Journal of Classification 35 (2018)
DOI: 10.1007/s00357-018-9251-4

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

Hailemichael M. Worku

Leiden University, The Netherlands

Mark de Rooij

Leiden University, The Netherlands

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 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.

Keywords: Multivariate binary data; Biplots; Multidimensional scaling; Multidimensional unfolding; Marginal model; Clustered bootstrap; Generalized estimating equations.

Corresponding Author's Address: H.M. Worku, Psychology Institute, Methodology and Statistics Unit, Leiden University, Wassenaarseweg 52, Box 9555, 2300 RB, Leiden, The Netherlands, e-mail: hmetiku@yahoo.com.

Published online

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 (Pennix et al., 2008; Spinhoven et al., 2009). Another study in which multivariate binary data arises is the Indonesian Children's Study (ICS: Sommer, Katz, and Tarwotjo, 1984; Liang, Zeger, and 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.

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, Kastner, and Blettner, 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 dimensions 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 this paper, we will develop a dimensional model for multivariate binary data within the marginal framework. The model does not make unverifiable assumptions. 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 visualizes the result.

De Rooij (2009) proposed the Ideal Point Classification (IPC) model for analyzing a multinomial response variable in the presence of predic-

tors. The IPC is a probabilistic distance model based on a two-mode distance function. De Rooij (2009) also showed that a simple logistic regression for binary response variable can be written as a unidimensional IPC model. Worku and De Rooij (2016) 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 and Wish, 1978; Borg and Groenen, 2005) and Generalized Linear Model (GLM: McCullagh and 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 and Hand, 1996; Gower, Lubbe, and Le Roux, 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.

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.

2. Multivariate Logistic Regression in a Distance Framework

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

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

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

De Rooij (2009) 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 coordinate of the position for 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,

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

With these two distances we can define the following probability model

$$\pi(x_i) = \frac{\exp(-0.5\delta_{i1})}{\exp(-0.5\delta_{i0}) + \exp(-0.5\delta_{i1})}. \tag{3}$$

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 $\eta_i = \beta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$. 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,

$$\begin{aligned}\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] &= 0.5\delta_{i0} - 0.5\delta_{i1} \\ &= \eta_i(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) \\ &= (\beta_0 + \beta_1 x_i)(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) \\ &= \beta_0(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) + \beta_1(\gamma_1 - \gamma_0)x_i.\end{aligned} \tag{4}$$

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

$$\begin{aligned}\log\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right] &= \beta_0(\gamma_1 - \gamma_0) + 0.5(\gamma_0^2 - \gamma_1^2) \\ &\quad + \beta_1(\gamma_1 - \gamma_0)x_{i1} + \beta_2(\gamma_1 - \gamma_0)x_{i2}.\end{aligned} \tag{5}$$

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) is not identified and therefore an 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 $\gamma_1 = 0$. The logistic distance model is now identified and its relationship with the univariate logistic model presented in (1) becomes

$$\begin{aligned}\beta_0^* &= \beta_0 - 0.5; \\ \beta_1^* &= \beta_1.\end{aligned} \tag{6}$$

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 $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{iJ})^{\mathsf{T}}$ denotes the multivariate responses observed on

Table 1. The structure of multivariate data in long format.

			Predictor variables			
SID	Index	Response	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}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
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}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
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}

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 1, we display the structure of multivariate data in long format. The first column, `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 1 five response variables are assumed, i.e., $R_1, R_2, \ldots, R_5$. The other columns represent measurements of the response variable and predictor variables, respectively.

A unidimensional space was used to represent the logistic regression model (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 dimension, 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, $\eta_{im} = \beta_{0m} + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_m$; 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 dysthimia; 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

$$\pi_j(\mathbf{x}_i) = \frac{\exp\left[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})\right]}{\exp\left[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{0j})\right] + \exp\left[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})\right]}. \tag{7}$$

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

$$\begin{aligned}\log\left[\frac{\pi_j(\mathbf{x}_i)}{1 - \pi_j(\mathbf{x}_i)}\right] = \sum_{m=1}^{M} \Big\{ &\beta_{0m}(\gamma_{1j,m} - \gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2 - \gamma_{1j,m}^2) \\ &+ \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_m(\gamma_{1j,m} - \gamma_{0j,m}) \Big\}.\end{aligned} \tag{8}$$

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, (8) simplifies to a single equation instead of a sum over dimensions.

This distance model for multivariate binary data (7 - 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.

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.,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \pi_j(\mathbf{x}_i)^{y_{ij}} [1 - \pi_j(\mathbf{x}_i)]^{(1-y_{ij})}, \tag{9}$$

where $\boldsymbol{\theta}$ is the concatenation of all the class points and all the regression weights.

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 sandwich estimator for the covariance matrix to correct for the bias (Liang and Zeger, 1986). Another method for obtaining correct standard errors is to apply a clustered bootstrap method (Sherman and Le Cessie, 1997; De Rooij and Worku, 2012; Cheng et al., 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

$$q = [M \times (p + 1)] + [(J - M) \times 2]. \tag{10}$$

The first term in (10), i.e., $[M \times (p + 1)]$, corresponds to the number of regression coefficients; the other term to the number of estimable 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.

2.4 The Relationship of the MLD Model to Generalized Estimating Equations

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 and 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 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,

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. \tag{11}$$

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), $\mathbf{U}_i = \mathbf{Z} \otimes \mathbf{x}_i^{\mathsf{T}}$, such that

$$\mathbf{U}_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}. \tag{12}$$

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.

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:

$$\text{QIC} = -2L(\boldsymbol{\theta}) + 2\ \text{trace}(\hat{\boldsymbol{\Omega}}_I^{-1}\hat{\mathbf{V}}_R), \tag{13}$$

where $\hat{\mathbf{V}}_R$ stands for 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.

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 and 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

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

Because each response is positioned on one and only one dimension, one of the columns in $\boldsymbol{\gamma}_j$ equals zero. So, $\boldsymbol{\gamma}_j$ 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 and 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.

$$\mathbf{B} = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{bmatrix}.$$

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.

3. Application: The NESDA Data

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 dysthmia (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

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

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.

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

$$\mathbf{Z}^{(1)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Z}^{(2)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{Z}^{(3)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \tag{14}$$

respectively. The superscript corresponds to a theory.

Table 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	$\mathbf{12,396.42}$
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	$\mathbf{12396.42}$
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.

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

The regression weights of this selected model are given in Table 3. The standard errors based on both the sandwich and the clustered bootstrap method are included in Table 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.

Table 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, $B = 1000$.

				Bootstrap	
Effect	Parameter	Estimate	SE (sandwich)	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, $p < 0.05$.

Although the total number of parameters of the final d/f model is $q = 21$, only sixteen of the parameters were displayed in Table 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 than for DYST. The model accounts for comorbidity because a high probability of DYST implies a high probability of GAD and MDD.

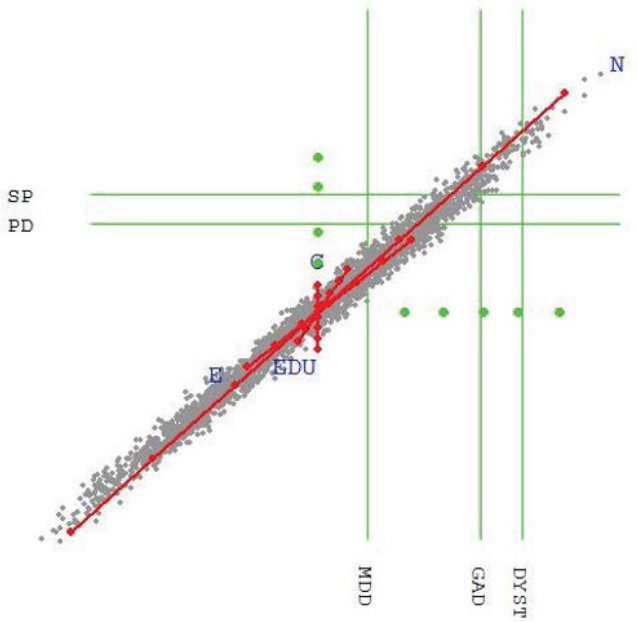

Figure 1. Biplot of the final "distress-fear" model fitted on the NESDA data. The plot is based on restrictions applied on the class points.

The results of the selected MLD model are displayed in a biplot shown in Figure 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 Figures 2 and 3, respectively. The positions of the two categories of all response variables are displayed in Figure 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.

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 3, both the variable axes (lines) and the subjects points (grey dots) are displayed. The space includes only statistically significant predic-

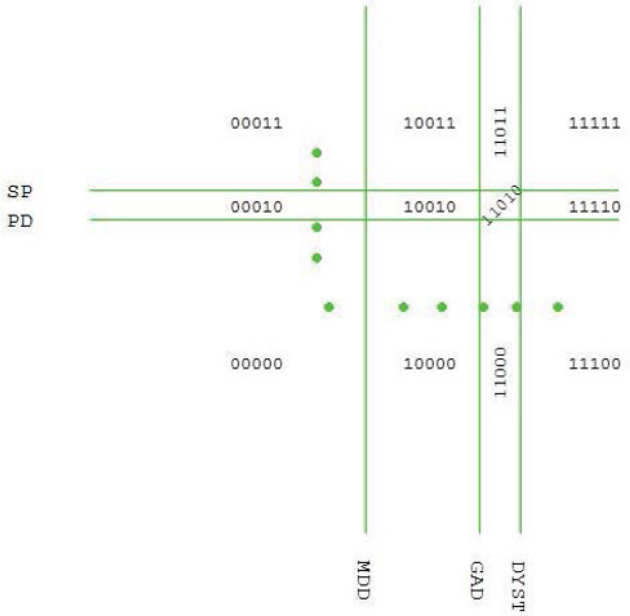

Figure 2. Representation of the binary response variables in the Euclidean space.

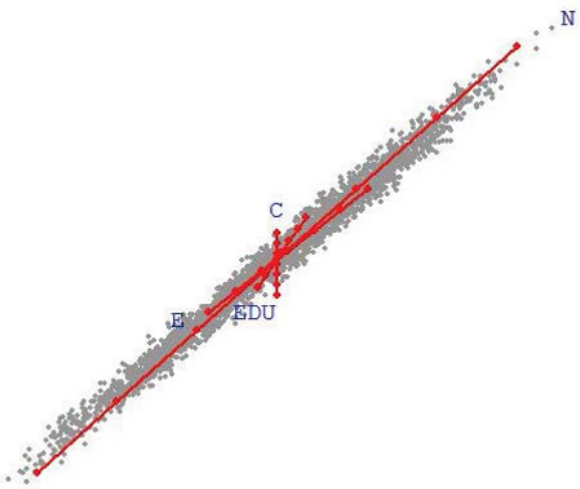

Figure 3. Variable axes representation of the predictor variables in the Euclidean space.

tors. 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 the variable axis.

Let us now interpret the biplot displayed in Figure 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), 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 dimension (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 esti-

Table 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.

			Bootstrap	
Effect	Parameter	Estimate	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, $p < 0.05$.

mates are shown in Table 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 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 (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 discriminates between the categories. In the fitted model, both DYST and GAD are posi-

tioned 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. 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^o 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 and 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 et al., 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. 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 and Anderson, 1986; Boomsma and 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 and 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.

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.

References

ACITO, F., and ANDERSON, R.D. (1986), "A Simulation Study of Factor Score Indeterminacy", *Journal of Marketing Research, 23,* 111–118.

AGRESTI, A. (2002), *Categorical Data Analysis* (2nd ed.), New York: John Wiley and Sons.

AKAIKE, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle", in *Proceedings of the Second International Symposium on Information Theory*, eds. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–281.

ASAR, Ö., and 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.

BEESDO-BAUM, K. et al. (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.

BOOMSMA, A., and HOOGLAND, J.J. (2001), "The Robustness of LISREL Modeling Revisted", in *Structural Equation Modeling: Present and Future,* eds. R. Cudeck, S. de Toit, and D.Sörbom, Chicago: Scientific Software International, pp. 139–168.

BORG, I. , and GROENEN, P.J.F. (2005), *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.), New York: Springer.

BULL, S.B. (1998), "Regression Models for Multiple Outcomes in Large Epidemiological Studies", *Statistics in Medicine, 17,* 2179–2197.

CHENG, G., YU, Z., and HUANG, J.Z. (2013), "The Cluster Bootstrap Consistency in Generalized Estimating Equations", *Journal of Multivariate Analysis, 115,* 33–47.

COSTA, P.T., and 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.

DE ROOIJ, M. (2009), "Ideal Point Discriminant Analysis with a Special Emphasis on Visualization", *Psychometrika, 74,* 317–330.

DE ROOIJ, M., and 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.

ELLIOT, D.S., HUIZINGA, D., and MENARD, S. (1989), *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems,* New York: Springer-Verlag.

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G., and 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 and Sons.

GOWER, J.C., and HAND, D.J. (1996), *Biplots,* London: Chapman and Hall.

GOWER, J.C., LUBBE, S., and LE ROUX, N. (2011), *Understanding Biplots*, Chichester: John Wiley and Sons Ltd.

HALEKOH, U., HOJSGAARD, S., and YAN, J. (2006), "The R Package geepack for Generalized Estimating Equations", *Journal of Statistical Software, 15(2)*, 1–11.

HUBBARD, A.E. et al. (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.

KRUEGER, R.F. (1999), "The Structure of Common Mental Disorders", *Archives of General Psychiatry, 56,* 921–926.

KRUSKAL, J.B., and WISH, M. (1978), *Multidimensional Scaling,* Sage Publications.

LIANG, K.Y., and ZEGER, S.L. (1986), "Longitudinal Data Analysis Using Generalised Linear Models", *Biometrika, 73,* 13–22.

LIANG, K.Y., ZEGER, S.L., and 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., and ZHAO, L.P. (1994), "Analysis of Repeated Categorical Data Using Generalized Estimating Equations", *Statistics in Medicine, 14,* 1149–1163.

MCCULLAGH, P., and NELDER, J.A. (1989), *Generalized Linear Models,* London: Chapman and Hall.

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.

PENNINX, B.W. et al. (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 (2013), "R: A Language and Environment for Statistical Computing", Computer Software Manual Version 3.0.2, Vienna, Austria, http://www.r-project.org/.

SAS INSTITUTE INC. (2011), "SAS/STAT Software", Computer Software Manual Version 9.3, Cary, NC, http://www.sas.com.

SHERMAN, M., and 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.

SOMMER, A., KATZ, J., and TARWOTJO, I. (1984), "Increased Risk of Respiratory Disease and Diarrhea in Children with Preexsting Mild Vitamin A Deficiency", *American Society for Clinical Nutrition, 40,* 1090–1095.

SPINHOVEN, P., DE ROOIJ, M., HEISER, W., PENNINX, B.W.J.H., and SMIT, J. (2009), "The Role of Personality in Comorbidity Among Anxiety and Depressive Disorders in Primary Care and Speciality Care: A Cross-Sectional Analysis", *General Hospital Psychiatry, 31,* 470–477.

SPINHOVEN, P., PENELO, E., DE ROOIJ, M., PENNINX, B,W., and 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.

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., and VERDONSCHOT, P.F.M. (1995), "Canonical Correspondence Analysis and Related Multivariate Methods in Aquatic Ecology", *Aquatic Sciences, 57(3),* 1015–1621.

VAN DER HEIJDEN, P.G.M., MOOIJAART, A., and TAKANE, Y. (1994), "Correspondence Analysis and Contingency Models", in *Correspondence Analysis in the Social Sciences*, eds. M.J. Greenacre and J. Blasius, New York: Academic Press, pp. 79–111.

VON OERTZEN, T., HERTZOG, C., LINDENBERGER, U., and 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., and STRAM, D. (1988), "Analysing 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., and DE ROOIJ, M. (2016), "Properties of Ideal Point Classification Models for Bivariate Binary Data", *Psychometrika* (accepted for publication).

ZIEGLER, A. (2011), *Generalized Estimating Equations,* New York: Springer.

ZIEGLER, A., and 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., and BLETTNER, M. (1998), "The Generalized Estimating Equations: An Annotated Bibliography", *Biometrical Journal, 40(2),* 115–139.