Summary

In this dissertation we developed statistical tools for analyzing multivariate binary data. In many disciplines multivariate binary data, in which there are multiple binary dependent variables and one or more independent variables, are often collected. In the Indonesian Children's Study (ICS), for example, over three-thousand children were medically examined to investigate whether they had respiratory infection, i.e., 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. 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. 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.

Statistical tools for analyzing multivariate continuous dependent variables are widely available due to the presence of multivariate normal distribution which is the building block of the methodology. Multivariate regression and MANOVA, to mention but a few, are among the popular methods that are applied in this area. For multivariate categorical data, however, only limited statistical tools are available. Moreover, existing methodology makes unverifiable assumptions (e.g., latent variable models and structural equation models) or requires the independent variables to to be categorized (e.g., GEE2 method for marginal

models). Using a Monte Carlo simulation study, we showed in Chapter 2 that latent variable models applied on multivariate binary data and with only a few indicators per latent variable (i.e., 2 or 3) performed poorly.

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). In Chapter 3 we studied properties of the IPC model for analyzing bivariate binary data. A bivariate logistic regression set-up 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; the second pertains to the prevalence of the second variable; and, the third dimension pertains to the association between the two dependent variables. Based on a simulation study we showed that the IPC model does not fully recover all the three parameters of bivariate binary data, i.e., the two marginal parameters and the association between the two dependent variables. In the same chapter we proposed the Bivariate IPC (BIPC) model which is a re-parameterization of the IPC model, and the required parameters of bivariate binary data are fully recovered.

A limitation of the BIPC model is that it is not straightforward to extend it for analyzing multivariate binary data (i.e., with more than two binary dependent variables). 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 unifies two domains of statistical methods, i.e., Multidimensional Scaling (MDS) and Generalized Linear Model (GLM). 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. For the NESDA data, for example, a researcher can use the MLD model to determine the dimensional structure of the mental disorders, and to investigate effect of the personality traits and the background variables on prevalence of the mental disorders.

To enhance interpretation, the results of the MLD model can be graphically represented in a biplot. Another 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.

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. 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 marginal model which can be fitted by its own right. In Chapter 5 we presented an **mldm** package that was developed in R statistical software for fitting the MLD model. The main function in the **mldm** package responsible for fitting the MLD model is `mldm.fit()`. 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 models. We made the **mldm** package available on the online repository system GitHub. The following link can be used to get access to the package: https://github.com/workuhm1/mldm-package-github.

Finally, we recommend applied researchers that they should be careful in using latent variable models (or structural equation models) for analyzing multivariate binary data and with only a few indicators per latent variable (i.e., 2 or 3) because these methods do perform poorly for this type of data. An alternative statistical model which requires less assumptions might be more appropriate, for example the multivariate logistic distance model.