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

Leiden University Repository

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

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 a 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 a 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, `depression`=1 if a given subject in the study has a depression, otherwise `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., `survival` = 1 if a patient survived; `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., categorical 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:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_p, \tag{1.1}$$

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., $\beta_1, \beta_2, \ldots, \beta_p$ corresponding to the p explanatory variables denoted by x). The link function is $g(.)$ and it connects the random part (i.e., μ) to the systematic part (i.e., $\beta_0 + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\mathsf{T}$).

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 $i = 1, 2, \ldots, N$. 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.,

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

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

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

$$\text{logit}[\pi(\mathbf{x}_i)] = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i, \tag{1.3}$$

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 $i = 1, 2, \ldots, N$.

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,

$$\pi_k(\mathbf{x}_i) = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k^\mathsf{T}\mathbf{x}_i)}{\sum_c \exp(\beta_{0c} + \boldsymbol{\beta}_c^\mathsf{T}\mathbf{x}_i)}. \tag{1.4}$$

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

$$\text{logit}[\pi_k(\mathbf{x}_i)] = \beta_{0k} + \boldsymbol{\beta}_k^\mathsf{T}\mathbf{x}_{ik}, \tag{1.5}$$

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),

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

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),

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

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),

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

where z_{tm} is the coordinate of object t on dimension m $(m = 1, 2, \ldots, M)$. 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, $\delta_{AD} = \left[(z_{A1} - z_{D1})^2 + (z_{A2} - z_{D2})^2\right]^{1/2} = \left[(6-3)^2 + (7-6)^2\right]^{1/2} = 3.16$. Similarly, $\delta_{AC} = \left[(z_{A1} - z_{C1})^2 + (z_{A2} - z_{C2})^2\right]^{1/2} = \left[(6-7)^2 + (7-3)^2\right]^{1/2} = 4.1$. 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.

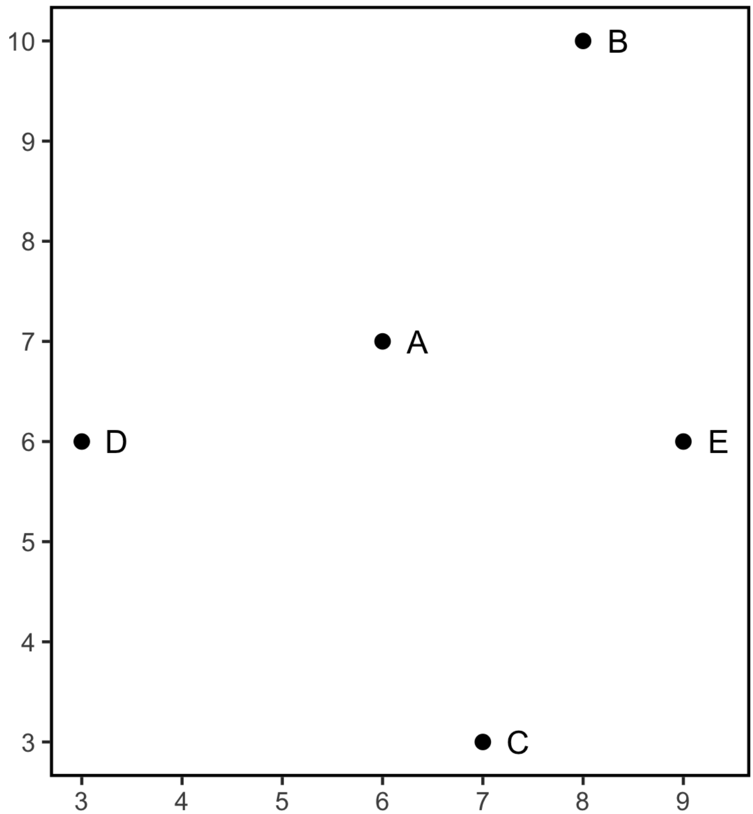

Figure 1.1: MDS Model: A two-dimensional configuration of dissimilarity data with five objects (i.e., `A, B, C, D` and `E`).

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.

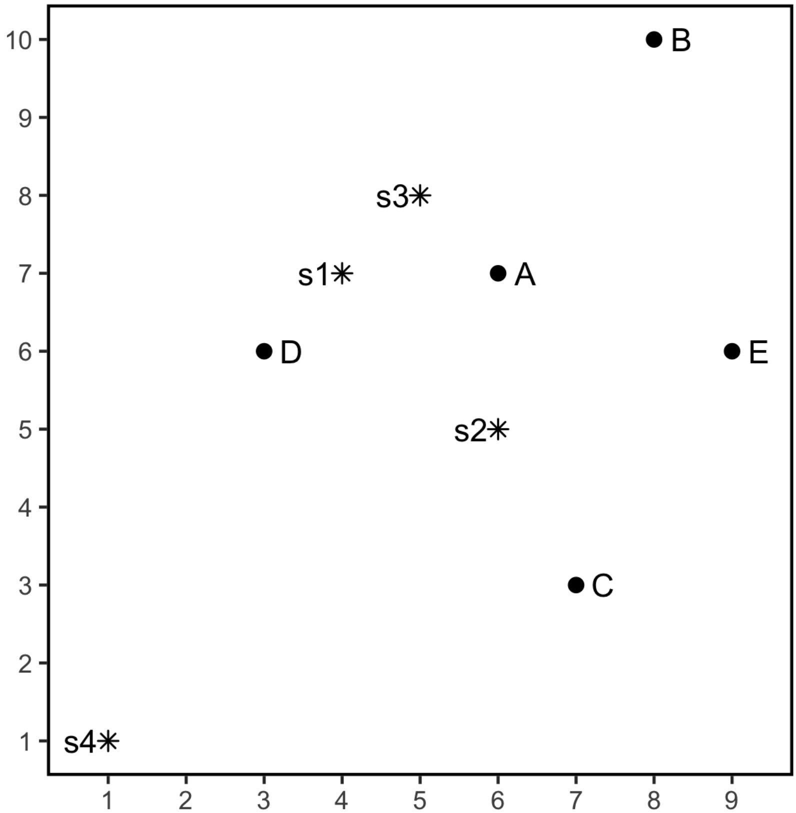

Figure 1.2: MDU Model: A two-dimensional configuration of preference data with four subjects (i.e., `s1, s2, s3` and `s4`) and five objects (i.e., `A, B, C, D` and `E`).

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, $\delta_{S4,A} = \left[(z_{S4,1} - z_{A1})^2 + (z_{S4,2} - z_{A2})^2\right]^{1/2} = \left[(1-6)^2 + (1-7)^2\right]^{1/2} = 7.81$. Similarly, $\delta_{S4,C} = \left[(z_{S4,1} - z_{C1})^2 + (z_{S4,2} - z_{C2})^2\right]^{1/2} = \left[(1-7)^2 + (1-3)^2\right]^{1/2} = 6.3$. 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),

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

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.,

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

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),

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

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,

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

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,

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

The class points of the *unidimensional* IPC model are given by $\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{01}, & \gamma_{11} \end{bmatrix}^\mathsf{T}$, 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,

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

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

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

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., $\boldsymbol{\beta}$) and an intercept with an offset of half (i.e., $\beta_0^{\mathsf{IPC}} = \beta_0^{\mathsf{LR}} + 0.5$).

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,

$$\gamma = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1.16}$$

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., $\boldsymbol{\gamma}_{2m} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$), the third category is on the $y-$axis (i.e., $\boldsymbol{\gamma}_{3m} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$), and the fourth category is on the $z-$axis (i.e., $\boldsymbol{\gamma}_{4m} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$). 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,

$$\begin{aligned}
\text{logit}[\pi_1(\mathbf{x}_i)] &= 0.5 * \delta_{i0}^2 - 0.5 * \delta_{i1}^2 \\
&= 0.5 * \sum_{m=1}^{3} (\eta_{im} - \gamma_{0m})^2 - 0.5 * \sum_{m=1}^{3} (\eta_{im} - \gamma_{1m})^2 \\
&= \sum_{m=1}^{3} (\gamma_{1m} - \gamma_{0m}) * \eta_{im} + 0.5 * \sum_{m=1}^{3} (\gamma_{0m}^2 - \gamma_{1m}^2) \\
&= \eta_{i1} - 0.5 \\
&= (\beta_{01} - 0.5) + \boldsymbol{\beta}_1^\mathsf{T} \mathbf{x}_i.
\end{aligned} \tag{1.17}$$

Similarly, the log-odds for the third category: $\text{logit}[\pi_2(\mathbf{x}_i)] = \eta_{i2} - 0.5 = (\beta_{02} - 0.5) + \boldsymbol{\beta}_2^\mathsf{T} \mathbf{x}_i$, and the log-odds for the fourth category: $\text{logit}[\pi_3(\mathbf{x}_i)] = \eta_{i3} - 0.5 = (\beta_{03} - 0.5) + \boldsymbol{\beta}_3^\mathsf{T} \mathbf{x}_i$. Thus, $\boldsymbol{\beta}_p^{\text{IPC}} = \boldsymbol{\beta}_p^{\text{MBCL}}$ for regression coefficients with dimension $(p \times M)$, and $\boldsymbol{\beta}_0^{\text{IPC}} = \boldsymbol{\beta}_0^{\text{MBCL}} - 0.5$ 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 ($1 = \texttt{yes}$; $0 = \texttt{no}$) and wheeze ($1 = \texttt{yes}$; $0 = \texttt{no}$), 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 $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{iJ})^T$ 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, ..., X_p.

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

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

$$\log(\tau_i) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i, \tag{1.18}$$

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 $\mathbf{y}_i = (\mathrm{y}_{i1}, \mathrm{y}_{i2}, \ldots, \mathrm{y}_{ij})$ 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 $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iq})$, and j unique factors (or measurement errors), denoted by $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{ij})$, with $j > q$

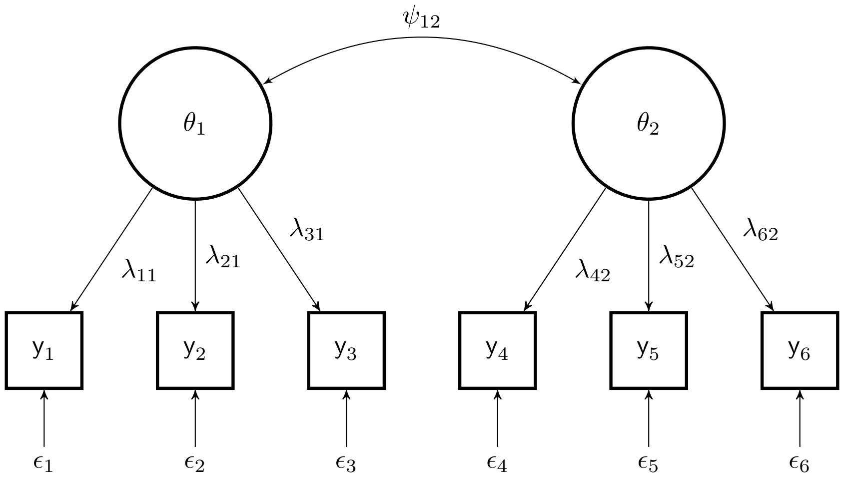

Figure 1.3: A path diagram of a CFA with six indicator variables represented by a square, and two latent variables represented by a circle.

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

$$
\begin{aligned}
\mathrm{y}_{i1} &= \lambda_{11}\theta_{i1} + \ldots + \lambda_{1q}\theta_{iq} + \epsilon_{i1} \\
\mathrm{y}_{i2} &= \lambda_{21}\theta_{i1} + \ldots + \lambda_{2q}\theta_{iq} + \epsilon_{i2} \\
&\vdots \\
\mathrm{y}_{ij} &= \lambda_{p1}\theta_{i1} + \ldots + \lambda_{jq}\theta_{iq} + \epsilon_{ij}
\end{aligned}
$$

or, in matrix form

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \tag{1.19}$$

where $\mathbf{\Lambda}$ is the matrix of factor loadings. Let $\mathbf{\Psi}$ be the covariance matrix of common factors, and let $\mathbf{\Phi}$ 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 distributions, i.e., $\boldsymbol{\theta} \sim \mathsf{N}_q(\mathbf{0}, \boldsymbol{\Psi})$ and $\boldsymbol{\epsilon} \sim \mathsf{N}_j(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ is a diagonal matrix. Given the model, the expected covariance matrix of the indicator variables becomes

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Phi}. \tag{1.20}$$

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 $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{iJ})$ 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 $\boldsymbol{y}_i^* = (y_{i1}^*, y_{i2}^*, \ldots, y_{ij}^*, \ldots, y_{iJ}^*)$. 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

$$\mathbf{y}_i^* = \boldsymbol{\Lambda}\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \qquad y_{ij} = \begin{cases} 1, & \text{if } y_{ij}^* \geq \tau_j, \\ 0, & \text{if } y_{ij}^* < \tau_j. \end{cases} \tag{1.21}$$

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

$$\boldsymbol{\Phi} = \mathbf{I} - \mathrm{diag}(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{\mathrm{T}}), \tag{1.22}$$

yielding $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}$. Therefore, the model has three sets of free parameters: $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}$ (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 $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ 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.,

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

where $\boldsymbol{\Gamma}$ 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

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

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

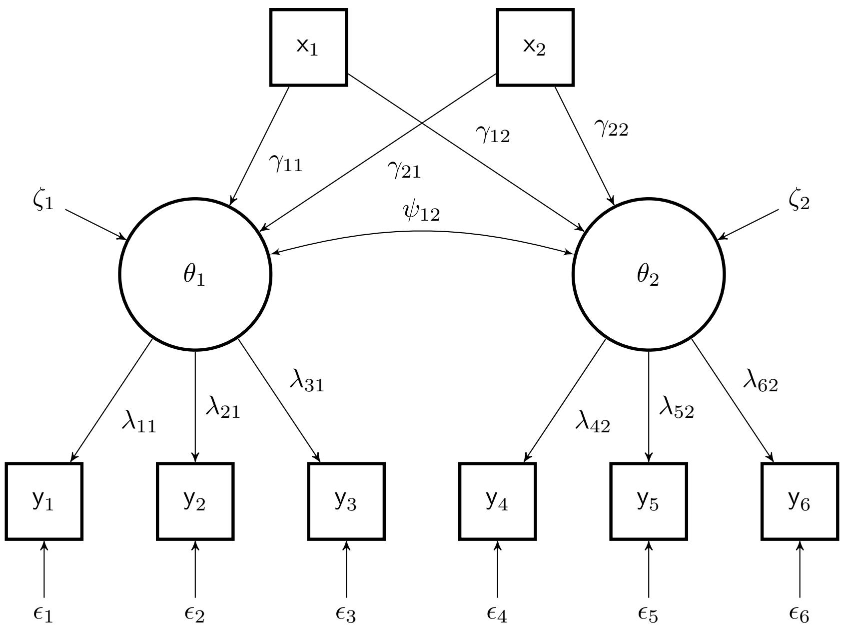

Figure 1.4: A path diagram for a MIMIC model with two external variables that are represented by a square.

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.