Universiteit
Leiden
The Netherlands

Distance models for analysis of multivariate binary data

Worku, H.M.

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

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/67140

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

Cover Page

Universiteit Leiden

The following handle holds various files of this Leiden University dissertation:
http://hdl.handle.net/1887/67140

Author: Worku, H.M.
Title: Distance models for analysis of multivariate binary data
Issue Date: 2018-12-20

Chapter 6

Conclusions and Discussions

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

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

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

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

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

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

marginal models and the association model.

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

A limitation of the BIPC model, however, is that it is not straightforward to extend it for analyzing multivariate binary data (i.e., with more than two binary or polytomous responses). This is due to the fact that both the pairwise and higher-order association structure parameters between the dependent variables must be specified in the likelihood function. With three binary responses (i.e., Y_1, Y_2, and $Y3$), for example, three pairwise associations and a three-way association parameters must be specified which makes the computation cumbersome. Due to this limitation of the BIPC model, we proposed a new distance-based marginal model in Chapter 4, namely the Multivariate Logistic Distance (MLD) model, for analyzing multivariate binary data.

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

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

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

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

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

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

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

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