the other one anxiety (d/a); and (3) an unidimensional structure where all the disorders are represented by a single dimension. The MLD model can be used to represent such theories within a unified framework, i.e., the candidate theories can be compared using appropriate statistics, and at the same time the MLD model allows for a direct relationship between co-morbidity of mental disorders and the predictor variables. It is assumed here that the covariates have the same effect on each of the responses that lie on the same dimension.

The paper is organized as follows. Section 2 develops the multivariate logistic distance model, investigates the link with marginal model for multivariate binary data estimated using a GEE method, and discusses the construction of biplots for the multivariate logistic model. In Section 3, the proposed model is fitted to empirical data and the results are interpreted using the estimated parameters and a graphical representation. We conclude in Section 4 with a discussion.

4.2 Multivariate Logistic Regression in a Distance Framework

4.2.1 Logistic Regression as a Distance Model

Logistic regression is a standard method for modelling dichotomous response data. Let y_i denote the observed value for a binary dependent variable Y for subject i , where $i = 1, 2, \dots, N$. Logistic regression models the probability of a category conditional on the value of a predictor variable x_i , $\Pr(y_i = 1|x_i) = \pi(x_i)$, i.e.,

$$\pi(x_i) = \frac{\exp(\beta_0^* + \beta_1^* x_i)}{1 + \exp(\beta_0^* + \beta_1^* x_i)}, \quad (4.1)$$

where β_0^* and β_1^* are the intercept and the regression coefficient, respectively. Logistic regression can easily be generalized to accommodate multiple predictors, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, and thus $\pi(\mathbf{x}_i) = \exp(\beta_0^* + \mathbf{x}_i^T \boldsymbol{\beta}^*) / (1 + \exp(\beta_0^* + \mathbf{x}_i^T \boldsymbol{\beta}^*))$, where $\boldsymbol{\beta}^*$ is now a vector with regression coefficients.

De Rooij (2009a) showed that logistic regression can be expressed as a distance model in a joint space with points representing the two categories of the response variable and points representing the subjects. In this section, we revisit this relationship and in Section 2.2 discuss an extension for multivariate binary responses.

Let us define a joint unidimensional space for subjects and the classes of the response variables. Denote by η_i the position of subject i and by γ_0 the coordinate of the position of one category and by γ_1 the coordinate of the position of the other category of the binary response variable. Define δ_{i0} and δ_{i1} to be the squared Euclidean distances between the

position of subject i and the two categories respectively. That is,

$$\begin{aligned}\delta_{i1} &= (\eta_i - \gamma_1)^2; \\ \delta_{i0} &= (\eta_i - \gamma_0)^2.\end{aligned}\tag{4.2}$$

With these two distances we can define the following probability model

$$\pi(x_i) = \frac{\exp(-0.5\delta_{i1})}{\exp(-0.5\delta_{i0}) + \exp(-0.5\delta_{i1})}.\tag{4.3}$$

The smaller the relative distance between a person point and a class point, the larger the probability that the subject belongs to that class. Therefore, the class probability is inversely related to the squared Euclidean distance between the points.

The coordinate for subject i , η_i , is assumed to be a linear combination of the predictor variable x_i , i.e., $\eta_i = \beta_0 + \beta_1 x_i$ or in case of multiple predictors $\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}$. The parameters of the distance model are the regression weights and the category points.

An important tool in the interpretation of probability models is the log-odds. The log-odds representation of the distance model becomes,

$$\begin{aligned}\log \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] &= 0.5\delta_{i0} - 0.5\delta_{i1} \\ &= \eta_i(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) \\ &= (\beta_0 + \beta_1 x_i)(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) \\ &= \beta_0(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) + \beta_1(\gamma_1 - \gamma_0)x_i.\end{aligned}\tag{4.4}$$

In the case of multiple predictors the logistic distance model takes the same form, having an intercept and extra slopes for the additional predictors. For example, with two

predictors $\mathbf{x}_i = (x_{i1}, x_{i2})^T$, the distance model becomes,

$$\begin{aligned} \log \left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] &= \beta_0(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) \\ &+ \beta_1(\gamma_1 - \gamma_0)x_{i1} + \beta_2(\gamma_1 - \gamma_0)x_{i2}. \end{aligned} \quad (4.5)$$

For a unit increase in x_{i1} the log-odds in the distance model changes by $\beta_1(\gamma_1 - \gamma_0)$, similarly for x_{i2} . By setting $\beta_0^* = \beta_0(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2)$ and $\beta_1^* = \beta_1(\gamma_1 - \gamma_0)$ a standard logistic regression is obtained.

The logistic distance model (4.4) is not identified and therefore identifiability constraint must be imposed. For example, with $\beta_1 = 2$ and $(\gamma_1 - \gamma_0) = 1$, $\beta_1^* = 2$. The same value $\beta_1^* = 2$ can also be obtained when $\beta_1 = 0.5$ and $(\gamma_1 - \gamma_0) = 2$. By imposing an identifiability constraint on the class points, the logistic distance model can be identified, for example by setting $\gamma_1 = 1$ and $\gamma_0 = 0$. The logistic distance model is now identified and its relationship with the univariate logistic model presented in (4.1) becomes

$$\begin{aligned} \beta_0^* &= \beta_0 - 0.5; \\ \beta_1^* &= \beta_1. \end{aligned} \quad (4.6)$$

4.2.2 Multivariate Extension of the Distance Model

In this section, the logistic distance model for a single response variable will be extended to handling multivariate binary data. Suppose $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{iJ})^T$ denotes the multivariate responses observed on the i -th subject, which is a $(J \times 1)$ -dimensional vector of all responses, where y_{ij} is the binary measurement of the j -th response variable observed on the i -th subject. It is not difficult to generalize the methodology to the case where the number of response variables differs over subjects, but that complicates the notation. As before, let $\mathbf{x}_i$ represent the multiple predictors observed on i -th subject. In Table 4.1, we display the structure of multivariate data in long format. The first column,

Table 4.1: The structure of multivariate data in long format.

SID	Index	Response	Predictor variables			
			x_1	x_2	...	x_p
1	R ₁	y_{11}	x_{11}	x_{12}	...	x_{1p}
1	R ₂	y_{12}	x_{11}	x_{12}	...	x_{1p}
1	R ₃	y_{13}	x_{11}	x_{12}	...	x_{1p}
1	R ₄	y_{14}	x_{11}	x_{12}	...	x_{1p}
1	R ₅	y_{15}	x_{11}	x_{12}	...	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	R ₁	y_{i1}	x_{i1}	x_{i2}	...	x_{ip}
i	R ₂	y_{i2}	x_{i1}	x_{i2}	...	x_{ip}
i	R ₃	y_{i3}	x_{i1}	x_{i2}	...	x_{ip}
i	R ₄	y_{i4}	x_{i1}	x_{i2}	...	x_{ip}
i	R ₅	y_{i5}	x_{i1}	x_{i2}	...	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	R ₁	y_{n1}	x_{n1}	x_{n2}	...	x_{np}
n	R ₂	y_{n2}	x_{n1}	x_{n2}	...	x_{np}
n	R ₃	y_{n3}	x_{n1}	x_{n2}	...	x_{np}
n	R ₄	y_{n4}	x_{n1}	x_{n2}	...	x_{np}
n	R ₅	y_{n5}	x_{n1}	x_{n2}	...	x_{np}

SID, is a variable which contains the subjects' identification number. The second column, Index, is a categorical indicator variable that indicates for which particular response variable the binary measurement y_{ij} is obtained. In Table 4.1 five response variables are assumed, i.e., $R_1, R_2, \dots, R_5$. The other columns represent measurements of the response variable and predictor variables, respectively.

A unidimensional space was used to represent the logistic regression model (4.3), which positions both the subjects and the two categories of the response variable. In the case of multiple responses $\mathbf{y}_i$, the distance model can be extended to accommodate the additional responses. Suppose there is a second response variable. One possibility for generalization is to add the two points representing the categories of the second response variable to the unidimensional space. In that case the predictor variables have a similar influence on the two response variables.

Another generalization is that the second response variable pertains to another dimen-

sion, giving rise to a two-dimensional model. The definition of the distance becomes

$$\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{kj}) = \sum_{m=1}^M (\eta_{im} - \gamma_{kj,m})^2,$$

where η_{im} is the coordinate for the point representing subject i on the m -th dimension and is defined as a linear combination of the predictors, $\eta_{im} = \beta_{0m} + \mathbf{x}_i^T \boldsymbol{\beta}_m$; and $\gamma_{kj,m}$ is the coordinate for category k ($k = \{0, 1\}$) of response variable j on dimension m . Each response variable belongs to one and only one dimension. This assumption is driven by theories often developed by applied scientists. In the Introduction section we discussed three different theories about comorbidity of mental disorders. Spinhoven et al. (2013), for example, found two dimensions of which the first dimension (distress) was represented by major depressive disorder, generalized anxiety disorder, and dysthymia; and the second dimension (fear) was represented by panic disorder and social phobia.

The probability for category 1 on response variable j given the predictors, i.e. $\Pr(y_{ij} = 1 | \mathbf{x}_i) = \pi_j(\mathbf{x}_i)$, is now defined by

$$\pi_j(\mathbf{x}_i) = \frac{\exp[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})]}{\exp[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{0j})] + \exp[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})]}. \quad (4.7)$$

The log-odds representation of the multivariate distance model becomes,

$$\log \left[\frac{\pi_j(\mathbf{x}_i)}{1 - \pi_j(\mathbf{x}_i)} \right] = \sum_{m=1}^M \left\{ \beta_{0m}(\gamma_{1j,m} - \gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2 - \gamma_{1j,m}^2) + \mathbf{x}_i^T \boldsymbol{\beta}_m(\gamma_{1j,m} - \gamma_{0j,m}) \right\}. \quad (4.8)$$

Because each response variable belongs to a single dimension, the log odds representation can be further simplified. Suppose response variable j belongs to dimension 1 so that $\gamma_{0j,m}$ and $\gamma_{1j,m}$ equal zero for all $m > 1$, i.e. all dimensions except the first one. In that case (4.8) simplifies to a single equation instead of a sum over dimensions. A hypothetical example is given in Appendix D to elaborate on the derivation of this simplified version.

This distance model for multivariate binary data (4.7 - 4.8) is called the Multivariate Logistic Distance (MLD) model. Because the MLD model is a type of bilinear model, for each dimension we have to fix the origin and scale. Like in the simple logistic regression representation we fix the class coordinates for one of the response variables on every dimension, e.g., $\gamma_{1j,m} = 1$ and $\gamma_{0j,m} = 0$.

The effect of a predictor variable on a specific response variable j is determined by the dimension the j -th response variable is positioned on. More specifically, the effect $\beta_m(\gamma_{1j,m} - \gamma_{0j,m})$. Therefore, for different response variables on the same dimension the size of the effect is different, depending on $(\gamma_{1j,m} - \gamma_{0j,m})$, but the direction is the same as long as $\gamma_{1j,m} \geq \gamma_{0j,m}, \forall j, m$, and defined by β_m . Furthermore, the larger $(\gamma_{1j,m} - \gamma_{0j,m})$ the bigger the effect is and vice versa. In other words, the larger the distance between the two points representing the categories of a single response variable, the better the predictor variables can discriminate between the categories.

4.2.3 Parameter Estimation

The parameters in the MLD model are estimated by maximizing the likelihood function for independent data, in the multivariate situation known as quasi-likelihood; i.e.,

$$L(\theta) = \prod_{i=1}^N \prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} [1 - \pi_j(\mathbf{x}_i)]^{(1-y_{ij})}, \quad (4.9)$$

where θ is the concatenation of all the class points and all the regression weights. The quasi-likelihood (4.9) is an approximation to a full likelihood function for multivariate binary data as there is no general parsimonious parameterization of the multivariate binary distribution.

Liang and Zeger (1986) showed that maximizing this quasi-likelihood provides consistent parameter estimates for the multivariate model. However, the standard errors based on the corresponding Hessian matrix are biased. The same authors proposed a sand-

wich estimator for the covariance matrix to correct for the bias (Liang & Zeger, 1986). Another method for obtaining correct standard errors is to apply a clustered bootstrap method (Sherman & Le Cessie, 1997; De Rooij & Worku, 2012; Cheng, Yu, & Huang, 2013). In this case, the re-sampling procedure is applied on the subject (cluster) level so that the correlation between the multivariate responses is retained in each bootstrap sample.

The number of independent parameters estimated in the MLD model, q , equals

$$q = [M \times (p + 1)] + [(J - M) \times 2]. \quad (4.10)$$

The first term in (4.10), i.e., $[M \times (p + 1)]$, corresponds to the number of regression coefficients; the other term to the number of estimable coordinates of class points. The identifiability constraints are accounted for in the second term, i.e., in each dimension the class coordinates for a single response variable are set to fixed values.

The MLD model can be fitted using the NLMIXED procedure in SAS software (SAS Institute Inc., 2011). Scripts for the analyses in this paper are available upon request from the first author.

4.2.4 The Relationship of the MLD Model to a Marginal Logistic Regression model

By setting the distance between the two categories of every response variable to be equal to one, i.e., $(\gamma_{1j,m} - \gamma_{0j,m}) = 1$, the MLD model becomes equivalent to a marginal model for multivariate binary data estimated using GEE method (Liang & Zeger, 1986). The restriction of the class points implies that predictor variables discriminate equally well for all response variables belonging to a specific dimension. Existing statistical packages with a GEE procedure (e.g., the genmod procedure from SAS or the geepack package from R) can be used for fitting this “restricted” MLD model on multivariate binary data.

Fitting the restricted MLD model using a GEE procedure involves a three-step approach: (1) construction of a design matrix for both the response and the predictor variables; and (2) applying the GEE method with the constructed design matrix; and (3) transforming the GEE parameters to MLD parameters. We now show construction of the design matrix using the example presented in Table 4.1.

Suppose we want to fit a 2-dimensional model on the five binary response variables. Further, suppose we like the first three response variables to be represented on the first dimension, and the fourth and the fifth on the second dimension. Therefore define a response indicator matrix, denoted by $\mathbf{Z}$, with dimension $(J \times M)$. The response indicator matrix specifies for each response variable to which dimension it pertains, with position (j, m) equal to one if the j -th response variable belongs to the m -th dimension and zero otherwise. For the structure layed-out above,

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. \quad (4.11)$$

The design matrix for subject i is then obtained by computing the Kronecker product between the response indicator matrix and the predictors vector (without intercept), $\mathbf{U}_i = \mathbf{Z} \otimes \mathbf{x}_i^T$, such that

$$\mathbf{U}_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}. \quad (4.12)$$

We concatenate $\mathbf{U}_i$ and the identity matrix to get the final design matrix, $\mathbf{S}_i = [\mathbf{I}_i, \mathbf{U}_i]$,

$$\mathbf{S}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}.$$

Then, a vertical concatenation of all $\mathbf{S}_i$ matrices will give us the final design matrix $\mathbf{S}$ on which the GEE method is finally applied to obtain parameter estimates of the marginal model. This results in five response specific intercepts $(\beta_{01}^*, \dots, \beta_{05}^*)$ corresponding to the first five columns of $\mathbf{S}$ and two sets of p regression weights $(\beta_{11}^*, \dots, \beta_{p1}^*$ and $\beta_{12}^*, \dots, \beta_{p2}^*)$. The MLD parameters can be derived from these as follows $\gamma_{0j,m} = -(\beta_{0j}^* + 0.5)$ for the dimension, m , to which disorder j belongs, zero otherwise. The regression weights β_{jm} are equal to the regression weights obtained from GEE method, $\beta_{jm} = \beta_{jm}^*$. The number of parameters in the “restricted” MLD model then becomes $q = [M \times (p + 1)] + (J - M)$ since additional constraints are imposed on the class points.

4.2.5 Model Selection

In statistical analysis we often select a parsimonious and best fitting model from a set of candidate models given the data. In the MLD model, we select not only predictor variables for the final model, but also the dimensionality of the model must be determined.

Pan (2001) proposed the quasi-likelihood under the independence model criterion (QIC) as an extension of Akaike Information Criterion (AIC) to GEE:

$$\text{QIC} = -2L(\boldsymbol{\theta}) + 2 \text{ trace}(\hat{\boldsymbol{\Omega}}_I^{-1} \hat{\mathbf{V}}_R), \quad (4.13)$$

where $L(\boldsymbol{\theta})$ is as defined in 4.9 above; $\hat{\mathbf{V}}_R$ stands for robust variance estimator obtained

under the assumption of a general “working” covariance structure R ; and, $\hat{\Omega}$ is for the naive variance estimator obtained under the assumption of an independence correlation structure. Pan (2001) also proposed a simplified version of QIC when $\text{trace}(\hat{\Omega}_I^{-1}\hat{\mathbf{V}}_R) \approx \text{trace}(\mathbf{I}) = q$, i.e.,

$$\text{QIC}_u = -2L(\theta) + 2q.$$

Yu and De Rooij (2013) studied the performance of QIC_u for determining the dimensionality of the Trend Vector Model (TVM). Both the Trend Vector model and the MLD model are marginal models in a distance framework, where the first is used for longitudinal multinomial response variables and the latter for multivariate binary responses. Yu and De Rooij (2013) recommended QIC_u for determining the dimensionality of the distance model.

In the MLD model, we use QIC_u fit statistics both for determining the dimensionality of the model and for variable selection. The model with the lowest QIC_u statistics is considered the most parsimonious and best fitting model. As recommended in Yu and De Rooij (2013), we first determine the dimensionality of the model and then proceed to the variable selection.

QIC (4.13) is a general formula for model selection and is used if there is also an interest to select the working correlation matrix, R . (Pan, 2001) In our case, we use QIC_u statistics as we are only interested on dimensionality and variable selection.

4.2.6 Biplot for the Multivariate Logistic Distance Model

To enhance interpretation of the model the results of a MLD model can be graphically represented in a biplot (Gabriel, 1971; Gower & Hand, 1996; Gower et al., 2011). The biplot represents the subjects, the response variables, and the predictor variables so that the relationship between predictors and responses can be read from the graph. We first

demonstrate how the response variables are included in the biplot, and then the predictors.

For a 2-dimensional MLD model the coordinates for a response variable are given by

$$\gamma_j = \begin{bmatrix} \gamma_{0j,1} & \gamma_{0j,2} \\ \gamma_{1j,1} & \gamma_{1j,2} \end{bmatrix}.$$

Because each response is positioned on one and only one dimension, one of the columns in γ_j equals zero. So, γ_j represents two points either on the first or second dimension. Halfway between the two points, a *decision line* is drawn indicating equal probabilities for the two categories of a response variable. Due to these lines (horizontal for response variables on the second dimension and vertical for response variables on the first dimension), the two dimensional space is partitioned into rectangles, each representing a most probable response profile.

The predictors are included in the biplot by variable axes (Gower & Hand, 1996). To derive the variable axis, first, a pseudo-design matrix $\tilde{\mathbf{X}}$ is constructed containing ones in the first column and zeros in the other columns except for the column representing the variable to be plotted. In this column, marker values are included within the range of the observed variable. Second, the matrix $\mathbf{B}$ with regression weights is defined, i.e.

$$\mathbf{B} = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_1 & \beta_2 \end{bmatrix}.$$

Finally we can compute the matrix $\mathbf{H}$ as

$$\mathbf{H} = \tilde{\mathbf{X}}\mathbf{B},$$

defining a straight line in our biplot. We will include variable axes for every statistically significant predictor. Positions of the subjects are computed as the linear combination of predictor variables and are included in the biplot as points.

4.3 Application: The NESDA Data

In order to illustrate the MLD model, the NESDA data (Penninx et al., 2008) introduced earlier were analysed. The sample comprised of $N = 2,938$ subjects aged 18 to 65 years (Mean = 42; S.D. = 13.1). About 66.5% were female and the average number of years of education attained was 12.2 with S.D. = 3.3. In this study, 37.1% of the subjects have major depressive disorder (MDD), 10.2% have dysthymia (DYST), 15.3% have generalized anxiety disorder (GAD), 22.4% have social anxiety disorder (SP), and 28.6% have panic disorder (PD). These five disorders are the response variables.

The predictors are Neuroticism (N), Extraversion (E), Openness to experience (O), Agreeableness (A), and Conscientiousness (C). We also took into account three background variables, i.e., age (AGE), years of education attained (EDU), and gender (GEN: 1 = female; 0 = male). The linear predictor part of the MLD model is

$$\begin{aligned}\eta_{im} = & \beta_{0m} + \beta_{1m}(\text{AGE})_i + \beta_{2m}(\text{EDU})_i + \beta_{3m}(\text{GEN})_i \\ & + \beta_{4m}\mathbf{N}_i + \beta_{5m}\mathbf{E}_i + \beta_{6m}\mathbf{O}_i + \beta_{7m}\mathbf{A}_i + \beta_{8m}\mathbf{C}_i,\end{aligned}$$

where η_{im} is a coordinate for the i -th subject position on the m -th dimension; and the β 's are regression weights. The candidate MLD models fitted on the NESDA data are

- (1) “distress-fear” (d/f) dimensions, in which MDD, GAD, are DYST are presumed to be indicators of distress, and PD and SP for fear;
- (2) “depression-anxiety” (d/a) dimensions, in which MDD and DYST are indicators of depression, and GAD, PD, and SP for anxiety;
- (3) “unidimensional” where all the five mental disorders are indicators of a single dimension.

Table 4.2: Results of fitting different MLD models to NESDA data. In the first block, dimensionality of the MLD model is assessed, and followed by variable selection in the second block.

Model	Dimension	Predictors	q	QIC_u
Model Selection for Dimensionality				
1	2 (d/f) [†]	All	21	12, 396.42
2	2 (d/a) [‡]	All	21	12, 398.08
3	1	All	13	12, 418.87
Model Selection for Predictors				
1a	2 (d/f)	All	21	12396.42
1b	2 (d/f)	AGE,EDU,GEN,N,E	15	12396.68
1c	2 (d/f)	AGE,EDU,GEN	11	14789.41

[†] d/f: distress/fear model.

[‡] d/a: depression/anxiety model.

These three theories are then translated into the following indicator matrices:

$$\mathbf{Z}^{(1)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Z}^{(2)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Z}^{(3)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad (4.14)$$

respectively. The superscript corresponds to a theory.

For illustration, we fitted both the MLD model with and without imposing equal distance restrictions on the class points. The results of the MLD model with the restrictions will be presented first, thereafter the solution without the restrictions will be discussed.

Table 4.2 shows the fit statistics of the candidate MLD models. As shown in the first block of Table 4.2 which compares different dimensionality, the 2-dimensional distress-fear (d/f) model fitted the data best ($QIC_u = 12, 396.42$). In the second block of Table 4.2, fit statistics for the comparison of different sets of predictor variables are given. The model with all predictor variables fitted the data best ($QIC_u = 12, 396.42$).

The regression weights of this selected model are given in Table 4.3. The standard

errors based on both the sandwich and the clustered bootstrap method are included in Table 4.3. Both methods resulted in similar estimates.

There is a strong positive association between neuroticism and the two dimensions: $\hat{\beta}_{41} = 0.1167$ with distress; and $\hat{\beta}_{42} = 0.1039$ with fear. With every unit increase in neuroticism the log odds for MDD, DYST, and GAD go up by 0.1167 while the log odds for SP and PD go up by 0.1039. There is a moderate negative association between extraversion and the two dimensions: $\hat{\beta}_{51} = -0.0419$ with distress; and $\hat{\beta}_{52} = -0.0320$ with fear. With every unit increase in extraversion the log odds for MDD, DYST, and GAD go down by 0.0419 while the log odds for SP and PD go down by 0.0320. From the background variables, only education has a statistically significant effect on both dimensions: $\hat{\beta}_{11} = -0.0386$ with distress; and $\hat{\beta}_{12} = -0.0575$ with fear. Less educated people have a higher risk of getting a mental disorder. The variable conscientiousness had a significant effect only on the second dimension (distress), $\hat{\beta}_{82} = 0.0189$, i.e. it only influences PD and SP.

Although the total number of parameters of the final d/f model is $q = 21$, only sixteen of the parameters were displayed in Table 4.3. The others are the intercepts obtained from GEE method which are response-specific, i.e., $\beta_{01}^{\text{MDD}} = -2.23$, $\beta_{02}^{\text{GAD}} = -3.73$, $\beta_{03}^{\text{DYST}} = -4.28$, $\beta_{04}^{\text{PD}} = -3.74$, and $\beta_{05}^{\text{SP}} = -4.14$. Using $\gamma_{0j,m} = -(\beta_{0j}^* + 0.5)$ as shown in Section 2.4 and $\gamma_{1j,m} = 1 + \gamma_{0j,m}$, the class point coordinates for each response variable can be obtained. Thus, $\gamma_{01,1} = 1.73$ for MDD, $\gamma_{02,1} = 3.23$ for GAD, $\gamma_{03,1} = 3.78$ for DYST, $\gamma_{04,2} = 3.24$ for PD, and $\gamma_{05,2} = 3.64$ for SP. We can use the estimated class points to compare the effect of predictors on the risk of developing disorders belonging to the same dimension. For example, MDD, DYST and GAD belong to the first dimension. Because $\gamma_{03,1} = 3.78$ for DYST is larger than both $\gamma_{01,1} = 1.73$ for MDD and $\gamma_{02,1} = 3.23$ for GAD, it means that starting from a very low subject position on the first dimension and then increasing this position will first lead to higher probabilities of MDD, followed by GAD, and then for DYST. The model accounts for comorbidity because a high probability

Table 4.3: Summarized results of the final “distress-fear” MLD model fitted on NESDA data. Restriction was applied on the class points, and thus it is a restricted MLD model. The reported standard errors are based on both sandwich and clustered bootstrap methods. The number of bootstraps, $B = 1000$.

Effect	Parameter	Estimate	SE (sandwich)	Bootstrap	
				SE	Wald
Distress dimension					
Education [†]	β_{11}	−0.0386	0.012	0.012	10.06
Gender	β_{21}	−0.1346	0.081	0.081	2.79
Age	β_{31}	0.0012	0.003	0.003	0.15
Neuroticism [†]	β_{41}	0.1167	0.006	0.006	413.84
Extraversion [†]	β_{51}	−0.0419	0.007	0.007	39.43
Openness to Experience	β_{61}	−0.0031	0.007	0.008	0.17
Agreeableness	β_{71}	−0.0074	0.008	0.007	1.03
Conscientiousness	β_{81}	−0.0071	0.007	0.007	1.06
Fear dimension					
Education [†]	β_{12}	−0.0575	0.012	0.011	26.18
Gender	β_{22}	0.0229	0.082	0.083	0.08
Age	β_{32}	−0.0008	0.003	0.003	0.08
Neuroticism [†]	β_{42}	0.1039	0.006	0.006	335.26
Extraversion [†]	β_{52}	−0.0320	0.007	0.006	25.56
Openness to Experience	β_{62}	0.0008	0.008	0.008	0.01
Agreeableness	β_{72}	−0.0003	0.008	0.008	0.00
Conscientiousness [†]	β_{82}	0.0189	0.007	0.007	6.72

[†] statistically significant effect, $p < 0.05$.

of DYST implies a high probability of GAD and MDD.

The results of the selected MLD model are displayed in a biplot shown in Figure 4.1. In order to interpret the biplot, let us first discuss how the biplot was constructed. The biplot is composed of two parts, i.e., the response space and the variable axes, as shown in Figure 4.2 and 4.3, respectively. The positions of the two categories of all response variables are displayed in Figure 4.2. For example, on the vertical dimension there are four points corresponding to no PD, no SP, having PD, and having SP from the bottom to the top, respectively. Included in the same representation are *decision lines* (vertical and horizontal lines) crossing the mid-points between the two categories. The decision lines partition the two-dimensional space into rectangles (regions), each representing a most probable response profile.

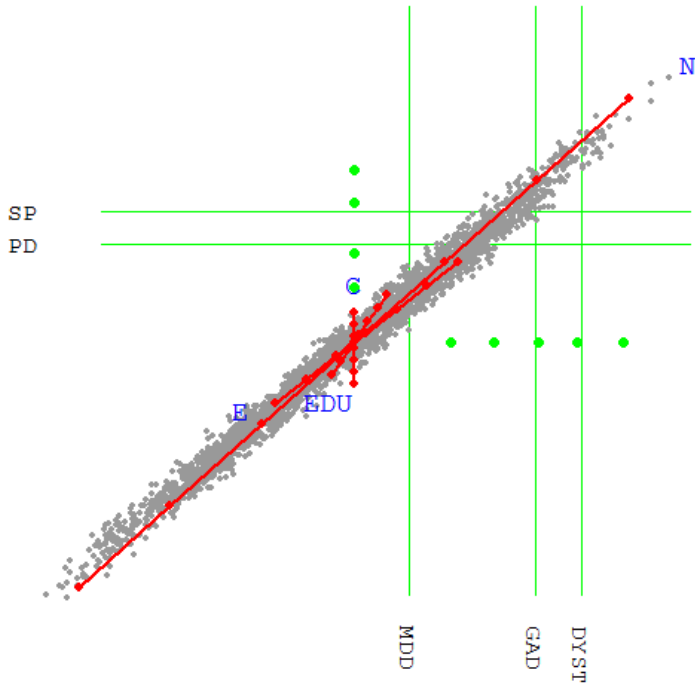


Figure 4.1: Biplot of the final “distress-fear” model fitted on the NESDA data, where the first dimension is represented by three disorders (MDD, GAD and DYST) and the second dimension by two disorders (SP and PD). The plot is based on restrictions applied on the class points.

Each region shows the disorder profile by 1’s and 0’s for the order MDD, GAD, DYST, PD, SP. An index ‘10011’, for example, corresponds to the presence of MDD, PD, and SP, but the absence of GAD and DYST. In the top-right, an index of ‘11111’ is used to represent a co-morbidity of all five mental disorders, while the region ‘00000’ in the bottom left representing the absence of disorders.

In Figure 4.3, both the variable axes (lines) and the subjects points (grey dots) are displayed. The space includes only statistically significant predictors. On the variable axes markers are placed that represent $\mu_x \pm t\sigma_x$, where μ_x is the mean of x , σ_x is the standard deviation, and $t = 0, 1, 2, 3$. Variable labels are included at the positive side of

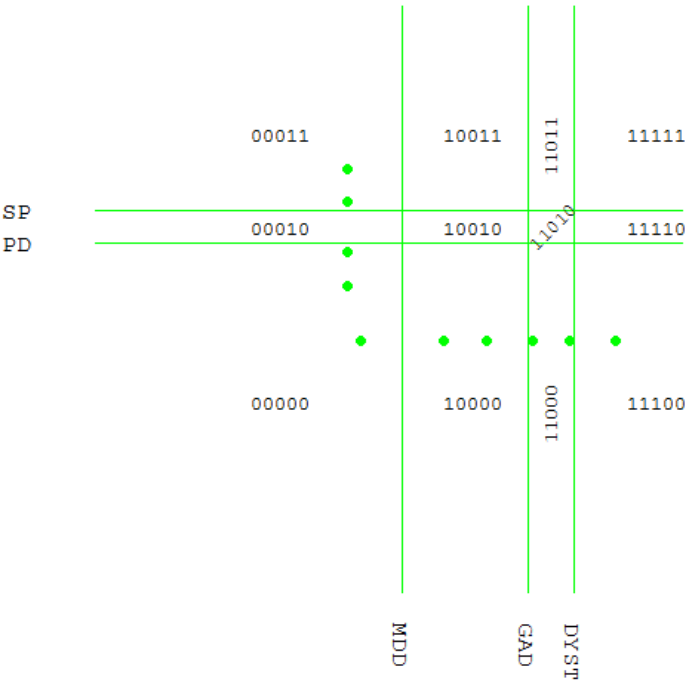


Figure 4.2: Representation of the binary response variables in the Euclidean space.

the variable axis.

Let us now interpret the biplot displayed in Figure 4.1. Most of the subjects are in the bottom left region representing absence of all the disorders. However, significant number of subjects are in other regions representing co-morbidity of mental disorders. The regions are '10000' corresponding to the presence of having only MDD; and '10010' corresponding to the presence of having both MDD and PD; '10011' corresponding to the presence of having MDD, PD, and SP; and, '11011' corresponding to the presence of all disorders, except DYST. Also a few subjects are in the upper right region having all the mental disorders.

Now let us interpret the variable axes. The variable axis for Neuroticism (N) runs from the lower left (low values of neuroticism) to the upper right (high values of Neuroticism),

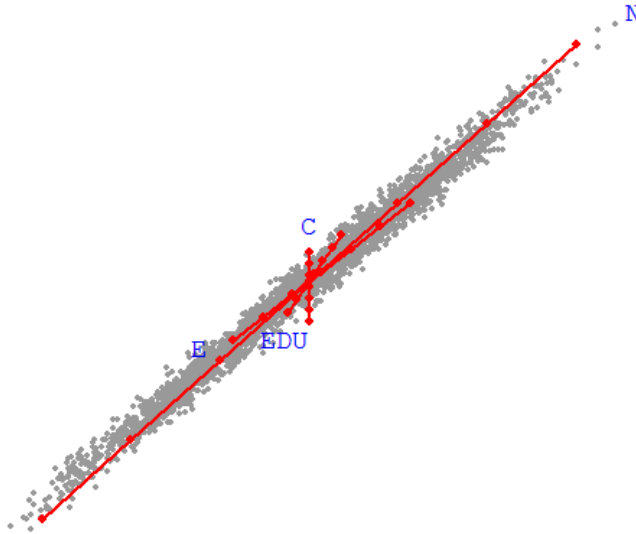


Figure 4.3: Variable axes representation of the predictor variables (i.e., N: Neuroticism, E: Extraversion, C: Conscientiousness, and EDU: EDUcation) in the Euclidean space.

indicating that persons with low values of Neuroticism are located in the '00000' region, whereas persons with very high values of neuroticism are located in the '11111' region. In short, the higher neuroticism the more disorders. Contrarily, the variable axes of extraversion points to the other direction.

The length of the variable axis indicates effect size; the longer the variable axis the larger the effect of the corresponding variable on the disorders.

The angle between the variable axis and the dimension measures the strength of their association. The smaller the angle between them, the stronger the association. For example, the angle of the extraversion variable axis with the first (horizontal) dimension is relatively small compared to the angle of extraversion with the second dimension. This indicates that extraversion has a larger effect on the disorders represented on the first dimension (MDD, DYST, and GAD) than on the disorders presented on the second di-

mension (PD and SP). The angle of neuroticism with both dimensions is about equal, although a bit smaller with the first dimension, indicating that the relationship of neuroticism with the disorders on the first dimension (MDD, GAD, and DYST) is slightly stronger than with the other two disorders. The variable conscientiousness is highly correlated to the second dimension and not to the first as its variable axis is orthogonal to the first dimension.

Finally, there is a strong correlation between the estimates of the subject points in the two dimensions, $\text{corr}(\hat{\eta}_{i1}, \hat{\eta}_{i2}) = 0.99$, indicating that the distress and fear dimensions are strongly correlated.

We now present the results of MLD model that does not impose restriction on the class points, i.e., the “unrestricted” MLD model, to address specifically the extra information from this model. The regression estimates are shown in Table 4.4. The estimates obtained from the “unrestricted” MLD model are slightly different compared to results obtained from the “restricted” MLD model fitted on NESDA data (shown in Table 4.3). However, both results lead to the same conclusion about significance of predictors, which is also indicated by the “Wald” statistics displayed in the last column of both tables. The class points for MDD are fixed for identification on the first dimension, i.e. the coordinates are 0 for no MDD and 1 for MDD. Similarly, the coordinates of PD on the second dimension are fixed to 0 for absence and 1 for presence of the disorder. The other parameters are the class points, i.e., $\gamma_{02,1} = 0.96$ and $\gamma_{12,1} = 1.73$ for GAD; $\gamma_{03,1} = 1.10$ and $\gamma_{13,1} = 1.99$ for DYST; and, $\gamma_{05,2} = -0.25$ and $\gamma_{15,2} = 1.28$ for SP. The distance between the two category points is 0.77 for GAD, 0.89 for DYST, and 1.53 for SP.

This unrestricted MLD model provides additional information about how well the predictors can discriminate between the response categories. According to equation (4.8), the effect of the predictor variables on each response is partially determined by the distance between class points of the response variable. The larger the distance between the class points of a response variable, the better the predictor variables are able to discriminate

Table 4.4: Regression weights of the final unrestricted “distress-fear” MLD model fitted on NESDA data. The number of bootstraps used to obtain standard errors equals 1000.

Effect	Parameter	Estimate	Bootstrap	
			SE	Wald
Distress dimension				
Education [†]	β_{11}	−0.0203	0.006	11.45
Gender	β_{21}	−0.0685	0.042	2.66
Age	β_{31}	0.0004	0.001	0.16
Neuroticism [†]	β_{41}	0.0605	0.004	228.77
Extraversion [†]	β_{51}	−0.0226	0.004	31.92
Openness to Experience	β_{61}	−0.0020	0.004	0.25
Agreeableness	β_{71}	−0.0037	0.004	0.86
Conscientiousness	β_{81}	−0.0041	0.004	1.05
Fear dimension				
Education [†]	β_{12}	−0.0202	0.005	16.32
Gender	β_{22}	0.0005	0.033	0.00
Age	β_{32}	−0.0007	0.001	0.49
Neuroticism [†]	β_{42}	0.0424	0.003	199.75
Extraversion [†]	β_{52}	−0.0141	0.003	22.09
Openness to Experience	β_{62}	0.0000	0.003	0.00
Agreeableness	β_{72}	0.0003	0.003	0.01
Conscientiousness [†]	β_{82}	0.0067	0.003	4.99

[†] statistically significant effect, $p < 0.05$.

between the categories. In the fitted model both DYST and GAD are positioned on the first dimension; because the distance for DYST (0.89) is larger than the distance for GAD (0.77), the effect of the predictor variables on DYST is stronger.

4.4 Conclusion and Discussion

We proposed a multivariate logistic distance (MLD) model for analyzing multivariate binary data that extends existing marginal models in a distance framework. The distance model for a single response variable was extended to analyzing multivariate binary data in the presence of predictors. The advantage of the MLD model over existing marginal model for multivariate data, is the possibility for dimension reduction as a form of regularization which simplifies the complexity of standard multivariate GLM model because less parameters are estimated. Moreover, using this dimension reduction substantial theories can be represented and investigated.

We have shown applications of both the “restricted” and the “unrestricted” MLD models using an empirical data set. The former MLD model imposes a restriction on the class points and the latter model does not. The “restricted” MLD model is equivalent to a marginal model for multivariate binary data estimated using GEE method, which is an advantage for our model because existing software package developed for GEE can be adopted to fit the MLD model. For the unrestricted case, the MLD model is a general model and can be fitted by its own right. The general MLD model provides us with additional information about how well the predictors can discriminate between the categories of the response variable.

The MLD model has a clear interpretation where both the odds ratio expressions as well as the biplot representation can be used. The space in the biplot is partitioned into different regions that indicate the most probable response profile. It is important to note that the assumption of which response variables belong to which dimension has a crucial

impact on which regions might occur. In a unidimensional model there are maximal 6 regions, whereas in the two dimensional solution in Figure 2 there are 12 regions. Having 5 response variables, a total of 32 different profiles can be defined. In a five dimensional model all these 32 profiles are present. Dimension reduction thus reduces the number of most probable response profiles. Moreover, the regions also account for comorbidity. In the solution of Figure 2 there is never a response profile where MDD is absent and DYST and GAD are present. Similarly, if PD is present then also SP is present in the response profile. Notice, however, that the model is a probability model not a deterministic model. So, a response profile is most probable but the model does not say that in that region only a profile must occur.

The effect size of predictor variables can be read from the biplot by the length of the variable axis. The longer the variable axis the stronger the effect. The differential effect on the two dimensions can be read from the angle of a variable axis with the dimension. The smaller the angle the stronger the effect. If a variable has a 90° angle with a dimension, the variable has no effect on the disorders belonging to that dimension.

The MLD model is related to Canonical Correspondence Analysis (CCA), as proposed by Ter Braak (1986), which is a multivariate method used for ordination axes that maximizes the separation among the multivariate binary responses (Ter Braak, 1986; Ter Braak & Verdonschot, 1995). There are two main differences between CCA and our model. The first is that these models work in different framework, i.e., the MLD model in a logistic framework where as CCA in a Gaussian framework. Due to this difference, the MLD can provide a clear interpretation in terms of (log)-odds and probabilities. The second is that unlike in CCA, the MLD model can position responses (e.g., mental disorders) on certain dimensions driven by the theories that we would like to test.

In areas like psychology, epidemiology, criminology, economics, political sciences, etc, researchers often use Structural Equation Models (SEM) for the analysis of data similar to the NESDA data (Plewis, 1996; Von Oertzen, Hertzog, Lindenberg, & Ghisletta,

2010; Spinhoven et al., 2013). Despite its popularity, SEM has limitations as it makes unverifiable assumptions about the underlying distributions of latent as well as observed variables (e.g., normality assumption for the latent variables). Moreover, SEM often suffers from improper solutions, non-convergent solutions, and the predicted factors are not determinate, i.e., for the same number of response variables multiple solutions can be obtained for the underlying latent variables. Therefore, they cannot be uniquely identified (Acito & Anderson, 1986; Boomsma & Hoogland, 2001; Hubbard et al., 2010). In the application section, we showed that the MLD model can be used for comparing theories of interest, without making unverifiable assumptions about underlying distributions.

Asar and Ilk (2013) proposed marginal model with shared-parameter within the GEE method (Asar & Ilk, 2013). To compare with our MLD model, they use the five dimensional model where each response variable pertains to a unique dimension. Then, they incorporate equality restrictions for certain predictors over different dimensions, giving a so-called shared parameter. In the restricted MLD model the regression weights are shared for all response variables belonging to a specific dimension.

Although our focus was on binary data, the model can be extended to polytomous data. Where for binary data there are two class points on each dimension for polytomous data there are multiple class points. Interpretation follows largely the binary model, although in the ordinal case we can derive odds ratios for every contrast of two categories of a response variable. These are formed by the difference of class points coordinates, just like in the binary case. The polytomous situation, however, is often more complicated than the binary one. The binary model for every response variable is by definition unidimensional, which is not the case for polytomous data. Therefore, the polytomous case needs further study.

Regarding model assumptions, it is worth mentioning the following two points. The first point is that the MLD model makes a strong linearity assumption regarding the explanatory variables, i.e., the model assumes that the explanatory variables are linearly

related to logit transform of the class probabilities. However, this assumption could be solved for example by using polynomial functions of the original explanatory variables. The second point is that, compared to structural equation models, the MLD model does not have the assumption of a normal distribution for the latent variables anymore.

We developed a package (an alpha version) in R, the `mldm` package, for fitting the MLD model on multivariate binary data in the presence of predictors. The package handles both the clustered bootstrap method and the sandwich estimators for correcting standard errors of model parameters. The package provides a biplot function for the fitted model. We also have SAS scripts for fitting the models. The first author can provide the package or the script upon request.

Chapter 5

mldm: An R Package for Analyzing Multivariate Binary Data

Abstract

We developed the **mldm** package in R (R Development Core Team, 2008) to fit a multivariate logistic distance model on multiple binary responses in the presence of explanatory variables. The package handles both the clustered bootstrap method and the sandwich estimators for obtaining the standard errors of model parameters. The package provides a `biplot` function to display results of the fitted model. In this chapter we illustrate the usage of the package using an empirical data.

This chapter is a user manual for the **mldm** package in R software developed by Worku, H. M. & De Rooij, M. (2018) for analyzing multivariate binary data.

5.1 Introduction

The Multivariate Logistic Distance (MLD) model is proposed to analyze multiple binary responses in the presence of explanatory variables (Worku & De Rooij, 2018). The MLD model unifies two domains of statistical methods, i.e., Multidimensional Scaling (MDS: Kruskal & Wish, 1978; Borg & Groenen, 2005) and the Generalized Linear Model (GLM: McCullagh & Nelder, 1989; Agresti, 2002). As a form of regularization, the MLD model allows for dimension reduction and therefore less parameters are estimated compared to the existing marginal models for multivariate data. Moreover, the model enhances interpretation by using a biplot (Gabriel, 1971; Gower & Hand, 1996; Gower et al., 2011) based on a distance interpretation.

For fitting the MLD model, we developed the **mldm** package in R statistical software. In this chapter we illustrate the usage of the package with an empirical data.

5.2 The Multivariate Logistic Distance Model

Suppose $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{ij}, \dots, y_{iJ})^T$ denotes the multivariate responses observed on the i -th subject, which is a $(J \times 1)$ -dimensional vector of all responses. The y_{ij} represents a binary measurement of the j -th response variable observed on the i -th subject.

The MLD model defines the probability for category 1 on response variable j given the explanatory variables, i.e. $\Pr(y_{ij} = 1 | \mathbf{x}_i) = \pi_j(\mathbf{x}_i)$, as

$$\pi_j(\mathbf{x}_i) = \frac{\exp[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})]}{\exp[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{0j})] + \exp[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})]}. \quad (5.1)$$

The log-odds representation of (5.1) becomes,

$$\log \left[\frac{\pi_j(\mathbf{x}_i)}{1 - \pi_j(\mathbf{x}_i)} \right] = \sum_{m=1}^M \left\{ \beta_{0m}(\gamma_{1j,m} - \gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2 - \gamma_{1j,m}^2) + \mathbf{x}_i^T \beta_m(\gamma_{1j,m} - \gamma_{0j,m}) \right\}. \quad (5.2)$$

Because each response variable belongs to a single dimension (see Chapter 4), the log odds representation can be further simplified. Suppose response variable j belongs to the first dimension so that $\gamma_{0j,m}$ and $\gamma_{1j,m}$ equal zero for all $m > 1$, i.e. all dimensions except the first one. In that case (5.2) simplifies to a single equation instead of a sum over dimensions. Moreover, as the MLD model is a type of bilinear model, for each dimension we have to fix the origin and scale.

5.2.1 Parameter Estimation

By setting the distance between the two categories of every response variable to be equal to one, i.e., $(\gamma_{1j,m} - \gamma_{0j,m}) = 1$, the MLD model can be fitted using the Generalized Estimating Equation method (Liang & Zeger, 1986). Therefore, existing statistical packages with a GEE procedure (e.g., the **geepack** package in R or the **genmod** procedure in SAS software) can be used for fitting the “restricted” MLD model on multivariate binary data. The restriction of the class points implies that explanatory variables discriminate equally well for all response variables belonging to a specific dimension.

Fitting the restricted MLD model using a GEE procedure involves a three-step approach: (1) construction of a design matrix for both the response and the explanatory variables; (2) applying the GEE method with the constructed design matrix; and (3) transforming the GEE parameters to MLD parameters.

We now show construction of the design matrix using the example presented in Table 4.1. Suppose we want to fit a 2-dimensional model on the five binary response variables. Further, suppose we like the first three response variables to be represented on the first

dimension, and the fourth and the fifth on the second dimension. Therefore define a response indicator matrix, denoted by $\mathbf{Z}$, with dimension $(J \times M)$. The response indicator matrix specifies for each response variable to which dimension it pertains, with position (j, m) equal to one if the j -th response variable belongs to the m -th dimension and zero otherwise. For the structure layed-out above,

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. \quad (5.3)$$

The design matrix for subject i is then obtained by computing the Kronecker product between the response indicator matrix and the explanatory variables vector (without intercept), $\mathbf{U}_i = \mathbf{Z} \otimes \mathbf{x}_i^T$, such that

$$\mathbf{U}_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}. \quad (5.4)$$

We concatenate $\mathbf{U}_i$ and the identity matrix to get the final design matrix, $\mathbf{S}_i = [\mathbf{I}, \mathbf{U}_i]$,

$$\mathbf{S}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}.$$

Then, a vertical concatenation of all $\mathbf{S}_i$ matrices will give us the final design matrix $\mathbf{S}$ on which the GEE method is finally applied to obtain parameter estimates of the marginal model. This results in five response specific intercepts $(\beta_{01}^*, \dots, \beta_{05}^*)$ corresponding to the first five columns of $\mathbf{S}$ and two sets of p regression weights $(\beta_{11}^*, \dots, \beta_{p1}^*$ and $\beta_{12}^*, \dots, \beta_{p2}^*)$, corresponding to the two dimensions. The MLD parameters can be derived from these as follows $\gamma_{0j,m} = -(\beta_{0j}^* + 0.5)$ for dimension, m , to which disorder j belongs, zero otherwise. The regression weights β_{jm} are equal to the regression weights obtained from GEE method, $\beta_{jm} = \beta_{jm}^*$. The number of parameters in the “restricted” MLD model then becomes $q = J + (M \times P)$ since additional constraints are imposed on the class points.

5.3 The NESDA Data

We used the Netherlands Study of Depression and Anxiety (NESDA: Penninx et al., 2008; Spinhoven et al., 2009) data as a working example in this chapter to demonstrate usage of the **mldm** package. NESDA is an ongoing cohort study designed to investigate determinants of depressive and anxiety disorders in a relatively large and representative sample of participants. In the current version of data we have, there are $N = 2,938$ subjects of age between 18 – 65 years with an average age of 42(S.D. = 13.1) in which 66.5% were female and the average number of years of education attained was 12.2(S.D. = 3.3).

The multivariate binary responses are major depressive disorder (MDD), dysthymia (DYST), generalized anxiety disorder (GAD), social phobia (SP), and panic disorder (PD). In this study, about 37.1% of the subjects had MDD, 10.2% had DYST, 15.3% had GAD, 22.4% had SP, and 28.6% had PD. The explanatory variables are the Big-Five personality traits (i.e., neuroticism, extroversion, openness to experience, agreeableness, and conscientiousness) and three background variables (i.e., age, years of education attained and

gender).

5.4 The mldm Package

5.4.1 Accessing the NESDA Data

The NESDA data is available in the **mldm** package. For demonstration purpose, we selected a random sample of 600 subjects. Thus, there are a total of 3000 records in the dataset with binary responses for the five disorders that are observed for each subject. Figure 5.1 shows how to extract the NESDA data from the **mldm** package. The dataset becomes available for us after we install and load the **mldm** package to the R environment.

```
# load the package  
library(mldm)  
  
# load the NESDA data  
data(NESDA)  
  
## View the dimension of the data  
dim(NESDA)  
  
## [1] 3000  11
```

Figure 5.1: Reading the NESDA data available in the **mldm** package.

The measurements from the first two subjects are displayed in Figure 5.2 below. The dataset is in a person-item order where measurements of the five mental disorders are nested within a subject. The first five rows represent measurements of the five mental disorders obtained for the first subject. The other measurements (i.e., row 6 – 10) are the same set of measurements for the second subject.

The last column (`pident`) has subject's identification number. The `Index` column

is used to identify which response variable (e.g., mental disorder) is measured in a given row. There are a total of five mental disorders measured for each subject. The Outcome column holds the actual binary measurements that belong to each response variable (e.g., mental disorder). For example, the first subject (`pident = 1`) has a social phobia disease because `SP = 1` while all the other measurements are zero. Whereas the second subject (`pident = 2`) has none of the mental disorders. All the other columns in Figure 5.2 represent measurements on the explanatory variables, i.e., the Big-Five personality traits and the three background variables.

```
# display measurements of the first ten subjects
print(head(NESDA, n=10), digits = 0, row.names = FALSE,
      right = FALSE)
```

GEN	AGE	EDU	N	E	O	A	C	Outcome	Index	pident
male	41	18	43	28	31	41	35	0	DYST	1
male	41	18	43	28	31	41	35	0	MDD	1
male	41	18	43	28	31	41	35	0	GAD	1
male	41	18	43	28	31	41	35	1	SP	1
male	41	18	43	28	31	41	35	0	PD	1
male	59	9	32	43	31	40	42	0	DYST	2
male	59	9	32	43	31	40	42	0	MDD	2
male	59	9	32	43	31	40	42	0	GAD	2
male	59	9	32	43	31	40	42	0	SP	2
male	59	9	32	43	31	40	42	0	PD	2

Figure 5.2: Excerpt of the NESDA data that shows records belonging to the first two subjects.

5.4.2 Model Specification and Fitting

Suppose we would like to fit a 2-dimensional MLD model on the NESDA data where the DYST and MDD response variables belong to the first dimension while the other response variables (i.e., GAD, SP and PD) belong to the second dimension. This model is also called the depression-anxiety model (Penninx et al., 2008).

This model can be fitted in the **mldm** package by first specifying a response indicator matrix (i.e., the matrix that indicates to which dimension every response belongs). Figure 5.3 shows the response indicator matrix for our 2-dimensional model.

```
# Indicator matrix
Z <- matrix(c(1,1,0,0,0,
              0,0,1,1,1), 5, 2, byrow=FALSE)
print(Z)

##      [,1] [,2]
## [1,]    1    0
## [2,]    1    0
## [3,]    0    1
## [4,]    0    1
## [5,]    0    1
```

Figure 5.3: Specification of an indicator matrix for the depression-anxiety model fitted on the NESDA data.

The multivariate logistic distance model for the NESDA data is given by

$$\begin{aligned} \eta_{im} = & \beta_{1m} \times \text{AGE} + \beta_{2m} \times \text{EDU} + \beta_{3m} \times \text{GEN} \\ & + \beta_{4m} \times \text{N} + \beta_{5m} \times \text{E} + \beta_{6m} \times \text{O} + \beta_{7m} \times \text{A} + \beta_{8m} \times \text{C}, \end{aligned} \quad (5.5)$$

where η_{im} is the i -th subject coordinate in dimension m . Since we want to fit the depression and anxiety model, the number of dimension becomes two, i.e., $M = 2$.

The R code below in Figure 5.4 shows the formula following (5.5) for every dimension. For simplicity reason, we only considered two of the Big-Five personality traits (N and E), and all the background variables. The left side of the formula (i.e., before the tilde sign) shows the response variable (Outcome) separated by a pipe operator. For a unidimensional MLD model, we would not use the pipe operator. At the right side of the formula, the explanatory variables are displayed which are again separated by the pipe operator. The pipe operator tells R that the components are dimension-specific. In our case, the MLD model is a 2-dimensional model.

```
## specify model formula
mf <- Outcome | Outcome ~ EDU + GEN + AGE + N + E |
      EDU + GEN + AGE + N + E
mf <- Formula(mf)
```

Figure 5.4: A two-dimensional representation of model formula for depression-anxiety model fitted on the NESDA data.

As model specification in MLD is dimension specific, it is possible to allow a given explanatory variable to have an effect in one of the dimension but not on the other one. For example, we can test the hypothesis that agreeableness has an effect on the first dimension (depression), but not on the second dimension (anxiety) while both openness to experience and conscientiousness having an effect on anxiety, but not on depression.

Once the response indicator matrix (Z) and the the model formula ($\mathbf{mf}$) are specified, we are now ready to fit the MLD model using the `mldm.fit` function as shown in Figure 5.5 below.

```

# fit the MLD model on NESDA data
fit <- mldm.fit(formula=mf, index = Index, resp.dim.ind = Z,
               data = NESDA, id = pident, scale=TRUE)

# display fitted model result
fit

Call:
mldm.fit(formula = mf, index = Index, resp.dim.ind = Z, data = NESDA,
        id = pident, scale = TRUE)

Formula:
Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE +
      N + E

QIC: [1] 2542.707

```

Figure 5.5: Application of the `mldm.fit` function for fitting the depression-anxiety model on the NESDA data.

Except the `scale` parameter in the `mldm.fit()` function, the input values for the other parameters are obtained both from the dataset itself (i.e., `index = Index`, `data = NESDA` and `id = pident`) and model specification (i.e., `formula = mf` and `resp.dim.ind = Z`). The `scale` argument in the `mldm.fit()` function is used for transforming the explanatory variables to z-scores.

The `mldm.fit()` function returns the model formula, and the Quasi-Information Criterion (QIC) statistics of the fitted model.

The other model outputs (e.g., the parameter estimates and the sandwich standard errors) can be obtained using the `summary()` function in R as shown in Figure 5.6 below. The `class coordinates` section of the output presents parameter estimates for $\hat{\gamma}_{kj,m}$. It represents an estimate for the coordinate of the k -th category that belong to the j -th

response variable in dimension m . The class point restriction (i.e., $\gamma_{1j,m} - \gamma_{0j,m} = 1$) can be seen from the estimates. For example, for the first response (DYST) on the first dimension, its class point estimates are $\hat{\gamma}_{01} = 2.286$ and $\hat{\gamma}_{11} = 3.286$.

The estimates of regression coefficients per dimension (i.e., β_m) then follows the class points as shown in Figure 5.6. These estimates show effect of the explanatory variables on each dimension, specifically on depression and anxiety dimensions. For example, $\hat{\beta}_{41} = 1.1286$ indicates that there is a strong positive association (p -value = 0.0000) between neuroticism and depression; similarly, with anxiety (i.e., $\hat{\beta}_{42} = 0.9856$). Whereas, extraversion has a moderate negative association with both dimensions, i.e., $\hat{\beta}_{51} = -0.4393$ with depression and $\hat{\beta}_{52} = -0.2782$ with anxiety.

The Pearson correlation between the two dimensions is also shown in the result, i.e., $\text{Corr}(\hat{\eta}_{i1}, \hat{\eta}_{i2}) = 0.98$. This result shows that there is a strong linear relationship between the positions of the subjects on the first dimension (η_{i1}) and the second dimension (η_{i2}).

The results displayed in Figure 5.6 are the default outputs by the `summary()` function. However, there are additional outputs (e.g., `npar` for number of parameters of the fitted model, etc) which are `glm`-like results, and they can be obtained by the `str(fit)` function in R.

```
# display summary of the fitted model result
```

```
summary(fit)
```

```
Call:
```

```
mldm.fit(formula = mf, index = Index, resp.dim.ind = Z, data = NESDA,  
          id = pident, scale = TRUE)
```

```
Formula:
```

```
Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE +  
          N + E
```

```
Class Coordinates:
```

```
          dim1 dim2  
gamma01 2.286 0.000  
gamma11 3.286 0.000  
gamma02 1.781 0.000  
gamma12 2.781 0.000  
gamma03 0.000 0.050  
gamma13 0.000 1.050  
gamma04 0.000 0.884  
gamma14 0.000 1.884  
gamma05 0.000 1.311  
gamma15 0.000 2.311
```

```
[1] "Regression coefficients for Dimension 1"
```

```
          estimate san.se      wald      p  
EDU1  -0.0277 0.0894  0.0957 0.7570  
GEN1  -0.0136 0.1928  0.0049 0.9439  
AGE1  -0.0440 0.0917  0.2309 0.6308  
N1      1.1286 0.1187 90.3897 0.0000  
E1     -0.4393 0.1109 15.6925 0.0001
```

```
[1] "Regression coefficients for Dimension 2"
```

	estimate	san.se	wald	p
EDU2	-0.2192	0.0759	8.3427	0.0039
GEN2	0.2367	0.1631	2.1076	0.1466
AGE2	-0.0269	0.0735	0.1343	0.7140
N2	0.9856	0.0946	108.6077	0.0000
E2	-0.2782	0.0822	11.4533	0.0007

```
Correlation among dimensions:
```

	dim1	dim2
dim1	1.00	0.98
dim2	0.98	1.00

```
QIC:
```

```
[1] 2542.707
```

Figure 5.6: Summary of the depression-anxiety model fitted on the NESDA data.

The Clustered Bootstrap Method

The parameters in the MLD model are estimated using the GEE method. Thus, the Sandwich estimators are primarily used for hypothesis testing since the model-based standard errors are biased. The `mldm.fit()` function uses this procedure at the backend as a default estimation method.

The other alternative for obtaining the standard errors of model parameters in the MLD model is to apply a clustered bootstrap technique (Sherman & Le Cessie, 1997; De Rooij & Worku, 2012). In this case, the re-sampling procedure is applied on the subject (cluster) level so that the correlations between the measurements within each subject are retained.

The clustered bootstrap method is implemented in the **mldm** package. The `bootstrap` argument in the `mldm.fit()` function is used for this purpose. The function returns both the parametric and the non-parametric confidence intervals of the model parameters. Figure 5.7 shows application of the new argument, i.e., `bootstrap = 1000`. The number of replicates used here is also what is recommended in practice.

```
# fit the MLD model using Clustered Bootstrap method
fit_boot <- mldm.fit(formula=mf, index = Index, resp.dim.ind = Z,
                    data = NESDA, id = pident, scale=TRUE, bootstrap=1000)
# fit_boot
```

Figure 5.7: Application of the Clustered Bootstrap method with the MLD model.

The `summary()` function with an additional argument can be used to obtain the clustered bootstrapped standard errors as shown in Figure 5.8 below.

Generally, layout of the model results are very similar to the one presented before in Figure 5.6. What makes this result different is that the standard errors are estimated differently (clustered bootstrap version) with a 95% confidence interval (CI) for the parameter estimates. By default, the confidence intervals are the parametric ones. The nonparametric confidence intervals can be obtained by specifying an additional argument in the `summary()` function, i.e., `boot.nonparam=TRUE`.

```
summary(fit_boot, bootstrap=TRUE)
```

Call:

```
mldm.fit(formula = mf, index = Index, resp.dim.ind = Z, data = NESDA,
         id = pident, scale = TRUE, bootstrap = 1000)
```

Formula:

```
Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE +
      N + E
```

Class Coordinates:

	dim1	dim2
gamma01	2.286	0.000
gamma11	3.286	0.000
gamma02	1.781	0.000
gamma12	2.781	0.000
gamma03	0.000	0.050
gamma13	0.000	1.050
gamma04	0.000	0.884
gamma14	0.000	1.884
gamma05	0.000	1.311
gamma15	0.000	2.311

[1] "Regression coefficients for Dimension 1"

	estimate	boot.se	boot.wald	p	boot.ll	boot.ul
EDU1	-0.0277	0.0882	0.0983	0.7538	-0.2006	0.1452
GEN1	-0.0136	0.1990	0.0046	0.9457	-0.4035	0.3764
AGE1	-0.0440	0.0934	0.2224	0.6372	-0.2271	0.1390
N1	1.1286	0.1220	85.5033	0.0000	0.8893	1.3678
E1	-0.4393	0.1137	14.9184	0.0001	-0.6622	-0.2164

[1] "NB: The confidence intervals are the parameteric ones!"

```
[1] "Regression coefficients for Dimension 2"

      estimate boot.se boot.wald      p boot.ll boot.ul
EDU2  -0.2192  0.0724    9.1658 0.0025 -0.3610 -0.0773
GEN2   0.2367  0.1667    2.0172 0.1555 -0.0899  0.5634
AGE2  -0.0269  0.0748    0.1297 0.7188 -0.1736  0.1197
N2     0.9856  0.0973  102.6649 0.0000  0.7949  1.1762
E2    -0.2782  0.0832   11.1704 0.0008 -0.4413 -0.1150

[1] "NB: The confidence intervals are the parametric ones!"

Correlation among dimensions:

      dim1 dim2
dim1  1.00 0.98
dim2  0.98 1.00

QIC:
[1] 2542.707
```

Figure 5.8: Summary of the depression-anxiety model fitted on the NESDA data using the Clustered Bootstrap method.

5.4.3 The Biplot for MLD Model

To enhance interpretation of the model the results of an MLD model can be graphically represented in a biplot (Gabriel, 1971; Gower & Hand, 1996; Gower et al., 2011). The biplot represents the subjects, the response variables, and the predictor variables so that the relationship between predictors and responses can be read from the graph.

The `biplot()` function in the **mldm** package can be used to display the results of the MLD model in a biplot. Figure 5.9 shows the application of this function. The biplot for the depression-anxiety model is presented in Figure 5.10.

```
# biplot of the MLD model
biplot(fit)
```

Figure 5.9: Application of the `biplot()` function available in the **mldm** package.

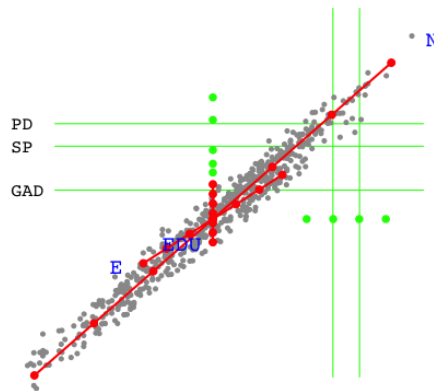


Figure 5.10: The biplot for depression-anxiety model fitted on the NESDA data.

5.4.4 Model Selection using QIC

In statistical analysis we often select a parsimonious and best fitting model from a set of candidate models given the data. In the MLD model, we select not only predictor variables for the final model, but also the dimensionality of the model must be determined.

Pan (2001) proposed the quasi-likelihood under the independence model criterion

(QIC) as an extension of Akaike Information Criterion (AIC) to GEE:

$$\text{QIC} = -2L(\boldsymbol{\theta}) + 2 \text{trace}(\hat{\boldsymbol{\Omega}}_I^{-1} \hat{\mathbf{V}}_R), \quad (5.6)$$

where $\hat{\mathbf{V}}_R$ represents the robust variance estimator obtained under the assumption of a general “working” covariance structure R ; and $\hat{\boldsymbol{\Omega}}$ is for the naive variance estimator obtained under the assumption of an independence correlation structure. Pan (2001) also proposed a simplified version of QIC when $\text{trace}(\hat{\boldsymbol{\Omega}}_I^{-1} \hat{\mathbf{V}}_R) \approx \text{trace}(\mathbf{I}) = q$, i.e.,

$$\text{QIC}_u = -2L(\boldsymbol{\theta}) + 2q.$$

Yu and De Rooij (2013) studied the performance of QIC_u for determining the dimensionality of the Trend Vector Model (TVM). Both the Trend Vector model and the MLD model are marginal models in a distance framework, where the first is used for longitudinal multinomial response variables and the latter for multivariate binary responses. Yu and De Rooij (2013) recommended QIC_u for determining the dimensionality of the distance model.

In the MLD model, we use QIC_u fit statistics both for determining the dimensionality of the model and for variable selection. The model with the lowest QIC_u statistics is considered the most parsimonious and best fitting model. As recommended in Yu and De Rooij (2013), we first determine the dimensionality of the model and then proceed to the variable selection.

The QIC_u fit statistics is implemented in the **mldm** package and its value can be extracted by specifying `fit$QIC`, where `fit` is an object of the fitted model.

Model selection for Dimensionality

For demonstration purpose we compare a unidimensional MLD model against a 2-dimensional MLD model fitted on the NESDA data with the same set of explanatory variables. For

dimensionality comparison, both the response indicator matrix and model formula should be redefined. Figure 5.11 shows specification of the indicator matrix for the unidimensional candidate model. For the 2-dimensional model, we use the same indicator matrix from the depression-anxiety model presented before in Figure 5.3.

```
# an indicator matrix for the unidimensional MLD model
Z1 <- matrix(c(1,1,1,1,1), 5, 1, byrow=FALSE)
Z1

[,1]
[1,] 1
[2,] 1
[3,] 1
[4,] 1
[5,] 1
```

Figure 5.11: Specification of an indicator matrix for candidate models with respect to dimensionality in the model.

The new model formula for the unidimensional candidate model is shown in Figure 5.12 where the explanatory variables are specified only in the first dimension. For the 2-dimensional MLD model we use the same model formula that was defined above for the depression-anxiety model (i.e., `mf`) in Figure 5.4.

```
# formula for the unidimensional model
mf1 <- Outcome ~ EDU + GEN + AGE + N + E
mf1 <- Formula(mf1)
```

Figure 5.12: Specification of model formula for a unidimensional MLD model.

Figure 5.13 displays the QIC fit statistics values for the two candidate models. We can conclude that the unidimensional MLD model fits the data well, although the two QIC values are very similar, i.e., $QIC^{1D} = 2541.235$ and $QIC^{2D} = 2542.707$. Note that only $N = 600$ subjects were used for fitting the candidate models. If all subjects (i.e., $N = 2938$) were included in the analysis, the 2-dimensional (depression-anxiety) model would fit the NESDA data better.

```
# fit the unidimensional MLD model
fit_dim1 <- mldm.fit(formula=mf1, index = Index, resp.dim.ind = Z1,
                    data = NESDA, id = pident, scale=TRUE)

# get the QIC value
fit_dim1$QIC

[1] 2541.235

# fit the 2-dimensional MLD model
fit_dim2 <- mldm.fit(formula=mf, index = Index, resp.dim.ind = Z,
                    data = NESDA, id = pident, scale=TRUE)

# get the QIC value
fit_dim2$QIC

[1] 2542.707
```

Figure 5.13: Model selection in MLD model for dimensionality.

Model selection for Explanatory variables

For demonstration purpose, let us compare two candidate 2-dimensional MLD models that only differs on the explanatory variables. That is, (1) a depression-anxiety model with only the background variables (i.e., education, gender and age); and, (2) a depression-anxiety

model with both the background variables and two of the Big-5 personality traits (i.e., neuroticism and extroversion). The model formula for each candidate model is presented below in Figure 5.14.

```
# the first model, i.e., only the background variables
mf2a <- Outcome | Outcome ~ EDU + GEN + AGE |
      EDU + GEN + AGE
mf2a <- Formula(mf2a)

# the second model, i.e., with the background variables and
# two of the personality traits
mf2b <- Outcome | Outcome ~ EDU + GEN + AGE + N + E |
      EDU + GEN + AGE + N + E
mf2b <- Formula(mf2b)
```

Figure 5.14: Model formula structure of the candidate MLD models.

In Figure 5.15 the candidate models are fitted and the QIC results are obtained. The QIC fit statistics of the candidate models are then extracted by specifying `fit$QIC`. It can be concluded that the 2-dimensional MLD model with both the background variables and the two personality traits fits the data better than the model without (N, E) since it has a smaller QIC value, i.e., $QIC = 2542.707$.

```

# fit the first model
fit_dim2a <- mldm.fit(formula=mf2a, index = Index, resp.dim.ind = Z,
                      data = NESDA, id = pident, scale=TRUE)

# get the QIC value for the first model
fit_dim2a$QIC

[1] 3056.799

# fit the second model
fit__dim2b <- mldm.fit(formula=mf2b, index = Index, resp.dim.ind = Z,
                      data = NESDA, id = pident, scale=TRUE)

# get the QIC value for the second model
fit_dim2b$QIC

[1] 2542.707

```

Figure 5.15: Model selection in MLD model for explanatory variables.

5.5 Conclusion and Discussion

In this chapter we showed an application of the **mldm** package using a psychological dataset. The **mldm** package fits the newly proposed Multivariate Logistic Distance (MLD) model for analyzing multivariate binary data.

The `mldm.fit()` function in the **mldm** package supports two different estimation techniques for obtaining standard errors for model parameters in the MLD model, namely the Sandwich estimator from GEE method and the clustered bootstrap method. Using the `biplot()` function, one can easily produce a biplot for the fitted model. The QIC object returned from the `mldm.fit()` function can be used to compare candidate MLD

models. The QIC fit statistics is used to determine: (1) the dimensionality of the model, and (2) the structure of the explanatory variables.

We made the **mldm** package available on the online repository system GitHub. The following link can be used to get access to the package: <https://github.com/workuhm1/mldm-package-github>.

Chapter 6

Conclusions and Discussions

In this dissertation our main aim was to develop a methodology, based on the Ideal Point Classification (IPC: De Rooij, 2009a) model, for analyzing multivariate categorical data which requires a less assumptions and takes the data as it is. Existing methodology makes unverifiable assumptions (e.g., latent variable models and structural equation models that make a normality assumption for latent variables) or requires the independent variables to to be categorized (e.g., the GEE2 method for marginal models). In Appendix C we showed limitation of latent variable models regarding normality assumption of factor scores using empirical data.

Structural equation models were originally proposed for analysis of continuous (or interval) indicator variables. Recently, confirmatory factor analysis and structural equation models have been applied for data with dichotomous indicators and with only a few indicators per latent variable, i.e., 2 or 3 (Krueger, 1999; Beesdo-Baum et al., 2009). Using a Monte Carlo simulation study, we showed in Chapter 2 that latent variable models applied on such type of data performed poorly with higher incidence of improper solutions, poor quality of recovering the true factor scores, too conservative or inflated type-I error rates, and weak power.

In this dissertation we further developed the IPC model for analyzing multivariate

binary data. The IPC model is a probabilistic multidimensional unfolding model and closely related to the Ideal Point Discriminant Analysis (IPDA: Takane et al., 1987). De Rooij (2009a) showed that the IPC model for a univariate dependent variable with C categories in $M = C - 1$ dimensions equals the Multinomial Baseline Category Logit (MBCL: Agresti, 2007, chap. 6) model. For a dichotomous dependent variable, it was shown by De Rooij (2009a) that the IPC model with only one dimension equals a simple Logistic Regression (LR: Agresti, 2007, chap. 4) model. The IPC model was further studied in Yu and De Rooij (2013). The IPC model has been successfully applied to investigate trends in living conditions for psychiatric homeless people over time (De Rooij, 2009b); to look at vote transitions between political parties (De Rooij, 2011); and, in preferential choice for television programs of children (De Rooij & Schouteden, 2009).

In Chapter 3 we studied properties of IPC model for analyzing bivariate binary data. A bivariate logistic regression set-up (Bahadur, 1961; Palmgren, 1989; Lipsitz et al., 1990) is used so that the Euclidean space of the dependent variables is three-dimensional. In this case the first dimension pertains to the prevalence of the first dependent variable (e.g., breathlessness in the Coalminers study); the second pertains to the prevalence of the second variable (e.g., wheeze); and, the third dimension pertains to the association between the two dependent variables (e.g., the association between breathlessness and wheeze in the Coalminers study).

Based on a simulation study and analytical derivations, we showed that the 3-dimensional IPC model for bivariate binary data fully recovered the association structure between the two dependent variables, but misspecified the univariate marginal models. On the other hand, the 2-dimensional IPC model with a “fixed” set of class points (i.e., the first two columns of the indicator matrix presented in 1.16) fully recovered the marginal models, but assumes an “independent” association structure between the dependent variables. With a “free” set of class points, the 2-dimensional IPC model represented the association model as a form of restricted model. However, this model misspecified both the

marginal models and the association model.

To fully recover both the marginal models and the association model of a bivariate binary data, a re-parameterization of the IPC model was proposed. The newly proposed model is called a Bivariate IPC (BIPC, Worku & De Rooij, 2017a) model, and using a simulation study it was shown that this model recovered both the marginal models and the association model well. Unlike existing marginal models for bivariate binary data, the BIPC model can provide us with a biplot which enhances the interpretation of the model. The BIPC model can be extended easily for bivariate polytomous data by adding class coordinates to accommodate the additional response categories.

A limitation of the BIPC model, however, is that it is not straightforward to extend it for analyzing multivariate binary data (i.e., with more than two binary or polytomous responses). This is due to the fact that both the pairwise and higher-order association structure parameters between the dependent variables must be specified in the likelihood function. With three binary responses (i.e., Y_1 , Y_2 , and Y_3), for example, three pairwise associations and a three-way association parameters must be specified which makes the computation cumbersome. Due to this limitation of the BIPC model, we proposed a new distance-based marginal model in Chapter 4, namely the Multivariate Logistic Distance (MLD) model, for analyzing multivariate binary data.

The MLD model can be used to simultaneously assess the dimensional structure of the data and to study the effect of the predictor variables on the response variables. The MLD model belongs to the family of marginal models for multivariate responses, as opposed to latent variable models and conditionally specified models. By setting the distance between the two categories of every response variable to be equal, the MLD model can be fitted using the GEE estimation method (Liang & Zeger, 1986). Therefore, existing statistical packages built for the GEE procedure, e.g., the **genmod** procedure in SAS or the **geepack** package in R, can be used for fitting the MLD model. Without the equality constraint, the MLD model is a general model which can be fitted by its own right (Worku & De

Rooij, 2018). The former is sometimes referred to as the “restricted” MLD model, and the later as the “unrestricted” MLD model.

The MLD model is related to Canonical Correspondence Analysis (CCA: Ter Braak, 1986; Ter Braak & Verdonschot, 1995) which is a multivariate method used for ordination axes that maximizes the separation among the multivariate binary responses. The main two differences between the CCA and the MLD model are the following. Firstly, the model set-up is different, i.e., the MLD model is built in a logistic framework where as the CCA is in a Gaussian framework. Due to this difference, the MLD can provide a clear interpretation in terms of (log)-odds and probabilities. The second reason is that unlike in CCA, the MLD model can position responses (e.g., mental disorders) on certain dimensions driven by the theories that we would like to test.

Like the BIPC model, the MLD model can also be extended for analyzing multivariate polytomous data. The polytomous situation, however, is often more complicated than the binary one. The binary model for every response variable in the MLD framework is by definition unidimensional, which is not the case for polytomous data. Therefore, we recommend further study to fully understand the behavior of the MLD model with multivariate polytomous data.

Regarding model assumptions, it is worth mentioning the following two points. The first point is that the MLD model makes a strong linearity assumption regarding the explanatory variables, i.e., the model assumes that the explanatory variables are linearly related to logit transform of the class probabilities. However, this assumption could be solved for example by using polynomial functions of the original explanatory variables. The second point is that, compared to structural equation models, the MLD model does not have the assumption of a normal distribution for the latent variables anymore.

In Chapter 5 we presented an **mldm** package that was developed in R for fitting the MLD model. The main function in the **mldm** package responsible for fitting the MLD model is `mldm.fit()`. The function supports the two approaches, namely the Sandwich

estimator from GEE method and the clustered bootstrap method, which are used for obtaining standard errors of model parameter estimates. These estimation techniques are applied in the MLD model to correct bias of the Hessian matrix. Using the `biplot()` function, one can produce a biplot for the fitted model. The QIC object returned by the `mldm.fit()` function can be used to compare different candidate MLD models. The fit statistics is mainly used to determine the structure of the fixed part in the MLD model (i.e., set of explanatory variables); and, the dimensionality of the MLD model.

We conclude the dissertation with some recommendations for future researchers. It is important to note that we used an advantageous design for our Monte Carlo simulation study in Chapter 2. The latent variables were generated from a bivariate normal distribution. Moreover, the population model was correctly specified. In empirical studies it is likely that assumptions are only partially valid. Moreover, the fitted model could be misspecified; for example, an important indicator variable may not have been included in the analysis. Under such conditions we would expect even more improper solutions and factor scores that are further off than what we found in our current study. Therefore, these methods performed poorly for this type of data and therefore must be used carefully. An alternative statistical model which requires less assumptions might be more appropriate, for example the multivariate logistic distance model (Worku & De Rooij, 2018).

Appendix A

List of Tables that belong to Chapter 2

Table A.1: Parameter estimates of the 2-way interaction logistic regression model fitted on the nonconvergence data. For simplicity, we denote the design variables as, a: type of indicators; b: number of indicators; c: factor structure; d: correlation between underlying latent variables; and, e: sample size.

Effect		Estimates (S.E.)	p-value	OR
Intercept		-22.14 (2.548)	< 0.0001	
Type of Indicators (a)	BLR	11.27 (1.578)	< 0.0001	78,766.54**
	BMR	5.96 (1.637)	0.0003	385.92**
Number of Indicators (b)	6	6.58 (1.451)	< 0.0001	717.83**
	10	3.27 (1.534)	0.0333	26.20**
Factor structure (c)	Weak	9.72 (1.555)	< 0.0001	16,724.23**
	Moderate	6.53 (1.573)	< 0.0001	683.68**
Correlation between Factors (d)	Independence	0.06 (0.437)	0.8878	1.06
	Moderate	-2.57 (0.729)	0.0004	0.08**
Sample size (e)	50	13.39 (2.438)	< 0.0001	655,792.70**
	100	10.03 (2.44)	< 0.0001	22,781.63**
	300	4.88 (2.471)	0.0481	131.95**
a × b	[a=BLR] × [b=6]	-1.70 (0.293)	< 0.0001	0.18**
	[a=BMR] × [b=6]	-0.34 (0.295)	0.2500	0.71
	[a=BLR] × [b=10]	-1.32 (0.300)	< 0.0001	0.27*
	[a=BMR] × [b=10]	-0.66 (0.302)	0.0286	0.52
a × c	[a=BLR] × [c=Weak]	-4.50 (0.646)	< 0.0001	0.01**
	[a=BMR] × [c=Weak]	-2.85 (0.647)	< 0.0001	0.06**
	[a=BLR] × [c=Moderate]	-3.27 (0.652)	< 0.0001	0.04**
	[a=BMR] × [c=Moderate]	-2.17 (0.651)	0.0009	0.11**

a × d	[a=BLR] × [d=Independence]	-1.31 (0.204)	< 0.0001	0.27*
	[a=BMR] × [d=Independence]	-1.15 (0.201)	< 0.0001	0.32
	[a=BLR] × [d=Moderate]	-0.72 (0.214)	0.0008	0.49
	[a=BMR] × [d=Moderate]	-0.84 (0.210)	< 0.0001	0.43
a × e	[a=BLR] × [e=50]	-2.51 (1.404)	0.0738	0.08**
	[a=BMR] × [e=50]	0.15 (1.470)	0.9213	1.16
	[a=BLR] × [e=100]	-0.65 (1.408)	0.6441	0.52
	[a=BMR] × [e=100]	1.09 (1.473)	0.4597	2.97
	[a=BLR] × [e=300]	0.85 (1.448)	0.5591	2.33
	[a=BMR] × [e=300]	1.68 (1.512)	0.2669	5.36**
b × c	[b=6] × [c=Weak]	-0.75 (0.149)	< 0.0001	0.47
	[b=10] × [c=Weak]	-0.16 (0.148)	0.2745	0.85
	[b=6] × [c=Moderate]	-0.02 (0.152)	0.9070	0.98
	[b=10] × [c=Moderate]	0.01 (0.150)	0.9258	1.01
b × d	[b=6] × [d=Independence]	1.07 (0.129)	< 0.0001	2.92
	[b=10] × [d=Independence]	0.56 (0.126)	< 0.0001	1.76
	[b=6] × [d=Moderate]	1.12 (0.132)	< 0.0001	3.06*
	[b=10] × [d=Moderate]	0.60 (0.130)	< 0.0001	1.82
b × e	[b=6] × [e=50]	-4.27 (1.413)	0.0025	0.01**
	[b=10] × [e=50]	-1.80 (1.498)	0.2300	0.17**
	[b=6] × [e=100]	-3.21 (1.413)	0.0230	0.04**
	[b=10] × [e=100]	-1.28 (1.497)	0.3927	0.28*
	[b=6] × [e=300]	-2.08 (1.417)	0.1418	0.13**
	[b=10] × [e=300]	-0.74 (1.502)	0.6208	0.48
c × d	[c=Weak] × [d=Independence]	-0.40 (0.131)	0.0024	0.67
	[c=Moderate] × [d=Independence]	0.00 (0.132)	0.9850	1.00
	[c=Weak] × [d=Moderate]	-0.31 (0.133)	0.0181	0.73
	[c=Moderate] × [d=Moderate]	-0.53 (0.133)	< 0.0001	0.59
c × e	[c=Weak] × [e=50]	-4.29 (1.404)	0.0022	0.01**
	[c=Moderate] × [e=50]	-2.79 (1.421)	0.0497	0.06**
	[c=Weak] × [e=100]	-3.37 (1.404)	0.0164	0.03**
	[c=Moderate] × [e=100]	-2.17 (1.421)	0.1259	0.11**
	[c=Weak] × [e=300]	-1.38 (1.410)	0.3288	0.25**
	[c=Moderate] × [e=300]	-0.76 (1.427)	0.5921	0.47
d × e	[d=Independence] × [e=50]	0.62 (0.330)	0.0602	1.86
	[d=Moderate] × [e=50]	2.89 (0.666)	< 0.0001	17.98**
	[d=Independence] × [e=100]	0.77 (0.329)	0.0185	2.17
	[d=Moderate] × [e=100]	2.92 (0.665)	< 0.0001	18.57**
	[d=Independence] × [e=300]	1.12 (0.331)	0.0007	3.06*

	[d=Moderate] × [e=300]	2.60 (0.666)	< 0.0001	13.40**
--	------------------------	--------------	----------	---------

*medium effect size, according to Ferguson (2009).

**large effect size, according to Ferguson (2009).

Table A.2: Parameter estimates of the 2-way interaction logistic regression model fitted on the *Heywood* data. For simplicity, we denote the design variables as, a: type of indicators; b: number of indicators; c: factor structure; d: correlation between underlying latent variables; and, e: sample size.

Effect		Estimates (S.E.)	p-value	OR
Intercept		-20.61 (2.813)	< 0.0001	
Type of Indicators (a)	BLR	8.73 (1.455)	< 0.0001	6,179.20
	BMR	0.33 (1.950)	0.8664	1.39
Number of Indicators (b)	6	4.13 (1.456)	0.0045	62.28
	10	-0.49 (1.934)	0.7982	0.61
Factor structure (c)	Weak	6.96 (1.504)	< 0.0001	1,049.37
	Moderate	1.65 (1.962)	0.3992	5.23
Correlation between Factors (d)	Independence	5.52 (1.450)	0.0001	248.30
	Moderate	3.75 (1.571)	0.0170	42.55
Sample size (e)	50	13.01 (2.708)	< 0.0001	445,359.70
	100	7.53 (2.709)	0.0054	1,865.31
	300	1.85 (2.794)	0.5075	6.37
a × b	[a=BLR] × [b=6]	-1.46 (0.285)	< 0.0001	0.23**
	[a=BMR] × [b=6]	0.72 (0.349)	0.0390	2.06
	[a=BLR] × [b=10]	-0.97 (0.300)	0.0013	0.38
	[a=BMR] × [b=10]	0.50 (0.360)	0.1687	1.64
a × c	[a=BLR] × [c=Weak]	-3.14 (0.322)	< 0.0001	0.04**
	[a=BMR] × [c=Weak]	-0.11 (0.414)	0.7826	0.89
	[a=BLR] × [c=Moderate]	-2.53 (0.325)	< 0.0001	0.08**
	[a=BMR] × [c=Moderate]	-0.04 (0.416)	0.9312	0.97
a × d	[a=BLR] × [d=Independence]	-1.61 (0.197)	< 0.0001	0.20**
	[a=BMR] × [d=Independence]	-0.96 (0.194)	< 0.0001	0.38
	[a=BLR] × [d=Moderate]	-1.25 (0.201)	< 0.0001	0.29*
	[a=BMR] × [d=Moderate]	-0.78 (0.197)	< 0.0001	0.46
a × e	[a=BLR] × [e=50]	-5.23 (1.376)	0.0001	0.01**
	[a=BMR] × [e=50]	-0.76 (1.861)	0.6843	0.47
	[a=BLR] × [e=100]	-2.61 (1.373)	0.0578	0.07**
	[a=BMR] × [e=100]	0.02 (1.861)	0.9928	1.02
	[a=BLR] × [e=300]	-0.65 (1.380)	0.6355	0.52
	[a=BMR] × [e=300]	0.51 (1.868)	0.7858	1.66
b × c	[b=6] × [c=Weak]	0.08 (0.370)	0.8364	1.08
	[b=10] × [c=Weak]	1.23 (0.474)	0.0096	3.41*
	[b=6] × [c=Moderate]	1.44 (0.434)	0.0009	4.22**
	[b=10] × [c=Moderate]	1.81 (0.530)	0.0006	6.12**
b × d	[b=6] × [d=Independence]	-0.14 (0.309)	0.6619	0.87

	[b=10] × [d=Independence]	0.28 (0.329)	0.3979	1.32
	[b=6] × [d=Moderate]	-0.37 (0.311)	0.2353	0.69
	[b=10] × [d=Moderate]	0.01 (0.333)	0.9721	1.01
b × e	[b=6] × [e=50]	-1.49 (1.337)	0.2662	0.23**
	[b=10] × [e=50]	0.54 (1.815)	0.7656	1.72
	[b=6] × [e=100]	0.50 (1.341)	0.7107	1.64
	[b=10] × [e=100]	1.83 (1.817)	0.3138	6.23**
	[b=6] × [e=300]	1.01 (1.363)	0.4581	2.75
	[b=10] × [e=300]	1.98 (1.835)	0.2806	7.24**
c × d	[c=Weak] × [d=Independence]	-1.37 (0.361)	0.0001	0.25**
	[c=Moderate] × [d=Independence]	-0.56 (0.367)	0.1261	0.57
	[c=Weak] × [d=Moderate]	-1.04 (0.373)	0.0051	0.35
	[c=Moderate] × [d=Moderate]	-0.63 (0.381)	0.0978	0.53
c × e	[c=Weak] × [e=50]	-2.55 (1.377)	0.0646	0.08**
	[c=Moderate] × [e=50]	0.26 (1.852)	0.8906	1.29
	[c=Weak] × [e=100]	-0.48 (1.378)	0.7273	0.62
	[c=Moderate] × [e=100]	2.10 (1.853)	0.2577	8.14**
	[c=Weak] × [e=300]	2.02 (1.503)	0.1793	7.52**
	[c=Moderate] × [e=300]	4.06 (1.948)	0.0372	57.89**
d × e	[d=Independence] × [e=50]	-2.78 (1.361)	0.0414	0.06**
	[d=Moderate] × [e=50]	-1.46 (1.486)	0.3263	0.23**
	[d=Independence] × [e=100]	-1.98 (1.358)	0.1456	0.14**
	[d=Moderate] × [e=100]	-0.87 (1.483)	0.5558	0.42
	[d=Independence] × [e=300]	-0.90 (1.364)	0.5115	0.41
	[d=Moderate] × [e=300]	-0.24 (1.489)	0.8705	0.78

* medium effect size, according to Ferguson (2009).

** large effect size, according to Ferguson (2009).

Table A.3: Effect size of the 2-way interaction ANOVA model fitted on the average correlations reported in Table 2.5. The design variables are denoted by letters, i.e., a: type of indicators; b: number of indicators; c: factor structure; d: correlation between latent variables; and e: sample size.

Effect	F (df)	p -value	η^2
Type of Indicators (a)	35549.88 (2)	0.000	0.736**
Number of Indicators (b)	13661.71 (2)	0.000	0.517**
Factor structure (c)	44178.4 (2)	0.000	0.776**
Correlation between Factors (d)	2643.23 (2)	0.000	0.172**
Sample size (e)	641.25 (3)	0.000	0.070*
Interactions			
a $\times$ b	270.02 (4)	0.000	0.041
a $\times$ c	1704.7 (4)	0.000	0.211**
a $\times$ d	7.76 (4)	0.000	0.001
a $\times$ e	9.03 (6)	0.000	0.002
b $\times$ c	74.34 (4)	0.000	0.012
b $\times$ d	4.53 (4)	0.001	0.001
b $\times$ e	3.09 (6)	0.005	0.001
c $\times$ d	18.07 (4)	0.000	0.003
c $\times$ e	46.14 (6)	0.000	0.011
d $\times$ e	7.03 (6)	0.000	0.002

*medium effect size, according to Cohen (1988).

**large effect size, according to Cohen (1988).

Table A.4: Observed power for the relationship between X_7 and the second factor, $\gamma_{72} = 0.10$. The number of replications per cell differ because of improper solutions. Dashed lines indicate no valid results were obtained for that cell.

		Sample Size															
		50				100				300				3000			
		Number of Indicators															
Type of Indicators	Factor structure	Correlation between factors															
		6	10	16	6	10	16	6	10	16	6	10	16	6	10	16	
BLR	Weak	Independence	—	—	—	0.00	0.06	0.00	0.07	0.08	0.13	0.39	0.37	0.39	0.37	0.57	
		Moderate	—	—	—	0.04	0.03	0.04	0.05	0.06	0.11	0.44	0.47	0.44	0.47	0.47	
		Strong	—	—	—	0.00	0.00	0.08	0.02	0.06	0.09	0.40	0.45	0.40	0.45	0.53	
	Moderate	Independence	—	—	—	0.00	0.00	0.00	0.03	0.05	0.05	0.40	0.47	0.40	0.47	0.59	
		Moderate	—	—	—	0.00	0.00	0.00	0.07	0.04	0.06	0.47	0.42	0.47	0.42	0.56	
		Strong	0.00	—	—	0.00	0.06	0.00	0.08	0.07	0.08	0.39	0.52	0.39	0.52	0.50	
	Strong	Independence	—	0.00	0.00	0.00	0.00	0.08	0.12	0.03	0.17	0.40	0.52	0.40	0.52	0.43	
		Moderate	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.06	0.08	0.50	0.49	0.50	0.49	0.50	
		Strong	—	0.00	—	0.00	0.00	0.00	0.00	0.00	0.11	0.54	0.53	0.54	0.53	0.38	
BMR	Weak	Independence	0.00	0.03	0.05	0.07	0.06	0.10	0.05	0.14	0.13	0.60	0.65	0.60	0.65	0.83	
		Moderate	0.06	0.00	0.01	0.05	0.08	0.05	0.09	0.11	0.14	0.66	0.66	0.66	0.66	0.68	
		Strong	0.03	0.07	0.03	0.06	0.08	0.04	0.11	0.12	0.11	0.63	0.66	0.63	0.66	0.77	
	Moderate	Independence	0.00	0.07	0.08	0.10	0.04	0.03	0.12	0.11	0.14	0.67	0.70	0.67	0.70	0.84	
		Moderate	0.00	0.06	0.08	0.05	0.07	0.09	0.08	0.15	0.11	0.67	0.75	0.67	0.75	0.68	
		Strong	0.06	0.07	0.03	0.04	0.12	0.07	0.13	0.16	0.14	0.67	0.78	0.67	0.78	0.82	
	Strong	Independence	0.00	0.00	0.00	0.14	0.00	0.13	0.14	0.13	0.18	0.60	0.64	0.60	0.64	0.70	
		Moderate	0.00	0.00	—	0.09	0.00	0.00	0.00	0.21	0.33	0.62	0.75	0.62	0.75	0.70	
		Strong	0.00	—	0.00	0.00	0.25	0.25	0.00	0.00	0.00	0.62	0.71	0.62	0.71	0.73	
Interval	Weak	Independence	0.05	0.12	0.05	0.12	0.10	0.14	0.12	0.06	0.06	0.65	0.80	0.65	0.80	0.81	
		Moderate	0.10	0.08	0.05	0.09	0.11	0.12	0.11	0.12	0.13	0.68	0.79	0.68	0.79	0.78	
		Strong	0.09	0.13	0.06	0.07	0.15	0.13	0.11	0.18	0.14	0.63	0.81	0.63	0.81	0.87	
	Moderate	Independence	0.09	0.08	0.06	0.10	0.14	0.11	0.13	0.15	0.11	0.82	0.80	0.82	0.80	0.76	
		Moderate	0.08	0.07	0.08	0.07	0.11	0.05	0.10	0.16	0.11	0.67	0.81	0.67	0.81	0.86	
		Strong	0.13	0.12	0.15	0.08	0.10	0.15	0.14	0.18	0.11	0.79	0.88	0.79	0.88	0.87	
	Strong	Independence	0.13	0.14	0.15	0.05	0.13	0.12	0.21	0.17	0.15	0.89	0.84	0.89	0.84	0.88	
		Moderate	0.14	0.09	0.06	0.06	0.11	0.06	0.21	0.17	0.16	0.80	0.94	0.80	0.94	0.89	
		Strong	0.11	0.18	0.11	0.12	0.09	0.12	0.13	0.13	0.13	0.84	0.89	0.84	0.89	0.85	

Table A.5: Observed power for the relationship between X_4 and the second factor, $\gamma_{42} = 0.95$. The number of replications per cell differ because of improper solutions. Dashed lines indicate no valid results were obtained for that cell.

Type of Indicators		Factor structure	Sample Size															
			50				100				300				3000			
			Number of Indicators															
		Correlation between factors	6	10	16	6	10	16	6	10	16	6	10	16	6	10	16	
BLR	Weak	Independence	—	—	—	0.38	0.43	0.62	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	—	—	—	0.38	0.53	0.64	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	—	—	—	0.17	0.43	0.72	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Moderate	Independence	—	—	—	0.31	0.50	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	—	—	—	0.53	0.70	0.80	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.00	—	—	0.26	0.56	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Strong	Independence	—	0.00	0.00	0.00	0.70	0.58	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.00	0.00	0.00	0.80	0.62	0.36	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	—	0.00	—	1.00	0.33	0.50	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
BMR	Weak	Independence	0.71	0.67	0.70	0.92	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.50	0.69	0.79	0.89	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.74	0.72	0.79	0.90	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Moderate	Independence	0.64	0.72	0.88	0.97	1.00	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.55	0.64	0.75	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.55	0.75	0.84	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Strong	Independence	1.00	0.67	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.50	0.33	—	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.50	—	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Interval	Weak	Independence	0.67	0.82	0.89	0.90	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.53	0.79	0.85	0.97	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.58	0.81	0.85	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Moderate	Independence	0.77	0.91	0.89	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.82	0.89	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.80	0.88	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	Strong	Independence	0.93	0.92	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Moderate	0.90	0.97	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
		Strong	0.91	0.94	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

Appendix B

Identification of the BIPC Model in Chapter 3

In this Section, we address identification issues of the BIPC model. As shown in equation (3.25), the marginal models are represented in the BIPC model as,

$$\pi_{i1\cdot} = \frac{\exp(-0.5\delta_{i1\cdot})}{\exp(-0.5\delta_{i0\cdot}) + \exp(-0.5\delta_{i1\cdot})},$$

$$\pi_{i\cdot 1} = \frac{\exp(-0.5\delta_{i\cdot 1})}{\exp(-0.5\delta_{i\cdot 0}) + \exp(-0.5\delta_{i\cdot 1})},$$

where $\delta_{il\cdot} = \sum_{m=1}^2 (\eta_{im} - \gamma_{l\cdot m})^2$ and $\delta_{i\cdot l} = \sum_{m=1}^2 (\eta_{im} - \gamma_{l\cdot m})^2$, $l = 0, 1$. Like simple logistic regression (LR) model, the BIPC model can also be represented using log-odds, i.e.,

$$\log \left[\frac{\pi_{i1\cdot}}{1 - \pi_{i1\cdot}} \right] = 0.5\delta_{i0\cdot} - 0.5\delta_{i1\cdot},$$

$$\log \left[\frac{\pi_{i\cdot 1}}{1 - \pi_{i\cdot 1}} \right] = 0.5\delta_{i\cdot 0} - 0.5\delta_{i\cdot 1}.$$

In the BIPC model, each binary response variable is positioned on one and only dimension, and thus the class coordinates are specified as follows:

$$\gamma_{l\cdot} = \begin{bmatrix} \gamma_{0\cdot 1} & 0 \\ \gamma_{1\cdot 1} & 0 \end{bmatrix} \text{ and } \gamma_{\cdot l} = \begin{bmatrix} 0 & \gamma_{\cdot 02} \\ 0 & \gamma_{\cdot 12} \end{bmatrix},$$

where the first binary response (Y_{i1}) is positioned on the first dimension and the second binary response variable (Y_{i2}) on the second dimension.

Suppose x_{i1} represents one of the predictor variables. Let us now simplify the log-odds

model by replacing the above class points. That is,

$$\begin{aligned}
 \log \left[\frac{\pi_{i1\cdot}}{1 - \pi_{i1\cdot}} \right] &= \sum_{m=1}^2 (\eta_{im} \gamma_{1\cdot m}) - \sum_{m=1}^2 (\eta_{im} \gamma_{0\cdot m}) + 0.5 \sum_{m=1}^2 \gamma_{0\cdot m}^2 - 0.5 \sum_{m=1}^2 \gamma_{1\cdot m}^2 \\
 &= \eta_{i1}(\gamma_{1\cdot 1} - \gamma_{0\cdot 1}) + 0.5 \times (\gamma_{0\cdot 1}^2 - \gamma_{1\cdot 1}^2) \\
 &= (\beta_{01} + \beta_{11} x_{i1})(\gamma_{1\cdot 1} - \gamma_{0\cdot 1}) + 0.5 \times (\gamma_{0\cdot 1}^2 - \gamma_{1\cdot 1}^2) \\
 &= \beta_{01}(\gamma_{1\cdot 1} - \gamma_{0\cdot 1}) + \beta_{11}(\gamma_{1\cdot 1} - \gamma_{0\cdot 1})x_{i1} + 0.5 \times (\gamma_{0\cdot 1}^2 - \gamma_{1\cdot 1}^2) \\
 &= \beta_{01}^* + \beta_{11}^* x_{i1},
 \end{aligned}$$

where $\beta_{01}^* = \beta_{01}(\gamma_{1\cdot 1} - \gamma_{0\cdot 1}) + 0.5 \times (\gamma_{0\cdot 1}^2 - \gamma_{1\cdot 1}^2)$ and $\beta_{11}^* = \beta_{11}(\gamma_{1\cdot 1} - \gamma_{0\cdot 1})$. So, for a unit increase in x_{i1} the log-odds in the BIPC model changes by $\beta_{11}(\gamma_{1\cdot 1} - \gamma_{0\cdot 1})$. Similarly, the simplified log-odds form of the second binary response variable becomes $\log[\pi_{i\cdot 1}/(1 - \pi_{i\cdot 1})] = \beta_{02}(\gamma_{\cdot 12} - \gamma_{\cdot 02}) + \beta_{12}(\gamma_{\cdot 12} - \gamma_{\cdot 02})x_{i1} + 0.5 \times (\gamma_{\cdot 02}^2 - \gamma_{\cdot 12}^2) = \beta_{02}^* + \beta_{12}^* x_{i1}$.

At this stage, the BIPC model is not identified since both the regression weights and the class coordinates influence the distance model parameters. For unique identification of model parameters, we must impose restrictions on the class coordinates. This can be achieved, for example, by setting a unit difference between the class coordinates. That is, $\gamma_{1\cdot 1} = 1$ and $\gamma_{0\cdot 1} = 0$ for the first response variable, Y_{i1} ; and $\gamma_{\cdot 12} = 1$ and $\gamma_{\cdot 02} = 0$ for Y_{i2} . Thus, $\beta_{1s}^* = \beta_{1s}$ and $\beta_{0s}^* = \beta_{0s} - 0.5$, where $s = 1, 2$.

Appendix C

Chapter 4: Exploring latent variable models with few number of indicators per factor

Recently, in clinical psychological research factor analysis and structural equation models have been proposed for the analysis of comorbidity of depressive and anxiety disorders (Krueger, 1999; Beesdo-Baum et al., 2009). A typical characteristic of these models is that the indicators are dichotomous, i.e. whether someone has or does not have a particular disorder, and that there are only a few indicators per latent variable, i.e. 2 or 3. These authors found for depressive and anxiety disorders two underlying factors: 'fear' with indicators social phobia (SP), and panic disorder (PD), and 'distress' with indicators major depressive disorder (MDD), dysthymia (DYST), and generalized anxiety disorder (GAD).

We tried to replicate these findings using data from the Netherlands Study for Depression and Anxiety (NESDA, Penninx et al., 2008). In this study, measurements were obtained on N=2,938 subjects of age between 18 – 65 years with an average age of 42 (S.D. = 13.1), 66.5% were female, and the average number of years of education attained was 12.2 (S.D. = 3.3). In the NESDA study measurements are obtained on each of the five disorders, i.e., about 37.1% of the subjects in the study had MDD, 10.2% had DYST,

Table C.1: Fit statistics for the factor models fitted on the NESDA data.

Model	# parameters	RMSEA	CFI
1-factor	10	0.082	0.956
2-factor (fear-distress)	11	0.034	0.994
2-factor (anxiety-depression)	11	0.071	0.974

15.3% had GAD, 22.4% had SP, and 28.6% had PD.

With these data it is possible to investigate the interplay between personality traits and co-morbidity of mental disorders. We considered Big-Five personality traits, i.e., Neuroticism (N), Extraversion (E), Openness to experience (O), Agreeableness (A), and Conscientiousness (C). After establishing the measurement model we would like to see the influence of the personality traits and the background variables (age, gender, and education) in explaining the latent variables.

The first step in the analysis was to establish the measurement model using a Confirmatory Factor Analysis (CFA). Besides the fear-distress theory there are two other competing theories (Krueger, 1999; Beesdo-Baum et al., 2009; Spinhoven et al., 2013): the first is the anxiety-depression theory which differs from the distress-fear with respect to generalized anxiety disorder. The anxiety-depression theory places GAD on the other latent factor. Finally, there is a single factor theory where all five disorders are indicators of a single underlying factor.

In Table C.1 we present fit statistics of the three measurement models applied to the NESDA data. The 2-factor model with fear and distress factors fitted the NESDA data best with fit statistics of RMSEA=0.034 and CFI=0.994. In this selected model, there are a total of eleven (11) parameters: three (3) loadings on the first factor, two (2) loadings on the second factor, a covariance between the factors, and a threshold for each of the five manifest variables.

Table C.2: Parameter estimates with the corresponding standard errors (S.E.) presented in parenthesis for the final 2-factor (fear-distress) model.

Effect	Parameter	Estimate (S.E.)
Fear (F1)		
Gender	γ_{11}	-0.127(0.072)
Education [†]	γ_{21}	-0.113(0.035)
Age	γ_{31}	0.012(0.036)
Neuroticism [†]	γ_{41}	0.985(0.062)
Extraversion [†]	γ_{51}	-0.289(0.045)
Agreeableness	γ_{61}	-0.026(0.035)
Conscientiousness	γ_{71}	-0.039(0.038)
Openness	γ_{81}	-0.014(0.035)
Distress (F2)		
Gender	γ_{12}	0.003(0.079)
Education [†]	γ_{22}	-0.166(0.039)
Age	γ_{32}	-0.020(0.038)
Neuroticism [†]	γ_{42}	0.954(0.069)
Extraversion [†]	γ_{52}	-0.245(0.047)
Agreeableness	γ_{62}	0.002(0.037)
Conscientiousness [†]	γ_{72}	0.112(0.041)
Openness	γ_{82}	0.002(0.038)
Covariance for Factors		
Var(F1)	ψ_{11}	1.000(—)
Var(F2)	ψ_{22}	1.000(—)
Cov(F1, F2)	ψ_{12}	0.333(0.053)
Threshold		
MDD	τ_1	0.417(0.048)
DYST	τ_2	1.533(0.070)
GAD	τ_3	1.181(0.058)
SP	τ_4	0.943(0.053)
PD	τ_5	0.696(0.047)

[†] statistically significant effect, i.e., $p < 0.05$.

In the second step, researchers would like to see the effect of the personality variables on the two latent factors. Suppose they would fit a Multiple Indicators and Multiple Causes (MIMIC) model for this purpose. Table C.2 shows the results. From the personality traits, only neuroticism and extraversion had statistically significant effects on both the fear and distress factor. The positive association between neuroticism and both dimensions implies that on average a subject having a higher score on neuroticism would score high on both fear and distress. In the case of extraversion, the association was negative implying a subject with lower score on extraversion would score higher on both fear and distress. Conscientiousness had an effect only on distress. From the background variables, only education had a strong negative association with both factors.

So far, the theory about comorbidity and the influence of personality variables on the latent factors replicates earlier findings. However, when computing and plotting the factor scores some notable results were found. The distribution of the factor scores from the factor analysis are shown in the left column of Figure C.1, and it can be seen that the distribution deviates strongly from normality. It is clear that the assumption of a normally distributed latent variable does not hold. Furthermore, in the right hand side the factor scores are shown for the two latent variables in the MIMIC model. Surprisingly, the distribution of these factor scores seems completely different from the distribution found in the factor analysis.

These surprising results cause doubts about the robustness of latent variable models with just a few dichotomous indicators per factor. However, we did not find large scale simulation studies for such models. The aim of Chapter 2 is to fill this gap. In Chapter 4 a Multivariate Logistic Distance (MLD) model is proposed to analyze multivariate binary data.

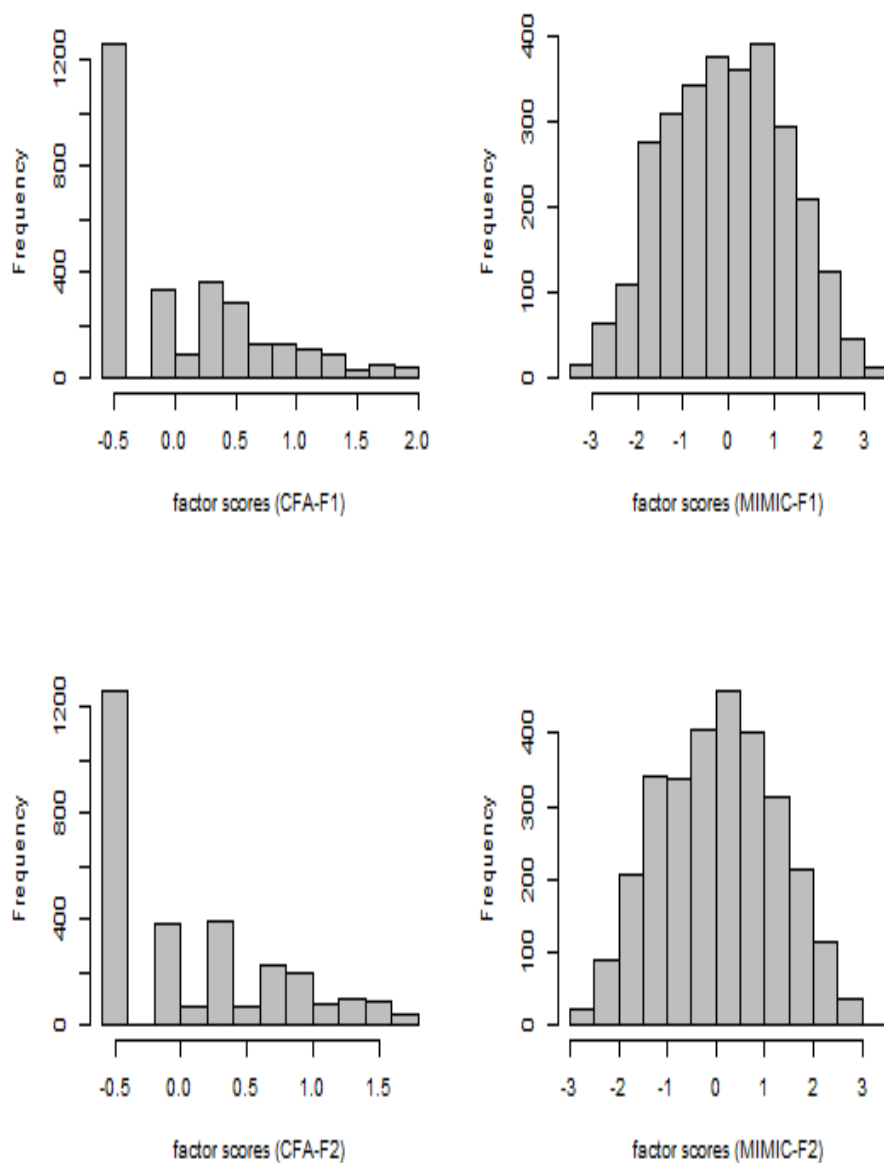


Figure C.1: The distribution of estimated factor scores obtained from the final 2-factor (fear-distress) model. The top panel representing the distribution of scores from the first factor (F1) before and after the inclusion of external variables, respectively; and, the bottom panel for those scores from the second factor (F2) before and after the inclusion of the external variables, respectively.

Appendix D

Chapter 4: Simplification of log-odds representation of the MLD model, Eq. 4.8

Let us simplify log-odds representation of the MLD model (Eq. 4.8) using distress-fear model, where the first dimension (distress) was represented by Major Depressive Disorder (MDD), Generalized Anxiety Disorder (GAD), and Dysthymia (DYST); and the second dimension (fear) was represented by Panic Disorder (PD) and Social Phobia (SP) (Spinhoven et al., 2013). Suppose class coordinates of GAD on the first dimension and PD on the second dimension are set to fixed for identification reasons. Thus, class coordinates of the other response variables become

$$\gamma_2^{\text{GAD}} = \begin{bmatrix} \gamma_{02,1} & 0 \\ \gamma_{12,1} & 0 \end{bmatrix}, \quad \gamma_3^{\text{DYST}} = \begin{bmatrix} \gamma_{03,1} & 0 \\ \gamma_{13,1} & 0 \end{bmatrix}, \quad \gamma_5^{\text{SP}} = \begin{bmatrix} 0 & \gamma_{05,2} \\ 0 & \gamma_{15,2} \end{bmatrix}, \quad (\text{D.1})$$

For demonstration purpose let us focus on GAD response variable of the first dimension and SP of the second dimension, and derive their corresponding simplified version of Eq. 4.8.

The log-odds representation of the multivariate distance model for GAD becomes,

$$\begin{aligned}
\log \left[\frac{\pi_2(\mathbf{x}_i)}{1 - \pi_2(\mathbf{x}_i)} \right] &= \sum_{m=1}^M \left\{ \beta_{0m}(\gamma_{12,m} - \gamma_{02,m}) + 0.5(\gamma_{02,m}^2 - \gamma_{12,m}^2) \right. \\
&\quad \left. + \mathbf{x}_i^T \boldsymbol{\beta}_m(\gamma_{12,m} - \gamma_{02,m}) \right\} \\
&= \left\{ \left\{ \beta_{01}(\gamma_{12,1} - \gamma_{02,1}) + 0.5(\gamma_{02,1}^2 - \gamma_{12,1}^2) \right. \right. \\
&\quad \left. + \mathbf{x}_i^T \boldsymbol{\beta}_1(\gamma_{12,1} - \gamma_{02,1}) \right\} \\
&\quad + \left\{ \beta_{02}(\gamma_{12,2} - \gamma_{02,2}) + 0.5(\gamma_{02,2}^2 - \gamma_{12,2}^2) \right. \\
&\quad \left. + \mathbf{x}_i^T \boldsymbol{\beta}_2(\gamma_{12,2} - \gamma_{02,2}) \right\} \Big\} \\
&= \beta_{01}(\gamma_{12,1} - \gamma_{02,1}) + 0.5(\gamma_{02,1}^2 - \gamma_{12,1}^2) \\
&\quad + \mathbf{x}_i^T \boldsymbol{\beta}_1(\gamma_{12,1} - \gamma_{02,1}).
\end{aligned}$$

Similarly, the log-odds representation of the multivariate distance model for SP becomes

$$\begin{aligned}
\log \left[\frac{\pi_5(\mathbf{x}_i)}{1 - \pi_5(\mathbf{x}_i)} \right] &= \sum_{m=1}^M \left\{ \beta_{0m}(\gamma_{15,m} - \gamma_{05,m}) + 0.5(\gamma_{05,m}^2 - \gamma_{15,m}^2) \right. \\
&\quad \left. + \mathbf{x}_i^T \boldsymbol{\beta}_m(\gamma_{15,m} - \gamma_{05,m}) \right\} \\
&= \left\{ \left\{ \beta_{01}(\gamma_{15,1} - \gamma_{05,1}) + 0.5(\gamma_{05,1}^2 - \gamma_{15,1}^2) \right. \right. \\
&\quad \left. + \mathbf{x}_i^T \boldsymbol{\beta}_1(\gamma_{15,1} - \gamma_{05,1}) \right\} \\
&\quad + \left\{ \beta_{02}(\gamma_{15,2} - \gamma_{05,2}) + 0.5(\gamma_{05,2}^2 - \gamma_{15,2}^2) \right. \\
&\quad \left. + \mathbf{x}_i^T \boldsymbol{\beta}_2(\gamma_{15,2} - \gamma_{05,2}) \right\} \Big\} \\
&= \beta_{02}(\gamma_{15,2} - \gamma_{05,2}) + 0.5(\gamma_{05,2}^2 - \gamma_{15,2}^2) \\
&\quad + \mathbf{x}_i^T \boldsymbol{\beta}_2(\gamma_{15,2} - \gamma_{05,2}).
\end{aligned}$$

Bibliography

- Acito, F., & Anderson, R. (1986). A simulation study of factor score indeterminacy. *Journal of Marketing Research*, 23, 111-118.
- Agresti, A. (2002). *Categorical data analysis* (Second ed.). New York: John Wiley and Sons.
- Agresti, A. (2007). *An introduction to categorical data analysis* (Second ed.). New Jersey: John Wiley and Sons.
- Agresti, A. (2013). *Categorical data analysis* (Third ed.). New York: John Wiley and Sons.
- Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119-126.
- Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, 29, 813-828.
- Aitchison, J., & Silvey, S. D. (1960). Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society, Series B*, 22, 154-171.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the second international symposium on information theory* (p. 267-281). Budapest: Akademiai Kiado.
- Anderson, J. A. (1984). Regression and order categorical variables. *Journal of Royal Statistics Society B*, 46(1), 1-30.

- Anderson, J. C., & Gerbing, D. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49, 155-173.
- Asar, Ö., & Ilk, Ö. (2013). **mmm**: An R package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine*, 112, 649-654.
- Ashford, J. R., Morgan, D. C., Rae, S., & Sowden, R. R. (1970). Respiratory symptoms in british coal miners. *The American Review of Respiratory Disease*, 102(3), 370-81.
- Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.), *Studies in item analysis and prediction* (p. 158-176). Stanford, California: Stanford University Press.
- Bartholomew, D., & Knott, M. (1999). *Latent variable models and factor analysis*. (2nd ed.). London: Arnold.
- Bartholomew, D., Steele, F., Moustaki, I., & Galbraith, J. (2002). *The analysis and interpretation of multivariate data for social scientists*. London: Chapman & Hall.
- Bartolucci, F., Colombi, R., & Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, 17, 691-711.
- Bartolucci, F., Forcina, A., & Dardanoni, V. (2001). Positive quadrant dependence and marginal modelling in two-way tables with ordered margins. *Journal of the American Statistical Association*, 96(456), 1497-1505.
- Beesdo-Baum, K., Höfler, M., Gloster, A. T., Klotsche, J., Lieb, R., Beauducel, A., ... Wittchen, H. U. (2009). The structure of common mental disorders: a replication study in a community sample of adolescents and young adults. *International Journal of Methods in Psychiatric Research*, 18, 204-220.
- Bergsma, W. P. (1997). *Marginal models for categorical variables* (Unpublished doctoral dissertation). Tilburg University.

- Bergsma, W. P., Croon, M. A., & Hagenaars, J. A. (2009). *Marginal models for dependent, clustered and longitudinal categorical data*. New York: Springer.
- Bergsma, W. P., & Rudas, T. (2002). Marginal models for categorical data. *The Annals of Statistics*, 30, 140-159.
- Bhuyan, M. J., Islam, M. A., & Rahman, M. S. (2018). A bivariate bernoulli model for analyzing malnutrition data. *Health Services and Outcomes Research Methodology*, 1-19. (<https://doi.org/10.1007/s10742-018-0180-9>)
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.
- Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12, 3-27.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review Psychology*, 53, 605-634.
- Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality*. Amsterdam: Sociometric Research Foundation. (Doctoral dissertation, University of Groningen, The Netherlands)
- Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, 50(2), 229-242.
- Boomsma, A. (2013). Reporting monte carlo studies in structural equation modeling. *Structural Equation Modeling*, 20, 518-540.
- Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future*. Lincolnwood: Scientific Software International.
- Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (Second ed.). New York: Springer.
- Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research*

- and Perspectives*, 6(1-2), 25-53.
- Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101-133.
- Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Bull, S. B. (1998). Regression models for multiple outcomes in large epidemiological studies. *Statistics in Medicine*, 17, 2179-2197.
- Carey, V., Zeger, S. L., & Diggle, P. (1993). Modeling multivariate binary data with alternating logistic regressions. *Biometrika*, 80, 517-526.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, 29(4), 468-508.
- Cheng, G., Yu, Z., & Huang, J. Z. (2013). The cluster bootstrap consistency in generalized estimating equations. *Journal of Multivariate Analysis*, 115, 33-47.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5-32.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). New York: Academic Press.
- Colombi, R., & Forcina, A. (2001). Marginal regression models for the analysis positive association of ordinal response variables. *Biometrika*, 88(4), 1007-1019.
- Coombs, C. H. (1976). *A Theory of Data*. Michigan: Mathesis Press.
- Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PR-I) and NEO five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cox, D. R. (1972). The analysis of multivariate binary data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21(2), 113-120.
- Davison, M. L. (1983). *Multidimensional scaling*. USA: John Wiley & Sons, Inc.
- De Leeuw, J. (2005). Multidimensional unfolding. In B. S. Everitt & D. C. Howelll (Eds.),

- Encyclopedia of statistics in behavioral science* (p. 1289-1294). Wiley.
- De Leeuw, J. (2006). *Pseudo-voronoi diagrams for multcategory exponential representations*. UCLA: Statistics Preprints # 463.
- De Rooij, M. (2009a). Ideal point discriminant analysis with a special emphasis on visualization. *Psychometrika*, 74, 317-330.
- De Rooij, M. (2009b). Trend vector models for the analysis of change in continuous time for multiple groups. *Computational statistical data analysis*, 53, 3209-3216.
- De Rooij, M. (2011). Transitional ideal point models for longitudinal multinomial outcomes. *Statistical Modelling*, 11(2), 115-135.
- De Rooij, M., & Heiser, W. J. (2005, March). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika*, 70(1), 99-122.
- De Rooij, M., & Schouteden, M. (2009). *Mixed effect ideal point models for longitudinal multinomial data*. (Submitted paper)
- De Rooij, M., & Worku, H. M. (2012). A warning concerning the estimation of multinomial logistic models with correlated responses in SAS. *Computer Methods and Programs in Biomedicine*, 107(2), 341-346.
- Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, 2(2), 119-143.
- Elffers, H., Bethlehem, J., & Gill, R. (1978). Indeterminacy problems and the interpretation of factor analysis results. *Statistica Neerlandica*, 32, 181-199.
- Elliot, D. S., Huizinga, D., & Menard, S. (1989). *Multiple problem youth: delinquency, substance use, and mental health problems*. New York: Springer-Verlag.
- Ferguson, C. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80,

27-38.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis*. London: Chapman and Hall.

Gabriel, K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, 58, 453-467.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: John Wiley & Sons.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both words: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10(4), 544-565.

Glonek, G. F. V. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, 83(1), 15-28.

Glonek, G. F. V., & McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, 57, 533-546.

Gower, J. C., & Hand, D. J. (1996). *Biplots*. London: Chapman and Hall.

Gower, J. C., Lubbe, S., & Le Roux, N. (2011). *Understanding biplots*. Chichester: John Wiley & Sons Ltd.

Green, B. F. (1976). On the factor score controversy. *Psychometrika*, 41(2), 263-266.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6(4), 430-450.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common factor theory. *British Journal of Statistical Psychology*, 8(65).

Halekoh, U., Hojsgaard, S., & Yan, J. (2006). The R package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15(2).

Hallquist, M. (2012). **MplusAutomation**: Automating **Mplus** model estimation and interpretation [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=MplusAutomation> (R package version 0.5-1)

- Heermann, E. F. (1964). The geometry of factorial indeterminacy. *Psychometrika*, 29(4).
- Heermann, E. F. (1966). The algebra of factorial indeterminacy. *Psychometrika*, 31(4).
- Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21, 2409-2419.
- Heiser, W. J. (1981). *Unfolding analysis of proximity data* (Unpublished doctoral dissertation). Leiden University.
- Heiser, W. J. (1987). Joint ordination of species and sites: The unfolding technique. In P. Legendre & L. Legendre (Eds.), *Developments in numerical ecology* (p. 189-221). Springer Verlag.
- Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., ... Satariano, W. A. (2010). To GEE or Not to GEE: Comparing population averaged and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, 21(4), 467-474.
- IBM SPSS. (2012). *IBM SPSS statistics version 21*. Boston, Massachusetts: International Business Machines Corp.
- Jöreskog, J. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631-639.
- Jöreskog, J. G., & Sörbom, D. (1981). *LISREL: Analysis of linear structural relationships by the method of maximum likelihood (version v)*. Chicago: National Educational Resources.
- Kenneth, A. (1989). *Structural equations with latent variables*. Canada: John Wiley & Sons, Inc.
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10(3), 333-351.
- Kirk, D. B. (1973). On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika*, 38(2), 259-268.

- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26(3), 457-477.
- Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, 56, 921-926.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Sage publications.
- Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *The Annals of Statistics*, 24(2), 726-752.
- Lang, J. B., & Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89, 625-632.
- Lawley, D. (1944). The factor analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh*, 62-A, 74-82.
- Ledermann, W. (1938). The orthogonal transformation of a factorial matrix into itself. *Psychometrika*, 3(181).
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Liang, K. Y., Zeger, S. L., & Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 54(1), 3-40.
- Lipsitz, S. R., Kim, K., & Zhao, L. P. (1994). Analysis of repeated categorical data using Generalized Estimating Equations. *Statistics in Medicine*, 14, 1149-1163.
- Lipsitz, S. R., Laird, N. M., & Harrington, D. P. (1990). Maximum likelihood regression methods for paired binary data. *Statistics in medicine*, 9, 1517-1525.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- MacCallum, R., & Austin, J. (2000). Applications of Structural Equation Modeling in

- psychological research. *Annual Review Psychology*, 51, 201-226.
- MacLean, R. R., Sofuoglu, M., & Rosenheck, R. (2018). Tobacco and alcohol use disorders: Evaluating multimorbidity. *Addictive Behaviors*, 78, 59-66.
- Mardia, K. V. (1967). Some contributions to the contingency-type bivariate distributions. *Biometrika*, 54, 235-249.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103(3), 391-410.
- Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? the number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181-220.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.
- Mislevy, R. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, 11(1), 3-31.
- Molenberghs, G., & Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association*, 89, 633-644.
- Molenberghs, G., & Lesaffre, E. (1999). Marginal modeling of multivariate categorical data. *Statistics in Medicine*, 18, 2237-2255.
- Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York: Springer.
- Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43(4), 551-560.
- Muthen, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22, 43-65.
- Muthen, B. (1984). A general structural equation model with dichotomous, ordered

- categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115-132.
- Muthen, L., & Muthen, B. (1998-2012). *Mplus user's guide*. (3rd ed.). Los Angeles, CA: Muthen & Muthen.
- Ottenbacher, K. (1991). Interpretation of interaction in factorial analysis of variance design. *Statistics in Medicine*, 10, 1565-1571.
- Palmgren, J. (1989). *Regression Models for Bivariate Binary Responses* (Working Paper No. 101). Seattle, WA: University of Washington.
- Pan, W. (2001). Akaike's information criterion in Generalized Estimating Equations. *Biometrics*, 57, 120-125.
- Park, T. (1994). Multivariate regression models for discrete and continuous repeated measurements. *Communications in Statistics - Theory and Methods*, 23, 1547-1564.
- Paxton, P., Curran, P., Bollen, K., Kirby, J., & Fen, C. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 35(8), 287-312.
- Pearson, K. (1900). On the correlation of characters not quantitatively measurable. *Royal Society of Philosophical Transactions, Series A*, 195, 1-47.
- Penninx, B. W., Beekman, A. T., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., ... Van Dyck, R. (2008). The Netherlands study of depression and anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, 17, 121-140.
- Plewis, I. (1996). Statistical methods for understanding cognitive growth: A review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology*, 49, 25-42.
- R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org> (ISBN 3-900051-07-0)

- R Development Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org>
- Rindskopf, D. (1984). Structural equation models: Empirical identification, heywood cases, and related problems. *Sociological Methods and Research*, 13, 109-119.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1-25.
- Robinson, P. M. (1974). Identification, estimation and large-sample theory for regressions containing unobservable variables. *International Economic Review*, 15(3), 680-692.
- SAS Institute Inc. (2011). SAS/STAT software, version 9.3 [Computer software manual]. Cary, NC. (<http://www.sas.com>)
- SAS Institute Inc. (2013). *Base SAS 9.4 procedures guide*. Cary, NC: SAS Institute Inc.
- Schonemann, P. H. (1971). The minimum average correlation between equivalent sets of uncorrelated factors. *Psychometrika*, 36(21).
- Schonemann, P. H., & Wong, M. M. (1972). Some new results on factor indeterminacy. *Psychometrika*, 37(61).
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.
- Sherman, M., & le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation*, 26, 901-925.
- Sherman, M., & Le Cessie, S. (1997). A comparison between bootstrap methods and Generalized Estimating Equations for correlated outcomes in Generalized Linear Models. *Communications in Statistics - Simulation and Computation*, 26, 901-925.
- Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, 35, 137-167.

- Skron dal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multi-level, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall.
- Sommer, A., Katz, J., & Tarwotjo, I. (1984). Increased risk of respiratory disease and diarrhea in children with preexisting mild vitamin A deficiency. *American Society for Clinical Nutrition*, 40, 1090-1095.
- Spinhoven, P., De Rooij, M., Heiser, W., Penninx, B. W. J. H., & Smit, J. (2009). The role of personality in comorbidity among anxiety and depressive disorders in primary care and specialty care: a cross-sectional analysis. *General Hospital Psychiatry*, 31, 470-477.
- Spinhoven, P., Penelo, E., De Rooij, M., Penninx, B. W., & Ormel, J. (2013). Reciprocal effects of stable and temporary components of neuroticism and affective disorders: results of a longitudinal cohort study. *Psychological Medicine*, 44, 337-348.
- Stapleton, D. C. (1978). Analyzing political participation data with a MIMIC model. *Sociological Methodology*, 9, 52-74.
- Takane, Y. (1998). Visualization in ideal point discriminant analysis. In J. Blasius & M. J. Greenacre (Eds.), *Visualization of categorical data* (p. 441-459). New York: Academic press.
- Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika*, 52, 371-392.
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393-408.
- Ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5), 1167-1179.
- Ter Braak, C. J. F., & Verdonschot, P. F. M. (1995). Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, 57(3), 1015-1621.

- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (p. 73-140). Mahwah, NJ: Erlbaum.
- Thurstone, L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
- Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, 1, 31-65.
- Van der Heijden, P. G. M., Mooijaart, A., & Takane, Y. (1994). Correspondence analysis and contingency models. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence analysis in the social sciences* (p. 79-111). New York: Academic Press.
- Vermunt, J. K., Rodrigo, M. F., & Ato-Garcia, M. (2001). Modeling joint and marginal distributions in the analysis of categorical panel data. *Sociological Methods and Research*, 30, 170-196.
- Von Oertzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology*, 63, 627-646.
- Wei, L., & Stram, D. (1988). Analyzing repeated measurements with possibly missing observations by modeling marginal distributions. *Statistics in Medicine*, 7, 139-148.
- Wei, X. (2012). %PROC.R: A SAS macro that enables native R programming in the base SAS environment. *Journal of Statistical Software*, 46.
- Worku, H. M., & De Rooij, M. (2017a). Properties of ideal point classification models for bivariate binary data. *Psychometrika*, 82(2), 308-328.
- Worku, H. M., & De Rooij, M. (2017b). **mldm**: An R package for analyzing multiple binary responses using a multivariate logistic distance model. (Software Manual)
- Worku, H. M., & De Rooij, M. (2018). A multivariate logistic distance model for the analysis of multiple binary responses. *Journal of Classification*, 35, 1-23. (<https://doi.org/10.1007/s00357-018-9251-4>)
- Yee, T. W. (2010). The **VGAM** package for categorical data analysis. *Journal of*

- Statistical Software*, 32(10), 1-34. Retrieved from <http://www.jstatsoft.org/v32/i10/>
- Yu, H. T., & De Rooij, M. (2013). Model selection for trend vector model. *Journal of Classification*, 30, 338-369.
- Zhao, L. P., & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77, 642-648.
- Ziegler, A. (2011). *Generalized estimating equations*. New York: Springer.
- Ziegler, A., & Arminger, G. (1995). Analyzing the employment status with panel data from GSOEP - a comparison of the MECOSA and the GEE1 approach for marginal models. *Vierteljahreshefte zur Wirtschaftsforschung*, 64, 72-80.
- Ziegler, A., Kastner, C., & Blettner, M. (1998). The Generalized Estimating Equations: An Annotated Bibliography. *Biometrical Journal*, 40(2), 115-139.

Samenvatting

In dit proefschrift beschrijven we recent ontwikkelde statistische tools voor het analyseren van multivariate binaire data. Multivariate binaire data, gedefinieerd als verzamelde gegevens van meerdere binaire afhankelijke variabelen en één of meer onafhankelijke variabelen, komen voor in allerlei onderzoeksdisciplines. Neem bijvoorbeeld de Indonesische Kinderen Studie (ICS). In deze studie is er data verzameld van meer dan drieduizend kinderen die medisch onderzocht zijn op luchtweginfectie, diarree-infectie, en xeroftalmie. Het doel van de ICS was om te achterhalen of kinderen met een deficiëntie in Vitamine A een verhoogd risico lopen op luchtweg- en diarree-infectie.

Een ander voorbeeld waarbij multivariate binaire wordt gebruikt is de Nederlandse Studie naar Depressie en Angst (NESDA). De gegevens die door NESDA verzamelt worden dienen ten doel om de interactie tussen persoonlijkheidseigenschappen enerzijds en de comorbiditeit van depressie- en angststoornissen anderzijds te kunnen onderzoeken. In dit onderzoeksgebied van psychologische stoornissen zijn psychologen en epidemiologen veelal genteresseerd in comorbiditeit en hoe comorbiditeit gerelateerd kan worden aan risicofactoren zoals persoonlijkheidseigenschappen en achtergrondkenmerken.

Er zijn talloze statistische methoden beschikbaar voor het analyseren van multivariate continue afhankelijke variabelen doordat er goed gebruik gemaakt kan worden van de multivariate normale kansverdeling. De multivariate regressie en de multivariate variantie analyse (MANOVA), om er maar een paar te noemen, behoren tot de populaire statistische methoden die hier worden toegepast. Echter, voor de multivariate categorische data is

het aanbod van methoden en technieken gering. De huidige beschikbare methoden en technieken bouwen voort op assumpties die niet gecontroleerd kunnen worden (zoals het bestaan van de latente variabelen in latent variable models en structural equation models), of komen met vereisten dat de onafhankelijke variabelen gecategoriseerd dienen te worden (zoals de GEE2 methode voor marginale modellen). Met behulp van een Monte Carlo simulatie studie laten in hoofdstuk 2 we zien dat het toepassen van een latente variable model op multivariate binaire data tot gebrekkige resultaten leidt wanneer er slechts twee of drie indicatoren per latente variabele zijn.

In dit proefschrift presenteren we een aangepaste versie van het ideal point classification (IPC) model waarmee multivariate binaire gegevens geanalyseerd kunnen worden. Het IPC model is een probabilistisch multidimensional “unfolding” model en veel lijkend Ideal Point Discriminant Analysis (IPDA). Hoofdstuk 3 begint eerst met een studie van de eigenschappen van het IPC model voor het analyseren van bivariate binaire gegevens. Door gebruik te maken van een kader gebaseerd op de bivariate logistische regressie, kunnen de afhankelijke variabelen worden gerepresenteerd in een drie-dimensionale Euclidische ruimte. In deze drie-dimensionale ruimte heeft de eerste dimensie betrekking op de prevalentie van de eerste afhankelijke variabele; de tweede heeft betrekking op de prevalentie van de tweede variabele; en, de derde dimensie heeft betrekking op de samenhang tussen de twee afhankelijke variabelen. Op basis van een simulatie studie kunnen we aantonen dat met het IPC model het niet volledig mogelijk is om de daadwerkelijke parameters van de binaire data te achterhalen, dat wil zeggen, de twee marginale prevalentie parameters en de parameter voor de associatie tussen de twee afhankelijke variabelen. In hoofdstuk 3 laten we vervolgens zien dat met een re-parameterisatie van het IPC model het wel mogelijk is om deze parameters terug te vinden. Dit aangepaste model noemen we het Bivariate IPC (BIPC) model.

Een beperking van het Bivariate IPC model is dat het niet toegankelijk is om uit te breiden naar multivariate binaire gegevens (meer dan twee binaire afhankelijke variabe-

len). Door deze beperking van het BIPC model, wordt in hoofdstuk 4 voorgesteld om het Multivariate Logistische Afstands (MLD) model te gebruiken voor het analyseren van multivariate binaire data. Het MLD model is een vereniging van twee soorten domeinen van statistische methoden: het domein van de Multidimensional Scaling (MDS) en het domein van het Generalized Linear Model (GLM). Het MLD-model kan tegelijkertijd gebruikt worden voor zowel het beoordelen van de dimensionale structuur van de data als het schatten van het effect van de onafhankelijke variabelen op de afhankelijke variabelen. Zo biedt het MLD-model de mogelijkheid om op NESDA data tegelijkertijd de dimensionale structuur van psychologische stoornissen te onderzoeken als het effect van persoonlijkheidseigenschappen en achtergrondkenmerken op de prevalentie van psychologische stoornissen.

Voor ondersteuning van interpretatie doeleinden lenen de resultaten de MLD analyse zich goed voor de grafische weergave in een biplot. Een ander voordeel van het MLD-model ten opzichte van marginale modellen is dat MLD-model toegepast kan worden in combinatie met dimensie reductie, waarmee de complexiteit van het standaard multivariate GLM wordt vereenvoudigd door minder parameters te hoeven schatten. Met deze dimensie-reductie methode wordt de deur geopend naar verder onderzoek.

Wanneer de afstanden tussen de twee categorieën op elke afhankelijke variabele eenzelfde waarde krijgen toegewezen, dan kan het MLD-model geschat worden door gebruik te maken van de GEE methode. Onder deze restrictie van 'gelijke afstanden' is het dan ook mogelijk om het MLD-model te schatten met behulp van bestaande statische software pakketten zoals de **genmod** procedure in SAS, of het **geepack**-pakket in R. Wanneer er geen gebruik wordt gemaakt van de gelijke afstanden restrictie, dan is het MLD-model een op zichzelf staand marginaal model. In hoofdstuk 5 presenteren we het **mldm**-pakket dat is ontwikkeld in R om het MLD-model op data te kunnen toepassen. De belangrijkste functie in dit pakket is `mldm.fit`, hiermee kunnen we het MLD-model schatten. Vervolgens kan met de functie `mldm.fit` het geschatte model grafisch worden weergegeven in een biplot. De functie `mldm.fit` heeft ook als output een object genaamd `QIC`, met dit

object kunnen verschillende kandidaat-modellen worden vergeleken. Het **mldm**-pakket is publiek toegankelijk en beschikbaar op het online database-systeem GitHub, te vinden via het URL adres: <https://github.com/workuhm1/mldm-package-github>.

Ten slotte raden we onderzoekers aan om voorzichtig te zijn met het toepassen van latent variable models of structural equation models op multivariate binaire gegevens. De prestatie van statistische methoden gebaseerd op deze modellen is ondermaats met slechts enkele indicatoren per latente variabele (d.w.z. 2 of 3). Een alternatief statisch model dat minder assumpties vereist is mogelijk beter toepasbaar, bijvoorbeeld het multivariaat logistische afstands model.

Acknowledgments

It would not be possible for me to successfully complete this dissertation without the support of many people. Let me use this opportunity to say something about you.

Firstly, I would like to thank my supervisors Mark de Rooij and Willem J. Heiser. Mark, I was lucky to get the chance working with you. As I came from a Biostatistics background, it took me sometime to understand the beauty of the other continent of science which is Psychometrics. You were there for me whenever I needed it. Thanks also for having confidence in me. Willem, it is a great honor for me to work with you. I would like to thank you for your contribution to the simulation paper. I wish you a long life and hope to see you somewhere in our scientific endeavors.

Secondly, I would like to thank members of my doctorate committee for their valuable feedback on the dissertation. I would like to thank all of my colleagues at Methodology and Statistics department in Leiden university. Special thanks to Cor Ninaber, Marije Fagginger Auer, Marian Hickendorff, Kees Verduin, Jacqueline Hartman-Dries, Bouk de Water, and Monique van der Geest. Guys, you made my time in Leiden fun and exciting. Thanks to my PhD paranymphs, Maarten Kampert and Yinebeb Tessema - I owe you one guys. I would also like to thank my colleagues at my current job, OCS Consulting BV. Special thanks to my line-managers Yasemin Atil and Jules van der Zalm, and my colleagues Evian Fernandez Garcia, Yves Poriau, and Marlies Nering Bogel.

When I first came to Leiden to do my PhD and later my wife joined me from home, setting up our life in Leiden was not as easy as we first anticipated. Luckily, we had a

chance to meet good friends from Leiden and from elsewhere that made our life full of joy. Special thanks to Tadu, Gash Tade, Dr. Azeb, Ermias Michael, Engida and Abeba, Solomon and Emu, Kesis Daniel and Maki, Yinebeb and Roza, and Mahlet and Janok. I would also like to thank our good Christian brothers and sisters of Ethiopian Orthodox church in Rotterdam and Amsterdam.

Honey, my wife (Meseret A. Kerga), you are my life and everything. Without your endless support, love, and prayer, it would not be possible to complete this dissertation. You are the blessing for me, and I am the lucky one to have you in my life. Haniel, our son, you are the little angel and a gift from God. You taught me love and patience; love you so much. I would also like to thank my parents-in-law for your continuous support. Special thanks to my late father-in-law Gashye (missing you so much...) and my mother-in-law Abuye, and Ast and Sam, Se and Masre, Dibo and Mak, Haile and Simegn, Alemu and Almaz, and the wonderful brothers Tse and Da.

Finally, I would like to thank my family back-home. Special thanks to my mother, Manalebish Desta Borbasa, my late sisters Nigmatwa and Gete, Zemen, Etagu and Tafe, Matrik, Seifi, Bermuda, and Andromeda; and, my best childhood friends Legnasil Tekabe and Dawit Girma. Sylvia, Thank you for your continuous support during the time I needed it most while I was doing my BSc degree in Hawassa, Ethiopia. May God bless you and your beloved ones! Lastly, but not the least, I would like to thank my high-school and university teachers. Special thanks to teacher Fasika, Belay, Mengistu and Christopher Rees from Chamo High school, Arbaminch, Ethiopia; and, to Dr. Ayele Taye, Hawassa university, Ethiopia. Thank you all for your support and advice.

Curriculum vitae

Hailemichael M. Worku was born on September 11, 1984 in Arbaminch, Ethiopia. Due to his passion for Mathematics, he chose Natural Science stream during his high school period and graduated in 2002 successfully from Arbaminch Comprehensive Secondary school. He then joined Hawassa University (HU) to do his BSc in Statistics and graduated with greatest distinction in 2005. For two years, from 2006 and 2007, he worked in HU as a graduate assistant in the department of Statistics where he taught introductory statistical courses to undergraduate students. In 2008, he went to Belgium to pursue his graduate study in Biostatistics from Hasselt University and successfully graduated in 2010 with distinction. After his master graduation, he came to The Netherlands for his postgraduate study in the department of Methodology and Statistics, Institute of Psychology, Leiden university. Currently, he lives in Leiden with his wife and a six-years old son, Haniel. In 2015, he started working for a consulting firm based in the Netherlands as analyst and statistical programmer. In his spare time, Hailemichael enjoys going to gym and playing soccer. He has a passion for programming and data science, and because of that he sometimes makes his hand dirty by learning new programming languages - whenever there is some spare time left for him.



```
error_ob = modifier_ob # set to mirror_ob , hope the other one is a mesh
modifier_ob = bpy.context.selected_objects[0]

error_ob = bpy.context.active_object
error_ob.select = False # pop modifier_ob from sel_stack
print("popped")

modifier_ob = bpy.context.selected_objects[0]
print("Modifier object:" + str(modifier_ob.name))

modifier_ob.select=1

print("mirror_ob",mirror_ob)
print("modifier_ob",modifier_ob)

mirror modifier on modifier_ob

error_mod = modifier_ob.modifiers.new("mirror_mirror","MIRROR")

mirror object to mirror_ob
error_mod.mirror_object = mirror_ob

generation == "MIRROR_X":
error_mod.use_x = True
error_mod.use_y = False
error_mod.use_z = False
error_mod.use_s = False
error_mod.use_t = False
error_mod.use_u = False
error_mod.use_v = False
error_mod.use_w = False
error_mod.use_x = True
error_mod.use_y = False
error_mod.use_z = False
error_mod.use_s = False
error_mod.use_t = False
error_mod.use_u = False
error_mod.use_v = False
error_mod.use_w = False
```