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
Invitation to attend the defense of the thesis
Distance Models for Analysis of Multivariate Binary Data
on Thursday 20 December 2018 at 3.00 pm in the Academiegebouw of Leiden University, Rapenburg 73 in Leiden
Paranymphs: Maarten Kampert & Yinebeb Tessema
Hailemichael Metiku Worku
Distance Models for Analysis of Multivariate Binary Data
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. Stolk,
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

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

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

6 Conclusions and Discussions ........................................ 143

Appendices .............................................................. 148

Bibliography ............................................................ 167

Samenvatting ............................................................ 181

Acknowledgments ......................................................... 185

Curriculum vitae .......................................................... 187
Though 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.


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. 

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.

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 dysthmia, 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.

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.

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

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

5.1 Reading the NESDA data available in the mldm package.

5.2 Excerpt of the NESDA data that shows records belonging to the first two subjects.
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 \texttt{mldm.fit} 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 \texttt{biplot()} function available in the \texttt{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 set-
tings. Each cell result is based on $R = 100$ simulated replications. .... 40
2.4 Percentage of Heywood 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, $\gamma_{31}$. 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 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. ...................... 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 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. 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