Distance models for analysis of multivariate binary data

Worku, H.M.

Chapter 5

mldm: An R Package for Analyzing Multivariate Binary Data

Abstract

We developed the **mldm** package in R (R Development Core Team, 2008) to fit a multivariate logistic distance model on multiple binary responses in the presence of explanatory variables. The package handles both the clustered bootstrap method and the sandwich estimators for obtaining the standard errors of model parameters. The package provides a `biplot` function to display results of the fitted model. In this chapter we illustrate the usage of the package using an empirical data.

This chapter is a user manual for the **mldm** package in R software developed by Worku, H. M. & De Rooij, M. (2018) for analyzing multivariate binary data.

5.1 Introduction

The Multivariate Logistic Distance (MLD) model is proposed to analyze multiple binary responses in the presence of explanatory variables (Worku & De Rooij, 2018). The MLD model unifies two domains of statistical methods, i.e., Multidimensional Scaling (MDS: Kruskal & Wish, 1978; Borg & Groenen, 2005) and the 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 the existing marginal models for multivariate data. Moreover, the model enhances interpretation by using a biplot (Gabriel, 1971; Gower & Hand, 1996; Gower et al., 2011) based on a distance interpretation.

For fitting the MLD model, we developed the **mldm** package in `R` statistical software. In this chapter we illustrate the usage of the package with an empirical data.

5.2 The Multivariate Logistic Distance Model

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.

The MLD model defines the probability for category 1 on response variable j given the explanatory variables, i.e. $\Pr(y_{ij} = 1|\mathbf{x}_i) = \pi_j(\mathbf{x}_i)$, as

$$\pi_j(\mathbf{x}_i) = \frac{\exp\left[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})\right]}{\exp\left[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{0j})\right] + \exp\left[-0.5\delta(\boldsymbol{\eta}_i, \boldsymbol{\gamma}_{1j})\right]}. \tag{5.1}$$

The log-odds representation of (5.1) becomes,

$$\log\left[\frac{\pi_j(\mathbf{x}_i)}{1-\pi_j(\mathbf{x}_i)}\right] = \sum_{m=1}^{M}\Big\{\beta_{0m}(\gamma_{1j,m}-\gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2-\gamma_{1j,m}^2) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_m(\gamma_{1j,m}-\gamma_{0j,m})\Big\}. \tag{5.2}$$

Because each response variable belongs to a single dimension (see Chapter 4), the log odds representation can be further simplified. Suppose response variable j belongs to the first dimension 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 (5.2) simplifies to a single equation instead of a sum over dimensions. Moreover, as the MLD model is a type of bilinear model, for each dimension we have to fix the origin and scale.

5.2.1 Parameter Estimation

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 can be fitted using the Generalized Estimating Equation method (Liang & Zeger, 1986). Therefore, existing statistical packages with a GEE procedure (e.g., the **geepack** package in `R` or the **genmod** procedure in `SAS` software) can be used for fitting the "restricted" MLD model on multivariate binary data. The restriction of the class points implies that explanatory variables discriminate equally well for all response variables belonging to a specific dimension.

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 explanatory variables; (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 4.1. 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,

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. \tag{5.3}$$

The design matrix for subject i is then obtained by computing the Kronecker product between the response indicator matrix and the explanatory variables vector (without intercept), $\mathbf{U}_i = \mathbf{Z} \otimes \mathbf{x}_i^T$, such that

$$\mathbf{U}_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}. \tag{5.4}$$

We concatenate $\mathbf{U}_i$ and the identity matrix to get the final design matrix, $\mathbf{S}_i = [\mathbf{I}, \mathbf{U}_i]$,

$$\mathbf{S}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & 0 & x_{i1} & x_{i2} & \dots & x_{ip} & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}.$$

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 ($\beta^*_{01}, \ldots, \beta^*_{05}$) corresponding to the first five columns of $\mathbf{S}$ and two sets of p regression weights ($\beta^*_{11}, \ldots, \beta^*_{p1}$ and $\beta^*_{12}, \ldots, \beta^*_{p2}$), corresponding to the two dimensions. The MLD parameters can be derived from these as follows $\gamma_{0j,m} = -(\beta^*_{0j} + 0.5)$ for 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 $q = J + (M \times P)$ since additional constraints are imposed on the class points.

5.3 The NESDA Data

We used the Netherlands Study of Depression and Anxiety (NESDA: Penninx et al., 2008; Spinhoven et al., 2009) data as a working example in this chapter to demonstrate usage of the **mldm** package. NESDA is an ongoing cohort study designed to investigate determinants of depressive and anxiety disorders in a relatively large and representative sample of participants. In the current version of data we have, there are $N = 2,938$ subjects of age between $18 - 65$ years with an average age of 42(S.D. $= 13.1$) in which 66.5% were female and the average number of years of education attained was 12.2(S.D. $= 3.3$).

The multivariate binary responses are major depressive disorder (MDD), dysthmia (DYST), generalized anxiety disorder (GAD), social phobia (SP), and panic disorder (PD). In this study, about 37.1% of the subjects had MDD, 10.2% had DYST, 15.3% had GAD, 22.4% had SP, and 28.6% had PD. The explanatory variables are the Big-Five personality traits (i.e., neuroticism, extroversion, openness to experience, agreeableness, and conscientiousness) and three background variables (i.e., age, years of education attained and

gender).

5.4 The mldm Package

5.4.1 Accessing the NESDA Data

The NESDA data is available in the **mldm** package. For demonstration purpose, we selected a random sample of 600 subjects. Thus, there are a total of 3000 records in the dataset with binary responses for the five disorders that are observed for each subject. Figure 5.1 shows how to extract the NESDA data from the **mldm** package. The dataset becomes available for us after we install and load the **mldm** package to the R environment.

```
# load the package
library(mldm)
# load the NESDA data
data(NESDA)
## View the dimension of the data
dim(NESDA)

## [1] 3000   11
```

Figure 5.1: Reading the NESDA data available in the **mldm** package.

The measurements from the first two subjects are displayed in Figure 5.2 below. The dataset is in a person-item order where measurements of the five mental disorders are nested within a subject. The first five rows represent measurements of the five mental disorders obtained for the first subject. The other measurements (i.e., row $6 - 10$) are the same set of measurements for the second subject.

The last column (`pident`) has subject's identification number. The `Index` column

is used to identify which response variable (e.g., mental disorder) is measured in a given row. There are a total of five mental disorders measured for each subject. The `Outcome` column holds the actual binary measurements that belong to each response variable (e.g., mental disorder). For example, the first subject (`pident` $= 1$) has a social phobia disease because `SP` $= 1$ while all the other measurements are zero. Whereas the second subject (`pident` $= 2$) has none of the mental disorders. All the other columns in Figure 5.2 represent measurements on the explanatory variables, i.e., the Big-Five personality traits and the three background variables.

```
# display measurements of the first ten subjects
print(head(NESDA, n=10), digits = 0, row.names = FALSE,
      right = FALSE)
```

GEN	AGE	EDU	N	E	O	A	C	Outcome	Index	pident
male	41	18	43	28	31	41	35	0	DYST	1
male	41	18	43	28	31	41	35	0	MDD	1
male	41	18	43	28	31	41	35	0	GAD	1
male	41	18	43	28	31	41	35	1	SP	1
male	41	18	43	28	31	41	35	0	PD	1
male	59	9	32	43	31	40	42	0	DYST	2
male	59	9	32	43	31	40	42	0	MDD	2
male	59	9	32	43	31	40	42	0	GAD	2
male	59	9	32	43	31	40	42	0	SP	2
male	59	9	32	43	31	40	42	0	PD	2

Figure 5.2: Excerpt of the NESDA data that shows records belonging to the first two subjects.

5.4.2 Model Specification and Fitting

Suppose we would like to fit a 2-dimensional MLD model on the NESDA data where the DYST and MDD response variables belong to the first dimension while the other response variables (i.e., GAD, SP and PD) belong to the second dimension. This model is also called the depression-anxiety model (Penninx et al., 2008).

This model can be fitted in the **mldm** package by first specifying a response indicator matrix (i.e., the matrix that indicates to which dimension every response belongs). Figure 5.3 shows the response indicator matrix for our 2-dimensional model.

```
# Indicator matrix
Z <- matrix(c(1,1,0,0,0,
              0,0,1,1,1), 5, 2, byrow=FALSE)
print(Z)

##      [,1] [,2]
## [1,]    1    0
## [2,]    1    0
## [3,]    0    1
## [4,]    0    1
## [5,]    0    1
```

Figure 5.3: Specification of an indicator matrix for the depression-anxiety model fitted on the NESDA data.

The multivariate logistic distance model for the NESDA data is given by

$$\begin{aligned}\eta_{im} = \beta_{1m} \times \mathsf{AGE} + \beta_{2m} \times \mathsf{EDU} + \beta_{3m} \times \mathsf{GEN} \\ + \beta_{4m} \times \mathsf{N} + \beta_{5m} \times \mathsf{E} + \beta_{6m} \times \mathsf{O} + \beta_{7m} \times \mathsf{A} + \beta_{8m} \times \mathsf{C},\end{aligned} \tag{5.5}$$

where η_{im} is the $i-$th subject coordinate in dimension m. Since we want to fit the depression and anxiety model, the number of dimension becomes two, i.e., $M = 2$.

The R code below in Figure 5.4 shows the formula following (5.5) for every dimension. For simplicity reason, we only considered two of the Big-Five personality traits (N and E), and all the background variables. The left side of the formula (i.e., before the tilde sign) shows the response variable (`Outcome`) separated by a pipe operator. For a unidimensional MLD model, we would not use the pipe operator. At the right side of the formula, the explanatory variables are displayed which are again separated by the pipe operator. The pipe operator tells R that the components are dimension-specific. In our case, the MLD model is a 2-dimensional model.

```
## specify model formula
mf <- Outcome | Outcome ~ EDU + GEN + AGE + N + E |
      EDU + GEN + AGE + N + E
mf <- Formula(mf)
```

Figure 5.4: A two-dimensional representation of model formula for depression-anxiety model fitted on the NESDA data.

As model specification in MLD is dimension specific, it is possible to allow a given explanatory variable to have an effect in one of the dimension but not on the other one. For example, we can test the hypothesis that agreeableness has an effect on the first dimension (depression), but not on the second dimension (anxiety) while both openness to experience and conscientiousness having an effect on anxiety, but not on depression.

Once the response indicator matrix (`Z`) and the the model formula (`mf`) are specified, we are now ready to fit the MLD model using the `mldm.fit` function as shown in Figure 5.5 below.

```
# fit the MLD model on NESDA data
fit <- mldm.fit(formula=mf, index = Index, resp.dim.ind = Z,
                data = NESDA, id = pident, scale=TRUE)
# display fitted model result
fit

 Call:
mldm.fit(formula = mf, index = Index, resp.dim.ind = Z, data = NESDA,
    id = pident, scale = TRUE)
 Formula:
Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE +
    N + E
 QIC: [1] 2542.707
```

Figure 5.5: Application of the `mldm.fit` function for fitting the depression-anxiety model on the NESDA data.

Except the `scale` parameter in the `mldm.fit()` function, the input values for the other parameters are obtained both from the dataset itself (i.e., `index = Index`, `data = NESDA` and `id = pident`) and model specification (i.e., `formula = mf` and `resp.dim.ind = Z`). The `scale` argument in the `mldm.fit()` function is used for transforming the explanatory variables to z-scores.

The `mldm.fit()` function returns the model formula, and the Quasi-Information Criterion (QIC) statistics of the fitted model.

The other model outputs (e.g., the parameter estimates and the sandwich standard errors) can be obtained using the `summary()` function in `R` as shown in Figure 5.6 below. The `class coordinates` section of the output presents parameter estimates for $\hat{\gamma}_{kj,m}$. It represents an estimate for the coordinate of the k-th category that belong to the j-th

response variable in dimension m. The class point restriction (i.e., $\gamma_{1j,m} - \gamma_{0j,m} = 1$) can be seen from the estimates. For example, for the first response (DYST) on the first dimension, its class point estimates are $\hat{\gamma}_{01} = 2.286$ and $\hat{\gamma}_{11} = 3.286$.

The estimates of regression coefficients per dimension (i.e., $\boldsymbol{\beta}_m$) then follows the class points as shown in Figure 5.6. These estimates show effect of the explanatory variables on each dimension, specifically on depression and anxiety dimensions. For example, $\hat{\beta}_{41} = 1.1286$ indicates that there is a strong positive association (p-value $= 0.0000$) between neuroticism and depression; similarly, with anxiety (i.e., $\hat{\beta}_{42} = 0.9856$). Whereas, extraversion has a moderate negative association with both dimensions, i.e., $\hat{\beta}_{51} = -0.4393$ with depression and $\hat{\beta}_{52} = -0.2782$ with anxiety.

The Pearson correlation between the two dimensions is also shown in the result, i.e., Corr$(\hat{\eta}_{i1}, \hat{\eta}_{i2}) = 0.98$. This result shows that there is a strong linear relationship between the positions of the subjects on the first dimension (η_{i1}) and the second dimension (η_{i2}).

The results displayed in Figure 5.6 are the default outputs by the `summary()` function. However, there are additional outputs (e.g., `npar` for number of parameters of the fitted model, etc) which are `glm`-like results, and they can be obtained by the `str(fit)` function in `R`.

```
# display summary of the fitted model result
summary(fit)

 Call:
mldm.fit(formula = mf, index = Index, resp.dim.ind = Z, data = NESDA,
    id = pident, scale = TRUE)
 Formula:
Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE +
    N + E
 Class Coordinates:
         dim1  dim2
gamma01 2.286 0.000
gamma11 3.286 0.000
gamma02 1.781 0.000
gamma12 2.781 0.000
gamma03 0.000 0.050
gamma13 0.000 1.050
gamma04 0.000 0.884
gamma14 0.000 1.884
gamma05 0.000 1.311
gamma15 0.000 2.311
[1] "Regression coefficients for Dimension 1"
     estimate san.se    wald      p
EDU1  -0.0277 0.0894  0.0957 0.7570
GEN1  -0.0136 0.1928  0.0049 0.9439
AGE1  -0.0440 0.0917  0.2309 0.6308
N1     1.1286 0.1187 90.3897 0.0000
E1    -0.4393 0.1109 15.6925 0.0001
```

```
[1] "Regression coefficients for Dimension 2"
     estimate san.se     wald      p
EDU2  -0.2192 0.0759   8.3427 0.0039
GEN2   0.2367 0.1631   2.1076 0.1466
AGE2  -0.0269 0.0735   0.1343 0.7140
N2     0.9856 0.0946 108.6077 0.0000
E2    -0.2782 0.0822  11.4533 0.0007
 Correlation among dimensions:
     dim1 dim2
dim1 1.00 0.98
dim2 0.98 1.00
 QIC:
[1] 2542.707
```

Figure 5.6: Summary of the depression-anxiety model fitted on the NESDA data.

The Clustered Bootstrap Method

The parameters in the MLD model are estimated using the GEE method. Thus, the Sandwich estimators are primarily used for hypothesis testing since the model-based standard errors are biased. The `mldm.fit()` function uses this procedure at the backend as a default estimation method.

The other alternative for obtaining the standard errors of model parameters in the MLD model is to apply a clustered bootstrap technique (Sherman & Le Cessie, 1997; De Rooij & Worku, 2012). In this case, the re-sampling procedure is applied on the subject (cluster) level so that the correlations between the measurements within each subject are retained.

The clustered bootstrap method is implemented in the **mldm** package. The `bootstrap` argument in the `mldm.fit()` function is used for this purpose. The function returns both the parametric and the non-parametric confidence intervals of the model parameters. Figure 5.7 shows application of the new argument, i.e., `bootstrap = 1000`. The number of replicates used here is also what is recommended in practice.

```
# fit the MLD model using Clustered Bootstrap method
fit_boot <- mldm.fit(formula=mf, index = Index, resp.dim.ind = Z,
           data = NESDA, id = pident, scale=TRUE, bootstrap=1000)
# fit_boot
```

Figure 5.7: Application of the Clustered Bootstrap method with the MLD model.

The `summary()` function with an additional argument can be used to obtain the clustered bootstrapped standard errors as shown in Figure 5.8 below.

Generally, layout of the model results are very similar to the one presented before in Figure 5.6. What makes this result different is that the standard errors are estimated differently (clustered bootstrap version) with a 95% confidence interval (CI) for the parameter estimates. By default, the confidence intervals are the parametric ones. The nonparametric confidence intervals can be obtained by specifying an additional argument in the `summary()` function, i.e., `boot.nonparam=TRUE` .

```
summary(fit_boot, bootstrap=TRUE)

 Call:
mldm.fit(formula = mf, index = Index, resp.dim.ind = Z, data = NESDA,
    id = pident, scale = TRUE, bootstrap = 1000)
 Formula:
Outcome | Outcome ~ EDU + GEN + AGE + N + E | EDU + GEN + AGE +
    N + E
 Class Coordinates:
         dim1  dim2
gamma01 2.286 0.000
gamma11 3.286 0.000
gamma02 1.781 0.000
gamma12 2.781 0.000
gamma03 0.000 0.050
gamma13 0.000 1.050
gamma04 0.000 0.884
gamma14 0.000 1.884
gamma05 0.000 1.311
gamma15 0.000 2.311
[1] "Regression coefficients for Dimension 1"
     estimate boot.se boot.wald      p boot.ll boot.ul
EDU1  -0.0277  0.0882    0.0983 0.7538 -0.2006  0.1452
GEN1  -0.0136  0.1990    0.0046 0.9457 -0.4035  0.3764
AGE1  -0.0440  0.0934    0.2224 0.6372 -0.2271  0.1390
N1     1.1286  0.1220   85.5033 0.0000  0.8893  1.3678
E1    -0.4393  0.1137   14.9184 0.0001 -0.6622 -0.2164
[1] "NB: The confidence intervals are the parameteric ones!"
```

```
[1] "Regression coefficients for Dimension 2"
     estimate boot.se boot.wald      p boot.ll boot.ul
EDU2  -0.2192  0.0724    9.1658 0.0025 -0.3610 -0.0773
GEN2   0.2367  0.1667    2.0172 0.1555 -0.0899  0.5634
AGE2  -0.0269  0.0748    0.1297 0.7188 -0.1736  0.1197
N2     0.9856  0.0973  102.6649 0.0000  0.7949  1.1762
E2    -0.2782  0.0832   11.1704 0.0008 -0.4413 -0.1150
[1] "NB: The confidence intervals are the parameteric ones!"
 Correlation among dimensions:
     dim1 dim2
dim1 1.00 0.98
dim2 0.98 1.00
 QIC:
[1] 2542.707
```

Figure 5.8: Summary of the depression-anxiety model fitted on the NESDA data using the Clustered Bootstrap method.

5.4.3 The Biplot for MLD Model

To enhance interpretation of the model the results of an MLD model can be graphically represented in a biplot (Gabriel, 1971; Gower & 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.

The `biplot()` function in the **mldm** package can be used to display the results of the MLD model in a biplot. Figure 5.9 shows the application of this function. The biplot for the depression-anxiety model is presented in Figure 5.10.

```
# biplot of the MLD model

biplot(fit)
```

Figure 5.9: Application of the `biplot()` function available in the **mldm** package.

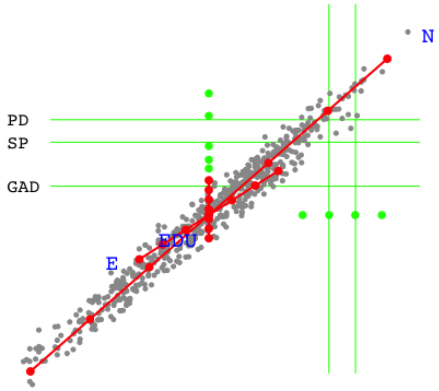

Figure 5.10: The biplot for depression-anxiety model fitted on the NESDA data.

5.4.4 Model Selection using QIC

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:

$$\text{QIC} = -2L(\boldsymbol{\theta}) + 2 \text{ trace}(\hat{\boldsymbol{\Omega}}_I^{-1}\hat{\mathbf{V}}_R), \tag{5.6}$$

where $\hat{\mathbf{V}}_R$ represents the 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.

The QIC_u fit statistics is implemented in the **mldm** package and its value can be extracted by specifying `fit$QIC`, where `fit` is an object of the fitted model.

Model selection for Dimensionality

For demonstration purpose we compare a unidimensional MLD model against a 2-dimensional MLD model fitted on the NESDA data with the same set of explanatory variables. For

dimensionality comparison, both the response indicator matrix and model formula should be redefined. Figure 5.11 shows specification of the indicator matrix for the unidimensional candidate model. For the 2-dimensional model, we use the same indicator matrix from the depression-anxiety model presented before in Figure 5.3.

```
# an indicator matrix for the unidimensional MLD model
Z1 <- matrix(c(1,1,1,1,1), 5, 1, byrow=FALSE)
Z1

[,1]
[1,]    1
[2,]    1
[3,]    1
[4,]    1
[5,]    1
```

Figure 5.11: Specification of an indicator matrix for candidate models with respect to dimensionality in the model.

The new model formula for the unidimensional candidate model is shown in Figure 5.12 where the explanatory variables are specified only in the first dimension. For the 2-dimensional MLD model we use the same model formula that was defined above for the depression-anxiety model (i.e., `mf`) in Figure 5.4.

```
# formula for the unidimensional model
mf1 <- Outcome ~ EDU + GEN + AGE + N + E
mf1 <- Formula(mf1)
```

Figure 5.12: Specification of model formula for a unidimensional MLD model.

Figure 5.13 displays the QIC fit statistics values for the two candidate models. We can conclude that the unidimensional MLD model fits the data well, although the two QIC values are very similar, i.e., $\text{QIC}^{1D} = 2541.235$ and $\text{QIC}^{2D} = 2542.707$. Note that only $N = 600$ subjects were used for fitting the candidate models. If all subjects (i.e., $N = 2938$) were included in the analysis, the 2-dimensional (depression-anxiety) model would fit the NESDA data better.

```
# fit the unidimensional MLD model
fit_dim1 <- mldm.fit(formula=mf1, index = Index, resp.dim.ind = Z1,
                     data = NESDA, id = pident, scale=TRUE)
# get the QIC value
fit_dim1$QIC

[1] 2541.235

# fit the 2-dimensionl MLD model
fit_dim2 <- mldm.fit(formula=mf, index = Index, resp.dim.ind = Z,
                     data = NESDA, id = pident, scale=TRUE)
# get the QIC value
fit_dim2$QIC

[1] 2542.707
```

Figure 5.13: Model selection in MLD model for dimensionality.

Model selection for Explanatory variables

For demonstration purpose, let us compare two candidate 2-dimensional MLD models that only differs on the explanatory variables. That is, (1) a depression-anxiety model with only the background variables (i.e., education, gender and age); and, (2) a depression-anxiety

model with both the background variables and two of the Big-5 personality traits (i.e., neuroticism and extroversion). The model formula for each candidate model is presented below in Figure 5.14.

```
# the first model, i.e., only the background variables
mf2a <- Outcome | Outcome ~ EDU + GEN + AGE |
        EDU + GEN + AGE
mf2a <- Formula(mf2a)


# the second model, i.e., with the background variables and
# two of the personality traits
mf2b <- Outcome | Outcome ~ EDU + GEN + AGE + N + E |
        EDU + GEN + AGE + N + E
mf2b <- Formula(mf2b)
```

Figure 5.14: Model formula structure of the candidate MLD models.

In Figure 5.15 the candidate models are fitted and the QIC results are obtained. The QIC fit statistics of the candidate models are then extracted by specifying `fit$QIC`. It can be concluded that the 2-dimensional MLD model with both the background variables and the two personality traits fits the data better than the model without (N, E) since it has a smaller QIC value, i.e., QIC $= 2542.707$.

```
# fit the first model
fit_dim2a <- mldm.fit(formula=mf2a, index = Index, resp.dim.ind = Z,
             data = NESDA, id = pident, scale=TRUE)
# get the QIC value for the first model
fit_dim2a$QIC

[1] 3056.799

# fit the second model
fit__dim2b <- mldm.fit(formula=mf2b, index = Index, resp.dim.ind = Z,
             data = NESDA, id = pident, scale=TRUE)
# get the QIC value for the second model
fit_dim2b$QIC

[1] 2542.707
```

Figure 5.15: Model selection in MLD model for explanatory variables.

5.5 Conclusion and Discussion

In this chapter we showed an application of the **mldm** package using a psychological dataset. The **mldm** package fits the newly proposed Multivariate Logistic Distance (MLD) model for analyzing multivariate binary data.

The `mldm.fit()` function in the **mldm** package supports two different estimation techniques for obtaining standard errors for model parameters in the MLD model, namely the Sandwich estimator from GEE method and the clustered bootstrap method. Using the `biplot()` function, one can easily produce a biplot for the fitted model. The `QIC` object returned from the `mldm.fit()` function can be used to compare candidate MLD

models. The QIC fit statistics is used to determine: (1) the dimensionality of the model, and (2) the structure of the explanatory variables.

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.