Advances in multivariate logistic distance modelling

Plaatsman, A.

Advances in Multivariate Logistic Distance Modeling

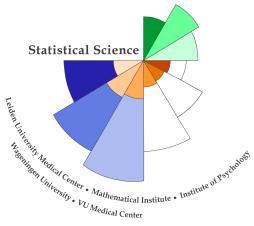
Author:
Amber Plaatsman
(S1885189)

Thesis advisors:
Prof. Dr. M.J. de Rooij
Dr. F.M.T.A. Busing

MASTER THESIS

Defended on June 11, 2019

Specialization: Statistical Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

Multivariate binary data are often collected in scientific fields such as psychology, economics and epidemiology. Worku and de Rooij (2018) proposed a marginal model for the analysis of this type of data in a distance framework: The multivariate logistic distance (MLD) model. Two different models were introduced by Worku and de Rooij: a restricted and an unrestricted MLD model. The interpretation of both models is clear, and a log-odds as well as a biplot representation can be used. In this work we proposed three extensions to the restricted model and showed the implications of the extensions for the interpretation of the corresponding biplot as well as for the log-odds. First, we showed how the model can be extended by making it possible for a response variable to belong to multiple dimensions. Consequently, the extended model can be used to examine other dimensionality structures compared to the original model. Second, we allowed for non-linear relationships of the predictor variables with the response variables in the model and therefore making the model more flexible. Finally, the dimensionality structure as well as the final predictor variables need to be selected. We showed how to use the prediction capability of a model as a selection criterion to select between competing models. This is a more versatile method to perform model selection, based on the bias-variance trade off, compared to the likelihood based criterion used in the original model. We fitted 16 variations of the model to an empirical data set to compare performance based on their prediction capability. All variations of the model can be estimated using standard statistical software for univariate models.

Contents

1

Introduction

1.1 General introduction

Worku and de Rooij (2018) proposed the *Multivariate Logistic Distance* (MLD) model to analyse multivariate binary responses in the presence of one or more predictor variables. This type of data is often collected in empirical sciences and over the years a variety of models for the analysis of this kind of data have been proposed. One way of dealing with multivariate binary responses is the *Generalized Linear Mixed-Effects Model* (GLMM: D. A. Anderson & Aitkin, 1985; Stiratelli, Laird, & Ware, 1984). GLMM is an extension of *Generalized Linear Models* (GLM) as proposed by Nelder and Wedderburn (1972). GLMM introduces a random effect in the model to capture within subject/cluster correlations. GLMM fully specifies the joint distribution of the responses and therefore inferences can be based on likelihood methods. However, estimation is computationally very difficult for non-normal multivariate data. In general, there is no simple closed-form solution to compute the likelihood (Chiou & Müller, 2005). This is why the software uses numerical integration techniques to compute the likelihood. Alternatively, Laplace-type approximations of the integrand can be used to obtain a closed-form expression of the approximated likelihood. For details about these techniques see Molenberghs and Verbeke (2005).

The absence of a multivariate distribution for binary responses renders the maximum likelihood estimation of the joint distribution of the responses computationally difficult. To address this problem Zeger and Liang (1986) proposed *Generalized Estimating Equations* (GEE), an estimation method modelling population averages of correlated categorical data. Contrary to GLMM, the model under GEE has a marginal and not a subject-specific interpretation. GEE can be seen as a multivariate extension of Wedderburn's (1974) quasi-likelihood method, and of Generalized Linear Models (Nelder & Wedderburn, 1972). An advantage of using GEE is that no assumptions are made for the joint probability of the correlated data. Another advantage of estimating a marginal mean model using GEE, is that the parameter estimators are consistent and asymptotically normal when the model for the mean response is correctly specified. This holds even when the dependence structure is misspecified (Halekoh, Højsgaard, & Yan, 2006; Zeger & Liang, 1986). The MLD model is part of a family of marginal models, like the GEE model, which

will be elaborated on further in this paper.

Comorbidity is a well-known phenomena in medical studies and behavioural studies and refers to the co-occurrence of two or more diseases or disorders at the same time. Especially in the field of mental disorders scientists are often interested in the underlying factors that are shared between disorders. Studies have consistency shown that disorders rarely occur in isolation of other disorders (e.g. Brown, Campbell, Lehman, Grisham, & Mancill, 2001; Spinhoven, van der Does, Ormel, Zitman, & Penninx, 2013). It is generally assumed that certain underlying factors, also referred to as dimensions or latent traits, are shared between disorders that tend to co-occur (Drost, Van der Does, van Hemert, Penninx, & Spinhoven, 2014). However, GEE cannot be used to access the dimensional structure of the response variables. To gain insight in these dimensions, latent variable models are often used. In general, latent variable models link continuous or categorical responses to unobserved latent traits. *Confirmatory factor analysis*, for example, can be used to test different theories about the number of latent variables and how the different response variables relate to these latent variables. Yet, for binary indicators, these models often make unverifiable distributional assumptions about the response variables and/or the underlying dimensions (e.g. Worku, 2018, Chapter 2).

The Multivariate Logistic Distance model of Worku and de Rooij (2018) can be used to access the dimensional structure of multivariate data without making distributional assumptions about the dimensions or the response variables. This is a clear advantage over latent variable models that are often used within empirical sciences to gain insight into the dimensional structure of the data. The MLD model has the possibility for dimension reduction as a form of regularization. Moreover, the model can be used to compare different theories about the dimensional structure of the data within one unified framework.

1.2 Problem statement

The MLD model has some advantages over other models for the analysis of multivariate data: The model can be used to access the dimensional structure of multivariate binary responses, as well as modelling the effect of the different predictor variables on the response variables; the interpretation of the model is clear, and both a log-odds as well as a biplot representation (Gabriel, 1971; Gower & Hand, 1995) can be given of the multivariate distance model. Although the MLD model has clear advantages, the model has some drawbacks too. The purpose of this thesis is to improve upon the current model by proposing three extensions to the model to overcome some of its pitfalls:

1. One of the advantages of latent variable models is the possibility of examining different theories about the dimensional structure of the model. The current MLD model can be used to access this dimensional structure as well, although the model is restricted. It only

allows for the assessment of dimensional structures in which every response variable relates to a single dimension. This limits the method in its flexibility to analyse comorbidity patterns in the data compared to latent variable models. We propose to extend the current model by making it possible for a response variable to relate to multiple dimensions.

2. The current model is part of a broad family of models in which the predictor variables have a linear relationship with the mean of the (transformed) response variables. However, because of the nature of the data it is possible that a linear relationship does not capture the true underlying pattern of the data. Therefore, we like to propose to extend the current MLD model by allowing for non-linear relationships of the predictor variables with the response variables, through the use of splines.

3. Finally, we desire to select a model that is parsimonious, but able to capture the underlying structure as well as the comorbidity patterns in the data. Therefore, the dimensionality structure as well as the final predictor variables need to be selected. We propose to use the prediction capability on independent test data to validate the model and to select between competing models. Comparing different candidate models in their ability to predict classes for independent test data will be used to select between these models.

1.3 Data and Software

We will use the Netherlands Study of Depression and Anxiety (NESDA; Penninx et al., 2008) data set to illustrate the proposed extensions of the model. The NESDA is an ongoing longitudinal cohort study in which data are collected to study personality traits, and their relationship to depressive and anxiety disorders. Furthermore, the studies interest lies in enhancing insight about the co-morbidity between depressive and anxiety disorders (Penninx et al., 2008; Spinhoven, de Rooij, Heiser, Smit, & Penninx, 2009). The NESDA study focuses on several disorders: Major Depresive Disorder (MDD), Dysthemia (DYST), General Anxiety Disorder (GAD), Panic Disorder (PD), and Social Phobia (SP). MDD and DYST are indicators of depression, whereas GAD, PD and SP are indicators of anxiety disorders. The personality traits in this data set are *extraversion* (E), *neuroticism* (N), *agreeableness* (A), *openness to experience* (O), and *conscientiousness* (C). Other (background) variables that are taken into account are age, years of education, and gender.

The sample used in this thesis consists of 1954 women and 984 men aged 18 through 65 with a mean age of 42 (*S.D.* = 13.1). The participants have had an average of 12.2 years of education (*S.D.* = 3.3). A comprehensive description of the design and sampling procedure of the NESDA is given by Penninx et al. (2008). The data consists of 1266 healthy people without a disorder and 1672 participants who suffered from one or more depressive or anxiety disorders. In this study the five disorders serve as response variables.

For implementation of the model throughout the thesis, the statistical software R (version 3.5.1) was employed (R Development Core Team, 2008). Additional used software packages will be named when used throughout the thesis. For reproducibility purposes, scripts for all the analyses are available upon request.

1.4 Organisation

The organisation of this thesis is as follows: In the next chapter an in-depth overview of the current model is documented. Thereafter, the next three chapters will address the limitations of the current model as well as the proposed solutions to each of them. The sixth chapter will provide an overview of the extended MLD model, the most important results, and a discussion about the proposed model.

2

Current MLD model

The MLD model is an extension of the Ideal Point Classification (IPC) model (de Rooij, 2009) to analyze multivariate binary responses. The IPC model is a classification model based on distances and is a simplification of ideal point discriminant analysis (IPDA) proposed by Takane, Bozdogan, and Shibayama (1987). The model was originally introduced to analyse univariate polytomous responses. In the IPC model coordinates representing the subjects and the classes are defined in a joint space. Let y_i denote the observed value for person i on a binary response variable, $y_i \in \{0, 1\}$ and let $\mathbf{x}_i$ be the observed values of person i on the p predictor variables, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$. When dealing with univariate binary responses, the probability for the first category given the predictor variables, $\pi(\mathbf{x}_i)$, can be defined in the IPC model as:

$$\pi(\mathbf{x}_i) = \frac{\exp[-0.5\,\delta_{1i}]}{\exp[-0.5\,\delta_{0i}] + \exp[-0.5\,\delta_{1i}]}, \tag{2.1}$$

with:

$$\begin{aligned} \delta_{1i} &= (\eta_i - \gamma_1)^2 \text{ and} \\ \delta_{0i} &= (\eta_i - \gamma_0)^2, \end{aligned} \tag{2.2}$$

where δ_{0i} and δ_{1i} are the squared Euclidean distances between the position of person i, and the points representing the two categories. By η_i the coordinate of the position for subject i is denoted. This coordinate is a linear combination of the predictor variables, i.e. $\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}$. By γ_c the coordinate for category c is denoted ($c \in \{0, 1\}$). The smaller the relative Euclidean distance between the person and a class point,the larger the probability that this person belongs to that class.

An univariate logistic regression model can be expressed as an unidimensional IPC model, as defined in (2.1). Before we discuss the extended MLD model, we will recapitulate the fundamentals of logistic regression and the relationship between the univariate logistic regression model and the unidimensional IPC model.

2.1 Logistic Regression

In the same way normal regression models are based on the Gaussian distribution, a binary response model is derived from a Bernouilli distribution. The Bernouilli *probability mass function* (pmf) of Y over possible outcomes y can be expressed as:

$$P(Y = y|\pi) = \pi^y(1-\pi)^{1-y} \qquad \text{for } y \in \{0, 1\}, \tag{2.3}$$

where y denotes the value on a binary response variable and π denotes the probability of $Y = 1$. Nelder and Wedderburn (1972) formulated *Generalized Linear Models* (GLMs), an extension of ordinary linear regression. GLMs are a class of models that allow exponentially distributed response variables to be linearly related to the predictor variables via a monotonic and differentiable link function, $g(.) = \eta^*$. The logistic regression model is part of the family of GLMs and is one of the most commonly used statistical methods for the analysis of binary response data (Hilbe, 2009). In logistic regression the Bernouilli distribution for every observation y_i is rewritten as

$$f(y_i; \pi(\mathbf{x}_i)) = \exp\left\{ y_i \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) + \log(1-\pi(\mathbf{x}_i))\right\}, \tag{2.4}$$

with:

$$\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \eta_i^* = \alpha + \mathbf{x}_i^T\boldsymbol{\beta}^*. \tag{2.5}$$

Using the inverse of this relationship we can write π in terms of η^*:

$$\pi(\mathbf{x}_i) = \frac{e^{\eta_i^*}}{1+e^{\eta_i^*}}. \tag{2.6}$$

One of the reasons why the logistic link function gained popularity, is because $\log\left(\frac{\pi}{1-\pi}\right)$ has a clear and nice interpretation in itself: The logistic transformation, also called the logit transformation or the log-odds, is the logarithm of the odds and maps the probability with range $[0, 1]$ to a scale with range $[-\infty, \infty]$. The odds are an important concept within probabilistic models and are an expression of the relative probability of a certain outcome over another outcome.

Within the GLM framework the regression parameters, α and $\boldsymbol{\beta}^*$, are estimated using *maximum likelihood* (ML) estimation. ML estimation is a technique to estimate the most likely values of the parameters, given the observed data. The idea of maximum likelihood estimation is to find a set of values for the parameters that maximize the likelihood. To find the maximum likelihood estimates we need to differentiate the (log) likelihood with respect to the parameters, set the derivatives equal to zero, and solve the obtained system of equations. However, as for most models in the GLM framework, the system of equations has no analytical solution. Therefore, it must be solved by numerical methods. Many optimization algorithms are available for such

problems.The method frequently used by `R` and other statistical packages like `SAS` to obtain the regression parameters of the logistic regression model, is the Newton-Raphson method. The Newton-Raphson algorithm is an iterative approach and begins with an initial guess for the regression coefficients. The updated coefficients are found by a second-order Taylor expansion evaluated at their current values. This process continues until convergence. In addition to an estimate of the regression parameters of our model, we can also obtain the standard errors of the coefficients by means of the Newton-Raphson method.

2.2 Relationship logistic regression with the IPC model

The IPC model (2.1) as defined by de Rooij (2009) can be expressed as an univariate logistic regression model. When writing the IPC model (2.2) in terms of the log-odds, this gives us the following expression:

$$\begin{aligned}
\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) &= 0.5\,\delta_{0i} - 0.5\,\delta_{1i} \\
&= 0.5\,(\eta_i - \gamma_0)^2 - 0.5\,(\eta_i - \gamma_1)^2 \\
&= 0.5\,\eta_i^2 - \gamma_0\eta_i + 0.5\,\gamma_0^2 - 0.5\,\eta_i^2 + \gamma_1\eta_i - 0.5\,\gamma_1^2 \\
&= \eta_i(\gamma_1 - \gamma_0) + 0.5\,(\gamma_0^2 - \gamma_1^2) \\
&= (\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})(\gamma_1 - \gamma_0) + 0.5\,(\gamma_0^2 - \gamma_1^2)
\end{aligned} \tag{2.7}$$

From equation (2.7) we can see the relationship with the univariate logistic model:

$$\begin{aligned}
\alpha &= \beta_0(\gamma_1 - \gamma_0) + 0.5\,(\gamma_0^2 - \gamma_1^2) \\
\boldsymbol{\beta}^* &= \boldsymbol{\beta}(\gamma_1 - \gamma_0)
\end{aligned} \tag{2.8}$$

To identify the IPC model, we need to define both the scale and the origin. This can be done by imposing a constraint on the coordinates of the category points. For example, by imposing the constraint that $\gamma_0 = 0$ and $\gamma_1 = 1$, both the scale and the origin are fixed. Other choices can be made regarding the identification of the model, we can for example center the data around zero to define the origin, and set the distance between the two categories of the response variable to one (i.e. $\gamma_1 - \gamma_0 = 1$) to define the scale. By imposing the latter constraints, the relationship with the univariate logistic model becomes: $\alpha = 0.5\,(\gamma_0^2 - \gamma_1^2)$ which can, because of the identifiability constraint, be simplified to $-\gamma_1 + 0.5$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}$. Hence, we can obtain the estimates of the univariate model using standard statistical software for logistic regression.

The regression coefficients, $\boldsymbol{\beta}$, represent the effect on the log-odds, eg. when x_1 increases with one unit, the log-odds of membership to category 1 versus the baseline category 0 changes by β_1. Note that henceforth, we will assume that our predictor variables are centered and scaled, i.e. the origin is fixed with $\beta_0 = 0$.

2.3 Multivariate extension

Worku and de Rooij (2018) extended the IPC model for a single response variable into a model with multiple binary response categories. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$ be the multivariate responses of subject i on the J response variables. When an unidimensional space is used, the different points representing the categories of the response variables all lie on the same line. When multiple dimensions are used to represent the class points, the definition of the distance becomes:

$$\delta(\boldsymbol{\eta_i}, \boldsymbol{\gamma_{cj}}) = \sum_{m=1}^{M} (\eta_{i,m} - \gamma_{cj,m})^2, \tag{2.9}$$

where $m = 1, \ldots, M$ are the dimensions, $\eta_{i,m}$ is the coordinate representing the point of subject i on the dimension m and $\gamma_{cj,m}$ is the coordinate for category $c \in \{0, 1\}$ of response variable j on dimension m. The coordinate representing the person on this dimension is a linear combination of the predictor variables on dimension m: $\eta_{i,m} = \mathbf{x}_i^T \boldsymbol{\beta}_m$, with $\boldsymbol{\beta}_m$ being a vector with regression coefficients of dimension m. In the model of Worku and de Rooij (2018), each response variable belongs to only one dimension. Let the probability for subject i for the first category of response variable j given the predictor variables be written as:

$$\pi_j(\mathbf{x}_i) = \frac{\exp[-0.5\, \delta(\boldsymbol{\eta_i}, \boldsymbol{\gamma_{1j}})]}{\exp[-0.5\, \delta(\boldsymbol{\eta_i}, \boldsymbol{\gamma_{0j}})] + \exp[-0.5 \delta(\boldsymbol{\eta_i}, \boldsymbol{\gamma_{1j}})]}. \tag{2.10}$$

Worku and de Rooij (2018) showed that the log-odds representation of the univariate model (2.7) can be extended into the log-odds representation of the multivariate distance model and can be expressed as:

$$\log\left(\frac{\pi_j(\mathbf{x}_i))}{1 - \pi_j(\mathbf{x}_i)}\right) = \sum_{m=1}^{M} \left\{ \mathbf{x}_i^T \boldsymbol{\beta}_m (\gamma_{1j,m} - \gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2 - \gamma_{1j,m}^2) \right\}. \tag{2.11}$$

When, in a multidimensional model, all response variables belong to a single dimension, like in the model of Worku and de Rooij (2018), the log-odds representation of the distance model can be further simplified: For a response variable j that does not belong to dimension m, $\gamma_{0j,m}$ and $\gamma_{1j,m}$ equal zero. Therefore, only the part of the dimension to which response variable j belongs, will contribute to the log-odds and hence equation (2.11) simplifies to a single equation, like (2.7), instead of a sum over multiple dimensions.

Worku and de Rooij (2018) proposed a *restricted* and an *unrestricted* variant of the MLD model. The former refers to a model in which the distance between the two categories of every response variable are set to be equal. The predictor variables in this model discriminate equally well between the categories of all response variables. The latter refers to a model without the equality constraint. All the extensions proposed in this manuscript concern the restricted model. The number of independent parameters estimated in the unrestricted MLD model equals $[(J - M) \times 2 + M \times p]$, in which p is the number of predictor variables. When fitting the restricted model, only $[M \times p + J]$ parameters have to be estimated.

2.4 Estimation MLD model

The absence of a distribution for multivariate binary responses that accounts for the dependence, renders the maximum likelihood estimation of the joint distribution of the responses computationally difficult. The parameters in the MLD model are therefore calculated by maximizing the *quasi-likelihood* (Wedderburn, 1974). For multivariate binary data, the log-likelihood under the assumption of independence and the quasi-likelihood function are identical. That is,

$$\ell\left(\boldsymbol{\beta}; \mathbf{y}\right) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left(y_{ij} \log\left(\frac{\pi_j(\mathbf{x}_i)}{1 - \pi_j(\mathbf{x}_i)} \right) + \log\left(1 - \pi_j(\mathbf{x}_i)\right) \right). \tag{2.12}$$

When dealing with multivariate data it is unreasonable to assume that responses are independent, because the observations of the same participant tend to be correlated. Parameter estimates obtained by maximizing the quasi-likelihood will still be consistent (under mild conditions), but standard errors derived from the Hessian matrix when fitting the quasi-likelihood will generally be incorrect (Sherman & Cessie, 1997; Wedderburn, 1974; Zeger & Liang, 1986).

By restricting the distance between the two categories of every response variable to be equal, the MLD model can be fitted using standard statistical software to fit logistic regression models. Correct standard errors can then be obtained by applying a clustered bootstrap method as proposed by Sherman and Cessie (1997). This method is based on a bootstrap procedure in which the correlation structure between the multivariate responses is retained. To employ this method the `ClusterBootstrap` package in `R` could, for example, be utilized (Deen & de Rooij, 2019).

Alternatively the restricted model can be fitted using *Generalized Estimating Equations* (GEE) as proposed by Zeger and Liang (1986). GEE extends the generalized linear model to allow for the analysis of correlated data, such as clustered data. The GEE method accounts for the dependency within clustered data, without fully specifying the likelihood or the dependence structure. Instead of optimizing a likelihood function, the parameters of the model are obtained by iteratively solving *estimation equations*. GEE is therefore not a likelihood-based approach, it is an estimation method in which the quasi-likelihood is constructed from the estimating equations (Pan, 2001). The estimation equations are a marginal formulation of the likelihood function that

use a working correlation matrix to adjust for the dependency within clusters. To obtain robust standard errors Zeger and Liang (1986) proposed a sandwich estimator. Originally the sandwich estimator was proposed by Huber (1967) and White (1982), Zeger and Liang extended the idea to longitudinal data. The sandwich estimator adopts a "working" assumption about the association structure of the data. Zeger and Liang showed that asymptotically correct standard errors are obtained by means of the sandwich estimator. This is true regardless of the true correlation structure of the data as long as the mean structure of the model is correctly specified and the data are sufficiently large. Under the *independence* working assumption, the estimated parameters equal those obtained with logistic regression. However, the standard errors will differ from the standard errors obtained by logistic regression (Molenberghs & Verbeke, 2005).

2.4.1 Implementation

Worku and de Rooij (2018) proposed an *unrestricted* and a *restricted* MLD model. The unrestricted model can be fitted in its own right. Because all the proposed extensions are related to the restricted model, fitting the unrestricted model is beyond the scope of the present thesis. By setting the distance between the two categories of all response variable to be equal, the restricted MLD model becomes equivalent to estimating a marginal model for multivariate binary data using Generalized Linear Models or Generalized Estimating Equations. Hence, standard statistical software for these methods can be utilized to fit the model. Alternatively, the `mldm` package can be employed to fit the restricted model (Worku, 2018).

For the implementation of the MLD model Worku and de Rooij (2018) used a technique proposed by Wright (1998). Wright showed that multivariate models may be estimated using software for univariate models (such as GLM and GEE). When using this technique, one has to modify the structure of the data: The multivariate responses need to be reordered in a vector and the matrix with predictor variables needs to be reorganized. We will illustrate Wright's method by means of an example: Consider three response variables for all n observations, we can restructure the multivariate responses into a vector Y. Suppose we want the first two response variables to be represented on the first dimension and the third response variable to be represented on second dimension. Let $\mathbf{Z}$ be the matrix that specifies to which dimension the response variables belong, with the rows representing the response variables and the columns representing the dimensions.

Further, let there be two predictor variables represented in predictor matrix **X**. We get:

$$Y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ \vdots \\ y_{n1} \\ y_{n2} \\ y_{n3} \end{bmatrix}, \qquad \mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \text{and} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}. \tag{2.13}$$

The MLD model may be estimated using the design matrix **S** which is obtained by taking the Kronecker product between the response indicator matrix **Z** and the predictor matrix **X** and concatenate it with an 3x3 identity matrix for each subject, i.e.

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & x_{11} & x_{12} & 0 & 0 \\ 0 & 1 & 0 & x_{11} & x_{12} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & x_{11} & x_{12} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & x_{n1} & x_{n2} & 0 & 0 \\ 0 & 1 & 0 & x_{n1} & x_{n2} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & x_{n1} & x_{n2} \end{bmatrix}. \tag{2.14}$$

Matrix **S** shows that the restricted model of Worku and de Rooij (2018) implies the predictor variables to discriminate equally well for all response variables belonging to the same specific dimension. It could be argued that such an assumption is not always justified. We will elaborate on this further in Chapter 5.

The predictor matrix **S** can be used together with response vector Y, to fit the model with, for example, the function `glm()` in `R`. We will obtain an intercept for each of the response variables and a coefficient per dimension for each of the predictor variables. From the three intercepts we derive the coordinates of the class points for response variables for the MLD model (see Equation 2.8). These are the coordinates of the class points for the dimension they belong to. Furthermore, the other four obtained regression parameters of the GLM correspond directly to the regression coefficients, $\boldsymbol{\beta}$, of the MLD model (see Equation 2.8).

2.5 Model selection

The assumption of which response variable belongs to which dimension has a crucial impact on the interpretation of the model. Therefore, not only do the predictor variables need to be selected, but also the dimensionality structure of the model has to be determined. Within the likelihood

framework, information criteria are typically used to compare and select between competing models. Methods like *Akaike's Information Criterion* (AIC) (Akaike, 1974) balance between the goodness of fit and the simplicity of the model by penalizing the fit for the number of estimators in the model.

Pan (2001) proposed an extension of the AIC criterion to select the best fitting model when using GEE: the *quasi-likelihood under the independence model criterion* (QIC). QIC is a modification of AIC, in which the likelihood is replaced by the quasi-likelihood and the penalty term is adjusted accordingly. Worku and de Rooij (2018) used QIC_u, a simplified version of QIC, to determine the dimensionality of the model. Furthermore, variable selection in the original MLD model is based on a Wald test performed on the regression parameters with standards errors obtained with the sandwich estimation method or the clustered bootstrap method.

2.6 Visualisation of the model

The MLD model is an appealing model since the interpretation based on distances is noticeably intuitive for classification purposes. In addition, a log-odds as well as a visual representation of the model can be given. We can visualize the MLD model by means of a *biplot* which can be seen as the multivariate equivalent of an ordinary scatterplot (Blasius, Eilers, & Gower, 2009; Gabriel, 1971). Traditionally the elements of a biplot are a set of axes representing the predictor variables and a set of points representing subjects, visualized in two or three dimensional space (Gower & Hand, 1995). To visualize the MLD model, another component needs to be included in the biplot, the categories of the response variables. Before introducing the biplot that accompanies the original model, we will first discuss the interpretation of the variable axes and the response space.

2.6.1 Variable axes

In the biplot accompanying the original model, the variable axes are derived by multiplying the obtained regression coefficients of the variable to be plotted (one coefficient per dimension) with a vector containing values ranging from -3 to 3, increasing with increments of one. We assume most of the scores on the predictor variables to be within three standard deviations of the mean. Because the predictor variables are centered and scaled this corresponds to a score between -3 and 3, hence the choice for the values in our multiplication vector. The obtained coordinates are connected and form the variable axis for the given variable. The coordinates that are used to form the variable axis can be interpreted as the coordinates of different subjects with scores ranging from -3 to 3 on this variable and a score of zero on the other predictor variables. The relative length of a variable axis projected on the dimension corresponds to the strength of the effect of the variable on the response variables belonging to this dimension. Let us explain this

by means of an example: Figure 2.1 visualizes class points of a response variable belonging to a single dimension, $D1$, and the position of subject ρ and τ in two dimensional Euclidean space. For notational simplicity, henceforth, in this example we will write γ_0 for $\gamma_{0j,1}$ an γ_1 for $\gamma_{1j,1}$ to denote the coordinates of the class points of variable j on $D1$. The two subjects differ one unit in their scores on variable X. Let ρ' and τ' be the orthogonal projections of the two subjects on the first dimension. These projections equal, by definition, the subjects coordinate on the first dimension, e.g. $\rho' \equiv \eta_{\rho 1}$.

The squared Euclidean distance in M-dimensional space between two points is defined as the sum of the squared differences per dimension. This follows directly from the Pythagorean formula. Thus, the squared Euclidean distance between ρ and γ_0 equals the sum of the squared line segment between ρ and ρ' and the squared line segment between ρ' and γ_0. The log-odds of subject ρ can be written as:

$$\log\left(\frac{\pi(X_\rho)}{1-\pi(X_\rho)}\right) = \frac{1}{2}\,\delta(\rho,\gamma_0) - \frac{1}{2}\,\delta(\rho,\gamma_1), \tag{2.15}$$

which simplifies to $\frac{1}{2}(\rho'-\gamma_0)^2 - \frac{1}{2}(\rho'-\gamma_1)^2$, since the response variable only belongs to the first dimension (Worku & de Rooij, 2018). When the constraint that $\gamma_0 - \gamma_1 = 1$ holds, the above can be rewritten as $\frac{1}{2}(\gamma_1^2 - \gamma_0^2) - \rho'$. Hence, the factor by which the log-odds of person ρ and τ differ, equals the difference of their orthogonal projections on $D1$. This is illustrated in Figure 2.1 by a blue arrow; when X increases with one unit the log-odds, change by β_x.

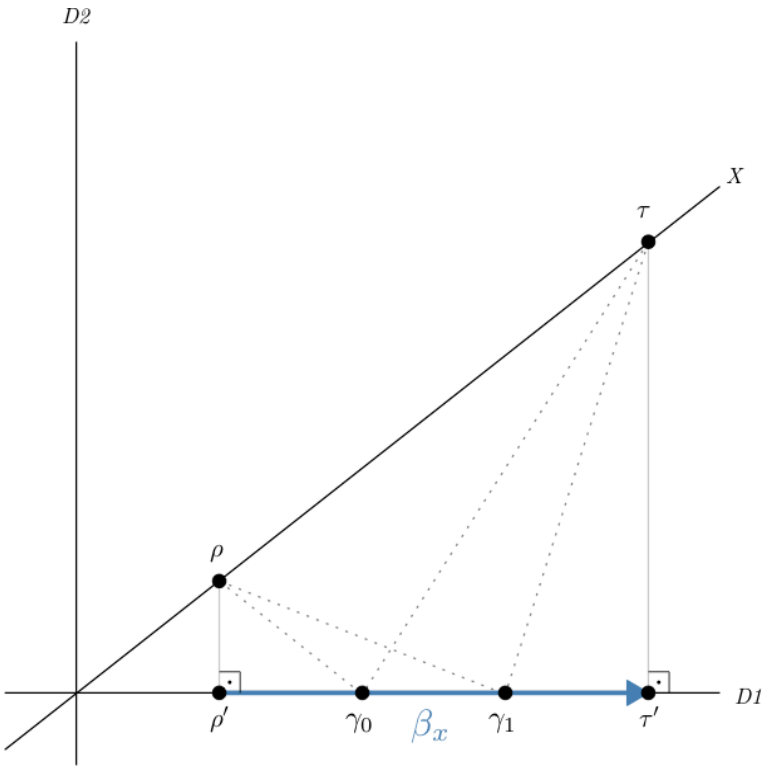


Figure 2.1 Projection of variable axis X on the first dimension.

2.6.2 Response space

Figure 2.2 shows the class points of the response variables of the fitted MLD model on the NESDA data in two-dimensional Euclidean space. Since the response variables belong to only one dimension in the model of Worku and de Rooij (2018), the coordinates of the response variables are zero for the other dimension. The category points of MDD, GAD and DYST are positioned on the first dimension. The category points of SP and PD are positioned on the second dimension. In addition, decision boundaries are displayed in Figure 2.2. A decision boundary is the set of points for which the log-odds of a response variable are zero. Because the Euclidean distance is used in the MLD model, the decision boundary equals a line orthogonal to the dimension on which γ_{0j} and γ_{1j} are positioned, going through the point halfway between γ_{0j} and γ_{1j}. The decision lines partition the space of Figure 2.2 into regions, each of which are representing an area in which the predicted odds are in favor of a specific response profile. The class points together with the decision boundaries compose the *response space*, as visualized in Figure 2.2. From the plot we can see that the response patterns account for comorbidity in the data. Each region shows a disorder profile; the bottom left region represents the absence of disorders, while the top right region represents the comorbidity of all five disorders.

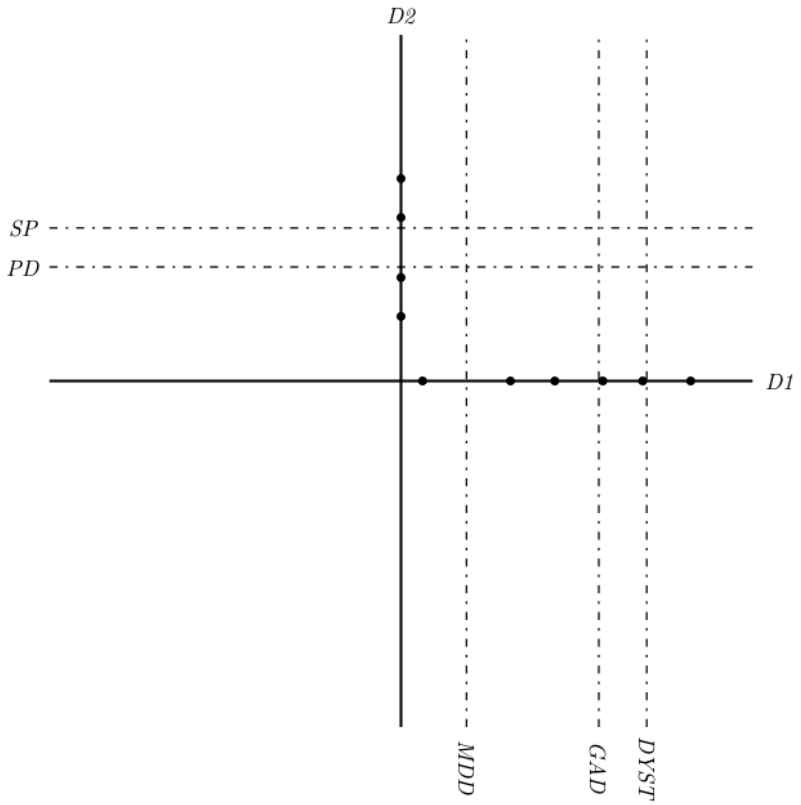


Figure 2.2 Response space of the MLD model fitted on the NESDA data with a two-dimensional structure.

2.6.3 Biplot

Figure 2.3 shows a biplot of the MLD model fitted on the NESDA data, similar to the model fitted in Figure 2.2. The positions of the subjects are constructed by taking a linear combination of their scores on the predictor variables and are included in the plot as points. Most of the subjects are positioned in the bottom left response region, corresponding to a most probable response profile without any disorders. From the plot we see that the two dimensions are positively correlated; subjects that have a higher probability of having one or more disorders on the first dimension also generally have a higher probability of having one or more disorders positioned on the other dimension.

Figure 2.3 only shows the variable axes of the final model as proposed by Worku and de Rooij (2018), i.e. *education, neuroticism, extraversion* and *conscientiousness*. On the variable axes, markers are placed at values ranging from -3 to 3, with increments of one. In addition labels are included at the positive side of the variable axes. As mentioned in Section 2.6.1, the relative length of a variable axis projected on the dimension corresponds to the strength of the effect of the predictor variable on the disorders belonging to this dimension. From the plot we can see that *neuroticism* has a large effect on disorders associated with both dimensions, while, for example, conscientiousness only has a minor effect on the disorders positioned on the second dimension.

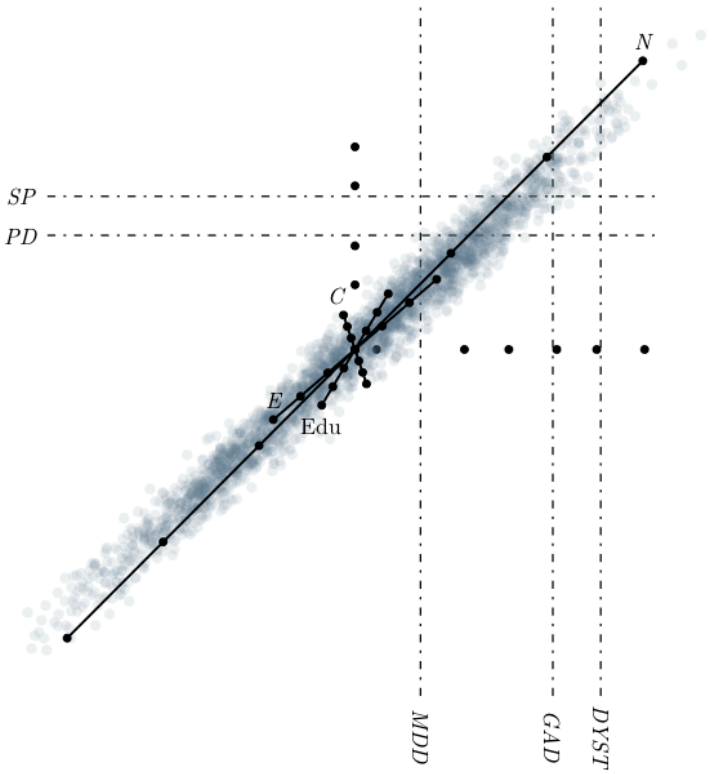


Figure 2.3 Biplot of the MLD model fitted on the NESDA data with a two-dimensional structure.

3

Response variable on multiple dimensions

The MLD model as proposed by Worku and de Rooij (2018) can be used to represent different dimensional structures. Moreover, the model can be used to compare different structures in an unified framework. In the current MLD model each response variable belongs to a single dimension. However, it should be taken into account that a response variable can relate to multiple dimensions, because of the nature of the data. Let us illustrate this phenomenon by means of an example based on the NESDA data set: Previous research suggests that comorbidity patterns of common mental disorders can be reflected using different structures, that is:

- A two-dimensional 'distress-fear' (d/f) structure with one dimension representing distress [MDD, GAD & DYST] and the other dimension representing fear [PD, & SP] (Beesdo-baum et al., 2009; Kotov, Gamez, Schmidt, & Watson, 2010; Krueger, 1999; Spinhoven, Penelo, De Rooij, Penninx, & Ormel, 2014).
- A two-dimensional 'depression-anxiety' (d/a) structure with one dimension representing depression [MDD & DYST] and the other dimension representing anxiety [PD, SP & GAD] (Penninx et al., 2008; Spinhoven et al., 2009, 2013).
- An 'uni-dimensional' structure with all five disorders represented on a single dimension (Penninx et al., 2008).

The MLD model can be used to represent and compare all of the structures presented above. Because of the discrepancy between the first two theories one could, for example, want to examine a model with a dimensional structure in which GAD is represented on both the first and the second dimension. This is currently not possible with the model, because of the restriction posed on the response variables. The following chapter provides a describtion of an extension of the current MLD model by making it possible for a response variable to belong to multiple dimensions. In addition, the impact of this extension on the log-odds and the biplot representation of the model will be discussed.

3.1 Extension current model

When employing the MLD model, it is assumed that the *logit* transformation of the response variables have a linear relationship with the predictor variables, as described in the previous chapter. The logit transformation, or the log-odds, are therefore an important concept when interpreting the MLD model. Like in the univariate case the log-odds of the multivariate case can be written as:

$$0.5\,\delta(\boldsymbol{\eta_i}, \boldsymbol{\gamma_{0j}}) - 0.5\,\delta(\boldsymbol{\eta_i}, \boldsymbol{\gamma_{1j}}), \tag{3.1}$$

where $\boldsymbol{\eta_i}$ denotes the coordinates representing the position of subject i. The position of the coordinate per dimension is determined by a linear combination of the predictor variables, e.g. $\eta_{im} = \mathbf{x}_i^T \boldsymbol{\beta_m}$, where η_{im} is the coordinate representing the point of subject i on dimension m. As stated in (2.11), the log-odds representation of the MLD model can be denoted as:

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

When all response variables relate to a single dimension, like in the model of Worku and de Rooij (2018), the log-odds representation simplifies to a single equation instead of a sum over multiple dimensions (see page 8). However, if we allow response variables to belong to multiple dimensions, the former simplification does not apply. Equation (3.2) shows the log-odds representation of the MLD model for which $\gamma_{1,m} - \gamma_{0,m} = 1 \;\forall\; m$ holds, without the restriction that response variables belong to a single dimension. The regression coefficients $\boldsymbol{\beta}_m$ represent the effect on the log-odds.

$$\begin{aligned}
\log\left(\frac{\pi_j(\mathbf{x}_i)}{1-\pi_j(\mathbf{x}_i)}\right) &= \sum_{m=1}^{M} \left\{\mathbf{x}_i^T \boldsymbol{\beta}_m(\gamma_{1j,m} - \gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2 - \gamma_{1j,m}^2)\right\} \\
&= \sum_{m=1}^{M} \left\{\mathbf{x}_i^T \boldsymbol{\beta}_m + 0.5((\gamma_{1j,m} - 1)^2 - \gamma_{1j,m}^2)\right\} \\
&= \sum_{m=1}^{M} \left\{\mathbf{x}_i^T \boldsymbol{\beta}_m + 0.5(\gamma_{1j,m}^2 - 2\gamma_{1j,m} + 1 - \gamma_{1j,m}^2)\right\} \\
&= \sum_{m=1}^{M} \left\{\mathbf{x}_i^T \boldsymbol{\beta}_m - \gamma_{1j,m} + 0.5\right\}.
\end{aligned} \tag{3.2}$$

When denoted as a logistic regression model, we write:

$$\begin{aligned} \alpha &= 0.5\,M - \sum_{m=1}^{M} \gamma_{1j,m} && \text{and} \\ \beta_X^* &= \sum_{m=1}^{M} \beta_{X,m} && \text{for } X = 1, \ldots, p, \end{aligned} \tag{3.3}$$

where regression coefficient β_X^* is the linear effect of variable X. The regression coefficients, $\beta_{X,1}, \beta_{X,2} \ldots \beta_{X,m}$, represent the effect on the log-odds; when X increases with one unit, the log-odds of membership to category 1 versus the baseline category 0 change by $\sum_{m=1}^{M} \beta_{X,m}$.

Because $\alpha = 0.5\,M - \sum_{m=1}^{M} \gamma_{1j,m}$, the multivariate logistic distance model in Equation (3.2) is not uniquely identified. Let us illustrate this by means of an example. Suppose we have a response variable j that belongs to two dimensions and one predictor variable. The log-odds representation of the model becomes:

$$\begin{aligned} \log\left(\frac{\pi_j(X_i)}{1-\pi_j(X_i)}\right) &= \sum_{m=1}^{2} \left\{\beta_m X_i(\gamma_{1j,m} - \gamma_{0j,m}) + 0.5(\gamma_{0j,m}^2 - \gamma_{1j,m}^2)\right\} \\ &= (\beta_1 + \beta_2)X_i - \gamma_{1j,1} - \gamma_{1j,2} + 1 \end{aligned}$$

Hence, the logistic regression representation equals: $\alpha = 1 - \gamma_{1j,1} - \gamma_{1j,2}$ and $\beta^* = \beta_1 + \beta_2$. Suppose we fit the model by means of standard GLM software and obtain a value for α, for example $\alpha = -3$. Figure 3.1 visualizes α as a line in two dimensional space for which $\alpha = -3$ holds. The coordinate of category one on the first dimension, $\gamma_{1j,1}$, could be 3 when the coordinate of category one on the second dimension, $\gamma_{1j,2}$, equals 1, yet $\gamma_{1j,1}$ could also be 2 when $\gamma_{1j,2} = 2$. Therefore, additionally to the equality constraint another identifiable constrained is needed to identify this model. By imposing the constrained that $\gamma_{1j,1} = \gamma_{1j,2}$, the two dimensional MLD model is identified for response variables lying on multiple dimensions. Although many identifiable constraints can be proposed, we choose to impose this particular constraint for visualization purposes. When the coordinates of the two categories of a response variable are equal for both dimensions, their coordinates can be combined to form the coordinates lying on a projection line, P, as visualized in Figure 3.1. P is orthogonal to the line visualizing α, it goes through the origin and the angle between the projection line and the two dimensions equals 45 degrees. Hence, the distance between the coordinates of the response categories on the different dimensions and the projection line are in the ratio $1 : 1 : \sqrt{2}$ which follows from the Pythagorean theorem.

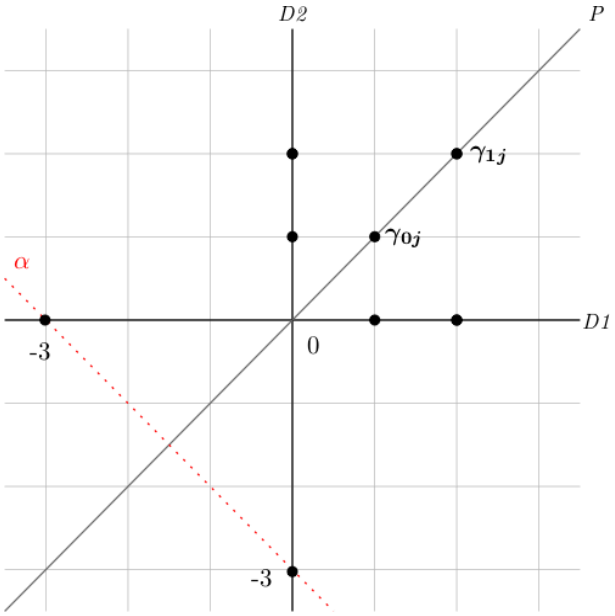


Figure 3.1 Visualization of a response variable on two dimensions with $\alpha = -3$.

3.2 Estimation extended model

The model can be estimated, employing standard statistical software to fit GLM or GEE models, by applying the same method as described in Section 2.4.1.

Compared to the original MLD model, an adapted response indicator matrix is used when fitting the extended MLD model. Let $\mathbf{Z}$ be the response indicator matrix used to fit the original model with a two-dimensional 'distress-fear' structure on the NESDA data. Furthermore, let $\mathbf{Z}'$ be the response indicator matrix of the extended model, where GAD is positioned on both dimensions. We get:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Z}' = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. \tag{3.4}$$

The MLD model may be estimated using a design matrix $\mathbf{S}$ which is obtained by taking the Kronecker product between the response indicator matrix $\mathbf{Z}'$ and the predictor matrix $\mathbf{X}$, $\mathbf{Z}' \otimes \mathbf{X}$, and concatenate it with an $J \times J$ identity matrix for each subject, such that

$$\mathbf{S} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & x_{11} & x_{12} & \dots & x_{1p} & 0 & 0 & \dots & 0 \\
0 & 1 & 0 & 0 & 0 & x_{11} & x_{12} & \dots & x_{1p} & 0 & 0 & \dots & 0 \\
0 & 0 & 1 & 0 & 0 & x_{11} & x_{12} & \dots & x_{1p} & x_{11} & x_{12} & \dots & x_{1p} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & x_{11} & x_{12} & \dots & x_{1p} \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & x_{11} & x_{12} & \dots & x_{1p} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & 0 & 0 & x_{n1} & x_{n2} & \dots & x_{np} & 0 & 0 & \dots & 0 \\
0 & 1 & 0 & 0 & 0 & x_{n1} & x_{n2} & \dots & x_{np} & 0 & 0 & \dots & 0 \\
0 & 0 & 1 & 0 & 0 & x_{n1} & x_{n2} & \dots & x_{np} & x_{n1} & x_{n2} & \dots & x_{np} \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & x_{n1} & x_{n2} & \dots & x_{np} \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & x_{n1} & x_{n2} & \dots & x_{np}
\end{bmatrix}. \tag{3.5}$$

3.3 Effect on the interpretation of the biplot

The extension of the model does not have implications for the interpretation of the points representing the subjects in the corresponding biplot. Yet, the interpretation of the variable axis, in relation to the response variable on projection line P, changes. Furthermore, the response space of the model changes when we allow response variables to pertain to multiple dimensions. To enhance understanding about the interpretation of the biplot of the extended model, we will first discuss the interpretation of the variable axes of the biplot accompanying the extended model.

3.3.1 Variable axes extended model

Let us consider a two dimensional joint space, in which the class points for response variable j and the variable axis of variable X are defined. We assume the following conditions:

1. The response variable j, belongs to the two dimensions.
2. The following constraints hold: $\gamma_{1j,m} - \gamma_{0j,m} = 1$ and $\gamma_{0j,1} = \gamma_{0j,2}$.

Moreover the class coordinates for response variable j have the following form:

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{0j,1} & \gamma_{0j,2} \\ \gamma_{1j,1} & \gamma_{1j,2} \end{bmatrix}. \tag{3.6}$$

Figure 3.2 is a visualisation of the above: The coordinates of the class points can be combined to form the class points lying on projection line, P. Furthermore, the the position of subjects ρ and τ in two dimensional Euclidean space are visualized, representing two subjects with a score of 1 and 2 on variable X respectively. In the biplot accompanying the original model of Worku and de Rooij (2018) the difference between the orthogonal projections of two subjects on the dimension,

relates to the difference in the log-odds of these two subjects. Because, in this example, the class points of variable j are positioned on both dimensions, we can project the position of subjects ρ and τ, perpendicular to projection line P. The projections are visualized by ρ' and τ' in Figure 3.2. β_1 and β_2 show respectively the difference on the first and second dimension of points ρ and τ. β_{12} shows the difference of the projections of ρ and τ, on the projection line.

A 45°- 45°- 90° triangle is visualized by red dotted lines, in Figure 3.2. This triangle consists of a line drawn from ρ parallel to β_{12} in combination with the prolongment of β_1 and the projection of τ on projection line P. It can be shown that the hypotenuse of this triangle equals $\beta_1 + \beta_2$, because a 45°- 45°- 90°triangle has a ratio of $1 : 1 : \sqrt{2}$. Therefore the congruent legs as well as β_{12} equal $\frac{1}{\sqrt{2}}(\beta_1 + \beta_2)$. From here we can see that the log-odds change by $\sqrt{2}\beta_{12}$ which equals $\beta_1 + \beta_2$. The multiplication follows directly from the Pythagorean theorem as mentioned before and is in line with Equation (3.2).

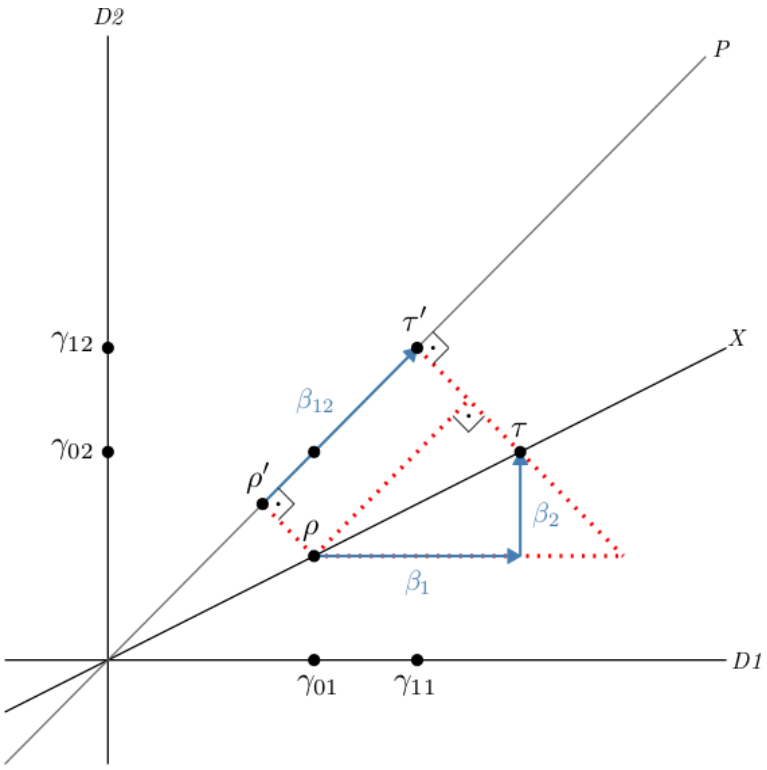


Figure 3.2 Visualization of a response variable on two dimensions and predictor variable X.

3.3.2 Decision regions of the biplot

As in the restricted MLD model, the decision boundary is the set of points for which the log-odds are zero. The decision boundary is perpendicular to the projection line that goes through the two class points, γ_{0j} and γ_{1j}. Similar to the restricted mode, the decision boundary is going through the point halfway between the class points.

Figure 3.3 is a visualisation of the response space of the fitted MLD model on the NESDA data in two-dimensional Euclidean space. Contrary to Figure 2.2, not all response variables belong

to only one dimension, GAD is positioned on both dimensions. The decision boundaries are displayed in Figure 3.3. The decision lines partition the space of Figure 3.3 into regions, each representing a most probable response profile. From the plot we can see that allowing response variables to belong to multiple dimension, has a crucial impact on which regions occur. For example, 12 response regions occurred while fitting the original MLD model on the NESDA data with a "distress-fear" structure (see Figure 2.2), while 14 response regions occur when fitting the extended MLD model on the NESDA data when GAD is positioned on both dimensions (see Figure 3.3). The maximum number of admissible response patterns with the dimensionality structure as presented in Figure 3.3 is 14 and the minimum number of response patterns is 12 (Coombs, 1964).

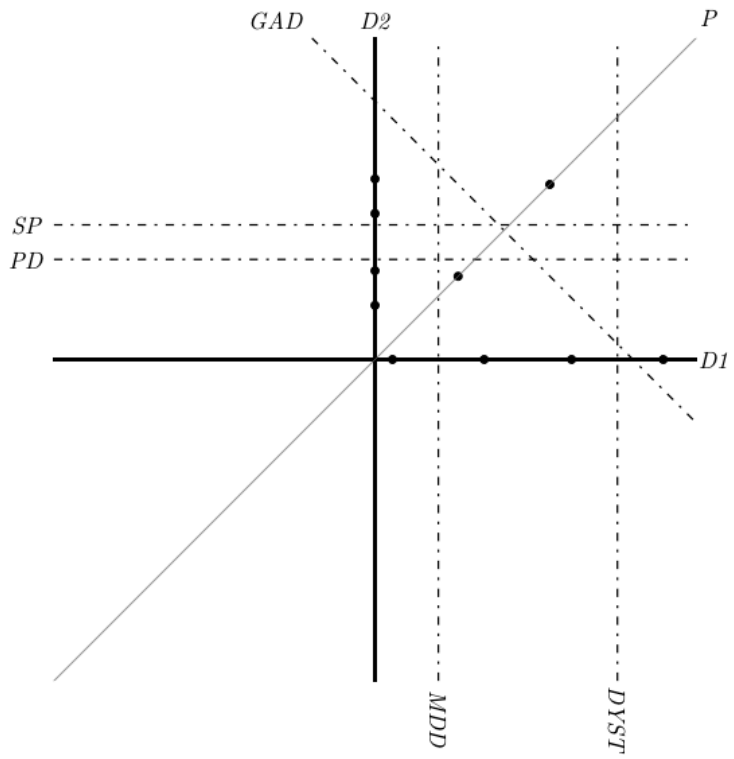


Figure 3.3 Response space of the extended MLD model fitted on the NESDA data.

Figure 3.4 visualizes a biplot of the extended MLD model fitted on the NESDA data similar to the model fitted in Figure 3.3. The positions of the subjects and the variable axes are obtained in a similar fashion as in the original model (see Section 2.6). The interpretation of the plot is similar to the interpretation of Figure 2.3, except for the interpretation of the variable axes (see Figure 3.2); in the original model the relative length of a variable axis projected on the dimension corresponds to the strength of the effect of the variable on the disorders belonging to the dimension. When we allow response variables to pertain to multiple dimensions, the relative length

of a variable axis projected on the projection line corresponds to the strength of the effect of the variable on the disorder, with a scaling factor of $\sqrt{2}$ (see Section 3.3.1). Figure 3.4 only shows the variable axes of the predictor variables *education, neuroticism, extraversion* and *conscientiousness*. From the plot we can see that the variable axes are almost similar to the variable axes in Figure 2.3.

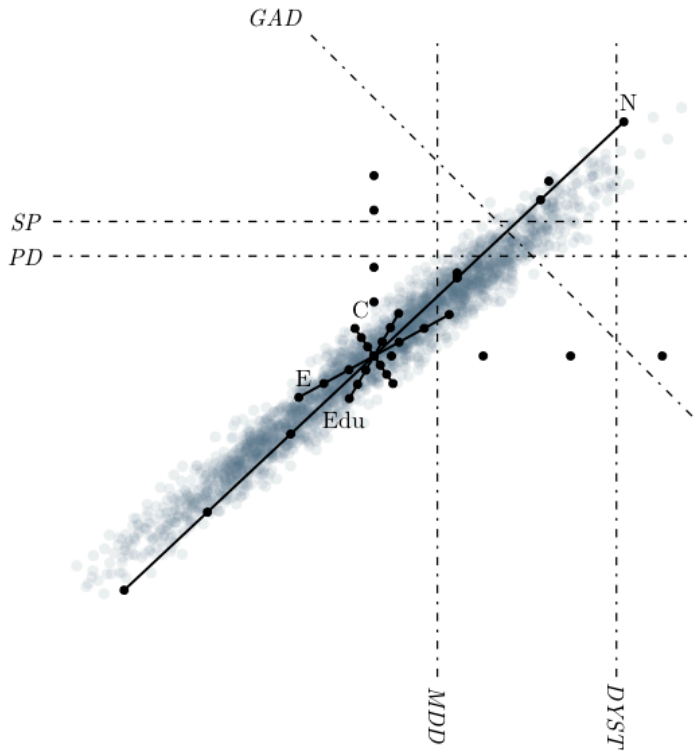


Figure 3.4 Biplot of the extended MLD model fitted on the NESDA data.

4

Non-linearity in the predictor variables

The multidimensional logistic distance model can be placed in a broader family of linear parametric models. These models assume that, on average, the change in the response variable is proportional to the change in the predictor variables. Linear models are sometimes an inevitable, approximation of the true function $f(\mathbf{X}) = E(Y|\mathbf{X})$. Inevitable because with a large number of predictors and a small number of observations, a linear model is all we can do without overfitting the data. Although linear models often tend to be easy to interpret, it is unlikely that the true function $f(\mathbf{X})$ is linear in its predictors. When we estimate a function that is not linear in its predictors by a standard (generalized) linear model, this can result in a model with very poor predictive power (Fox, 2015).

This chapter is concerned with the situation in which the assumption of linearity between the mean response and the predictor variables is not justified. Different approaches will be discussed on how to deal with non-linearity in the predictor variables. In essence all of the different approaches replace the predictor variables, $\mathbf{X}$, with transformations of these variables. The new variables are used to fit a (generalized) linear model and the model is therefore linear in its coefficients. We will assume until Section 4.3 that we only have one predictor variable X.

4.1 Global functions

Consider a model with one predictor variable X. The function $f(X)$ is represented by a linear combination of transformations of X:

$$f(X) = \sum_{b=1}^{B} \beta_b h_b(X), \tag{4.1}$$

with known transformations, also referred to as *basis functions*, $h_b(X)$, $b = 1, \ldots, B$ and parameters β_b that need to be estimated. When choosing a basis, we define the space of functions of which $f(X)$ is an element. Different choices can be made when choosing a basis: One way of

dealing with non-linearity is by fitting a polynomial regression model. Within polynomial regression the predictors are raised to a power e.g. $X, X^2, X^3, \ldots X^D$. For a polynomial regression model with degree D we get $h_d(X) = X^d$ and $d = 1, \ldots, D$, when denoted like equation (4.1).

Figure 4.1 shows the probability of having MDD against different levels of *conscientiousness* for the NESDA data set. The model fitted, is a logistic regression model of MDD using polynomial functions of the variable *conscientiousness* with $D = 4$ as predictors. The estimated point wise standard error is used to approximate a 95% confidence interval around the fitted model.

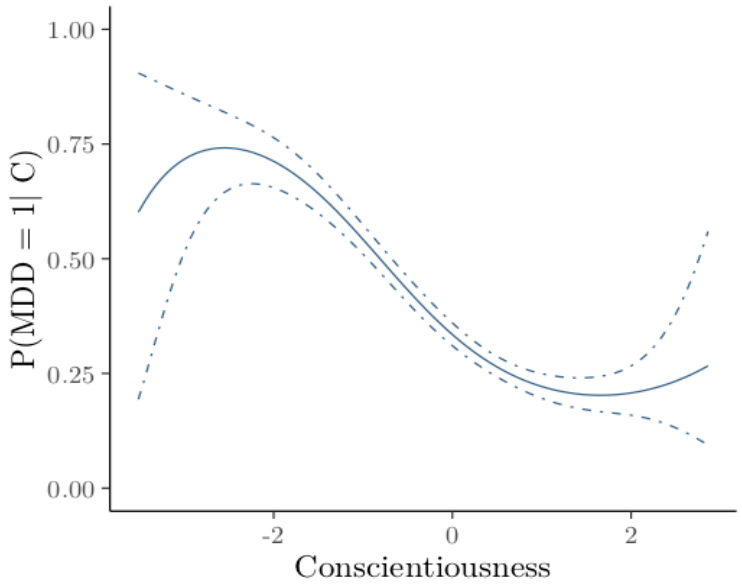


Figure 4.1 A 4th degree polynomial of the probability of having MDD as a function of conscientiousness with an estimated 95% CI.

When D is large enough, polynomial models offer a lot of flexibility without losing the interpretability of a linear model. Although, it must be noted that polynomials tend to be extremely flexible near the boundary of the domain of X, causing them to have unpredictable tail behavior (Friedman, Hastie, & Tibshirani, 2001). It can be seen from Figure 4.1 that the standard error near the boundaries of X tends to increase. Another disadvantage of using polynomial regression is, that polynomials are limited by their global nature. Changing the coefficients to model the form in one region can cause the function to change drastically in remote regions (Ramsay, 1988).

4.2 Piecewise Polynomials and Splines

As an alternative to fitting a single high degree polynomial over the whole domain of X, we can partition the data into distinct non-overlapping regions:

$$X < \xi_1,\ \xi_1 \leq X < \xi_2,\ \ldots,\ \xi_{K-1} \leq X < \xi_K,\ \xi_K \leq X, \tag{4.2}$$

fitting a different low degree polynomial in each region. The points between regions are known as *knots*, ξ_k, with $k = 1, \ldots, K$. The domain of X consists of $K + 1$ regions. Therefore, to indicate the different regions $K + 1$ basis functions are needed. When employing local basis functions instead of global basis functions, a given observation only affects the nearby fit, not the fit over the whole domain.

Although fitting a piecewise polynomial allows for a flexible local fit, it is generally desirable to restrict the resulting function to be continuous in value and sufficiently smooth at the knots. A *piecewise polynomial spline*, or spline for short, achieves these objectives by requiring the adjacent piecewise polynomials to join with a specified degree of smoothness at the knots. That is to say, a spline of degree D, with knots $\xi_k,\ k = 1, \ldots K$, consists of $K+1$ polynomial pieces of degree D and is required to be continuous and have continuous derivatives up to the $D-1$'th derivative at each of the knots. It is claimed that a cubic spline, i.e. a piecewise polynomial spline of degree 3, is the lowest order spline for which knot discontinuity is not noticable to the human eye, therefore it is the most commonly used spline in practice (de Boor, 1978; Friedman et al., 2001).

4.2.1 Spline Bases

The set of splines of order D over the knot sequence ξ_k, with $k = 1, \ldots K$ can be written as a linear combination of $D + K$ basis functions. Thus, the space of a spline is a vector space and therefore there are many basis functions to represent them, called equivalent bases. The design matrix obtained by a spline basis can be used to replace the column of the variable of interest in our predictor matrix. This modified predictor matrix can be used to fit the MLD model by applying the same method as described in Section 2.4.1.

The truncated power (TP) basis is a popular choice of basis functions, advocated by for example Ruppert, Wand, and Carroll (2003), because it is intuitive and conceptually simple. Generally, when using a TP basis of degree D and K knots, ξ_k, the function is given by:

$$f(X) = \sum_{j=1}^{D} \beta_j X^j + \sum_{k=1}^{K} \beta_{D+k} (X - \xi_k)_+^D \tag{4.3a}$$

with:

$$(X - \xi_k)_+^D = \begin{cases} 0 & X < \xi_k \\ (X - \xi_k)^D & X \geq \xi_k \end{cases}. \tag{4.3b}$$

It can be shown that the TP basis satisfies the constraints of a spline with continuous derivatives up to order $D-1$ at each of the knots. A verification of this is given in Appendix A. From equation (4.3a) and (4.3b) we can see that the TP basis representation has a nice natural interpretation of global polynomial of degree D, with local modifications to the right of each knot.

The design matrix obtained by the TP basis contains the values X to X^D in the first D columns, followed by K columns with the values of $(X - \xi_k)^D$. The obtained design matrix then, can be used to fit, for example, the MLD model.

Although the simplicity of the TP basis makes its use very attractive, their numerical properties are not favourable: Due to their construction of polynomials, the values in the design matrix of the TP basis can be very large or very small, which can lead to overflow errors and instabilities. Moreover, when the knots are very close together, the associated terms of the basis functions of the TP basis are almost similar for all observations, which makes them nearly co-linear (de Boor, 1978; Fahrmeir, Kneib, Lang, & Marx, 2013; Friedman et al., 2001).

Another way to represent splines, is throught the use of a B-spline basis (Curry & Schoenberg, 1947; de Boor, 1978). Despite the fact that they have a less intuitive interpretation, the B-spline basis functions are not linearly dependent and their values are always between 0 and 1, making them numerically more stable compared to the TP basis splines. Note that we use the term 'degree' to indicate splines that consist of piecewise polynomials of degree D. It should be mentioned that it is conventional, in the literature on splines, to use 'order', which equals $D + 1$. Therefore, henceforward we will not indicate a spline by its degree, but by its order, which will be denoted by Q. An order Q spline with K knots is characterized by $Q + K - 1$ parameters (Eilers & Marx, 2010). For B-splines of order Q, the basis functions consist of polynomial pieces of degree $Q - 1$, which are non-zero on a domain spanned by $Q + 1$ knots. To construct a B-spline representation of an order Q spline let us first define an augmented knot sequence:

$$\boldsymbol{\xi}^* = \xi_1, \xi_2, \ldots, \xi_{K+2Q-1},$$

resulting in $K + 2Q - 1$ knots for $K + Q - 1$ basis functions. By $B_{r,q}$ the B-spline basis r of order q is denoted, where $r = 1, \ldots, K + 2Q - 1$ and $q \leq Q$. The B-spline basis functions satisfy the recursive relation in terms of their divided differences and are given by:

$$B_{r,1} = \begin{cases} 1 & \text{when } X \in [\xi_r, \xi_{r+1}) \\ 0 & \text{otherwise,} \end{cases} \tag{4.4a}$$

for $r = 1, \ldots, K + 2Q - 1$ and

$$B_{r,q} = \frac{X - \xi^*_r}{\xi^*_{r+q-1} - \xi^*_r} B_{r,q-1}(X) + \frac{\xi^*_{r+q} - X}{\xi^*_{r+q} - \xi^*_{r+1}} B_{r+1,q-1}(X), \tag{4.4b}$$

for $r = 1, \ldots, K + 2Q - q$.

Figure 4.2 illustrates a sequence of B-spline basis functions up to order four with 7 equidistant knots. From the figure we can see that, at any given point in the domain of the B-spline, only Q basis functions are non-zero and they are constructed, such that

$$\sum_{r=1}^{K+Q-1} B_r(X) = 1 \quad \forall\, X. \tag{4.5}$$

The B-spline basis is locally defined, i.e. the basis functions are only positive on an interval based on $Q + 1$ knots. This differentiates them from the truncated polynomials of the TP basis which have positive values starting from a certain knot. Therefore, in contrast to the TP basis, we cannot make a distinction between global and local components. This makes the interpretation of the B-spline basis less intuitive (Fahrmeir et al., 2013).

Although the interpretation of the design matrix obtained by the B-spline basis is not as apparent as that of the one obtained by the TP basis, it can still be used to fit, for example, the MLD model, yielding the same fitted values as the model acquired by the TP basis.

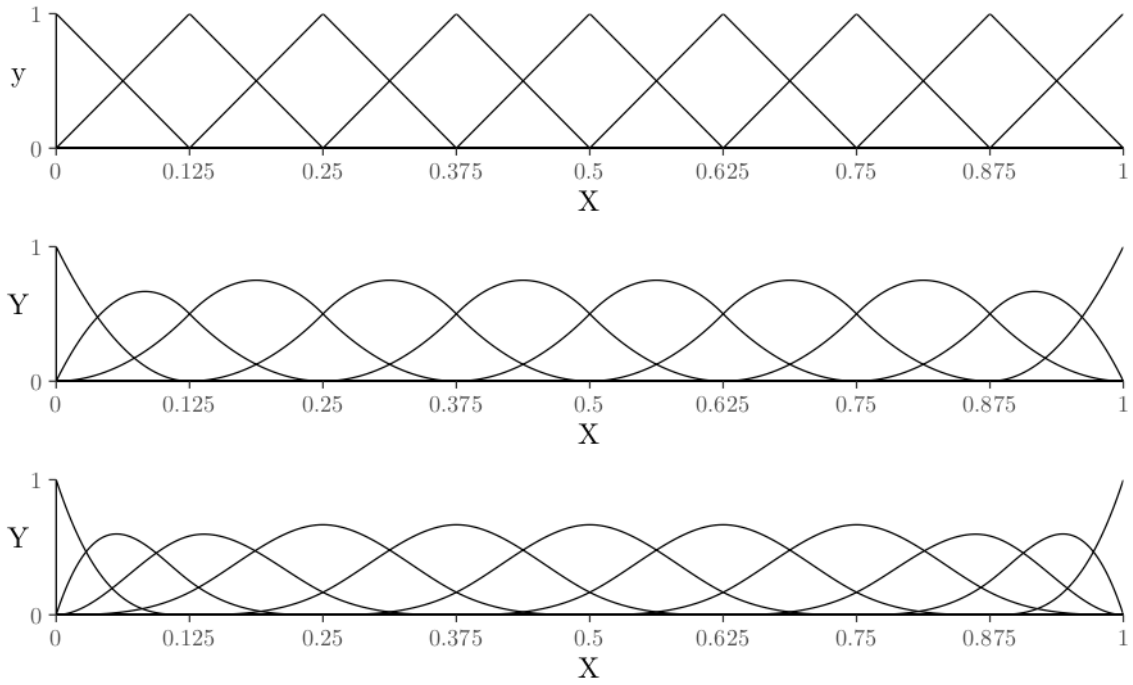


Figure 4.2 B-splines of order 2 to 4 from top to bottom, with 7 equidistant knots. The basis functions together span a spline space.

4.3 Multiple predictors

Until now we assumed X to be one-dimensional. Generalized additive models (T. Hastie & Tibshirani, 1986; T. J. Hastie & Tibshirani, 1990) can be used to identify and characterize non-linearity in the presence of multiple predictor variables. A generalized additive model (GAM) is a generalized linear model consisting of the sum of transformations of the predictors $\mathbf{X}$ (Wood, 2006). In general the model looks like:

$$g(\mu) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p), \tag{4.6}$$

where $g(.)$ is a smooth monotonic link function, $\mu = \mathbb{E}(Y|X_1, X_2 \cdots, X_p)$ with $Y \sim$ some exponential family distribution, and the functions $f_1(.)$, $f_2(.)$, ..., $f_p(.)$ are different non-linear

transformations of our predictor variables. The link function used in a multivariate logistic model is the logit link function. We model transformation $f_p(.)$ as:

$$f_p(X_p) = \sum_{b=1}^{B} \beta_{p,b} h_{p,b}(X_p), \tag{4.7}$$

where $h_{p,b}(.)$ is the b'th transformation of variable X_p. When the basis functions $h_{p,b}(.)$ are determined, the model is linear in these transformations, which allows for the same interpretation as in a generalized linear model. Employing GAM gives us the possibility to model a function $f(\mathbf{X})$ that is linear in some predictors, i.e. $h_{p,b}(X_p) = X_p$, and non-linear in others. This is particularly useful in modeling, for example with the MLD model, when we expect only some predictors to have a linear relationship with the response variable.

4.4 Visualization non-linear MLD model

To illustrate the impact of the proposed extension of the current MLD model by allowing for non-linearity, on the interpretation of the model, let us consider again the NESDA data set. From the personality traits, only *extraversion* and *neuroticism* had a statistically significant effect on both dimensions when fitting the original model with a two-dimensional 'distress-fear' structure. There is no indication of a linear association between, for example, *conscientiousness* and the log-odds ratio of the different response variables positioned at the two dimensions. However, it is theoretically possible for the trait *conscientiousness* to have a non-linear relationship with one or both dimensions and therefore with the disorders positioned on these dimensions. One could, for example, imagine that the effect of *conscientiousness* accelerates at some point along a dimension, increasing the probability of belonging to a certain category with respect to the other category. Moreover, it could even have a non-monotonic relationship with the dimensions, showing a negative effect on the log-odds when having a low score and a positive effect on the log-odds when having a higher score, or *vice versa*.

For illustration the MLD model was fitted on the NESDA data set, where the predictor variable *conscientiousness* was replaced in the design matrix by a cubic B-spline basis, with a single knot at zero, resulting in a spline basis with four basis functions. Figure 4.3 shows the relationship of the personality trait *conscientiousness* with the two dimensions separately. The figure indicates that, overall, a higher level of *conscientiousness* is associated with an increase in the probability of having a disorder that belongs to the 'distress dimension', that is, the log-odds of membership to category 1 versus the baseline category 0 increase, with higher scores on the personality trait *conscientiousness*. This effect seems to reverse at a score of approximately minus one and again around a score of one. The preceding indicates that, in between these scores, a higher score of *conscientiousness* is associated with a negative effect on the log-odds of the disorders positioned

on this dimension. The effect of *conscientiousness* on the 'fear dimension' seems to vary over the range of scores as can be seen from the plot. Overall a very high or a very low *conscientiousness* score is associated with a higher probability of having a disorder belonging to this dimension.

Figure 4.4 shows a plot of the variable axis as shown in Figure 4.3 for the two dimensions combined. The interpretation of the variable trajectory for both dimensions is similar to the interpretation of Figure 4.3. Along the variable axis an indication of the standard deviations of the trait is given. Because a score of three standard deviations falls outside the range of the observed data, it is not visualized.

Since the model is still linear in the basis functions, coordinates representing the position of subject i, $\boldsymbol{\eta_i}$, can still be obtained by taking a linear combination of the augmented matrix of the predictor variables. Figure 4.5 visualizes two variables, X_1 and X_2, having a non-linear and a linear relationship with the two dimensions respectively. Because the effect is additive in nature, we can obtain a subject's coordinates in two dimensional space by adding the two curves together, as in the original model. This can be done in a similar way as with completing parallelograms in vector addition: by shifting the origin of the vector to the point of the subjects score on the second variable trajectory, as illustrated in the figure.

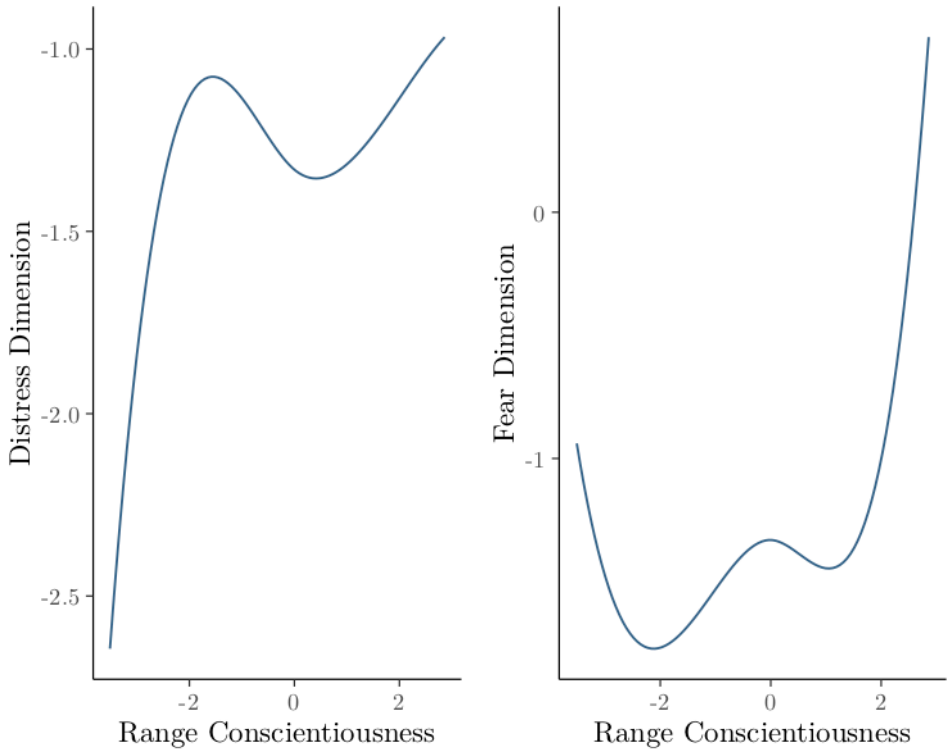


Figure 4.3 Cubic spline with one knot of the predictor variable conscientiousness fitted with the MLD model on the NESDA data set visualized for the distress dimension (left) and the fear dimension (right).

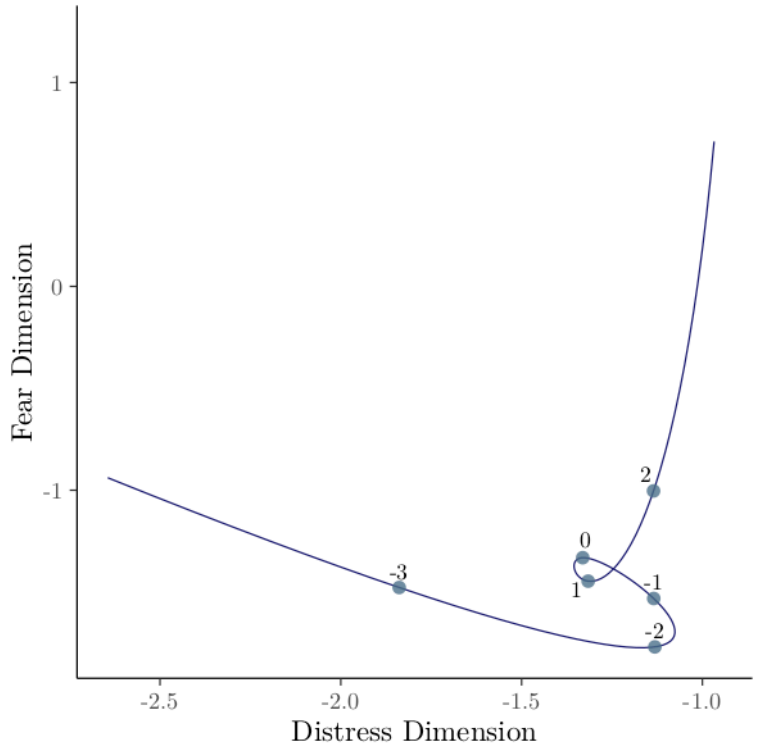


Figure 4.4 Cubic spline with one knot of the predictor variable conscientiousness fitted with the MLD model on the NESDA data set with an 'distress-fear' structure.

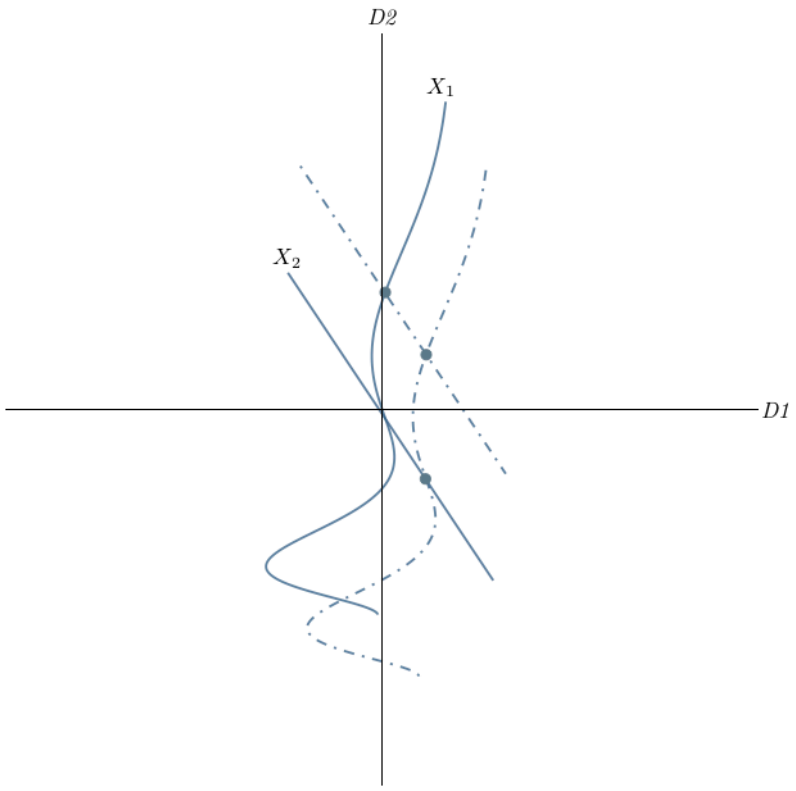


Figure 4.5 Two dimensional space with two variable trajectories represented by lines and scores on the variables represented by dots. Their combined coordinate in space is obtained by the intersection between the two projections of the variables on the other variable.

4.5 Equivalent Bases and Regularisation

Besides the choice of basis, spline modelling involves choosing the number and placement of the knots in the model. Choices regarding the number and placement of the knots can potentially have a substantial effect on the fit; too many knots can cause interpolation of the data, while too few knots can cause the model to be not flexible enough to capture the trend of the underlying data generating process.

To avoid overfitting of the model, we can use a penalization approach. Penalized (generalized) regression shrinks the regression coefficients of the model towards zero by putting a constraint on their size. Different penalties can be employed, although constraining the L_1 norm or the L_2 norm of the regression coefficients is most frequently used in practice (Friedman et al., 2001).

Tibshirani (1996) proposed LASSO (Least Absolute Shrinkage and Selection Operator) as a tool to perform variable selection, as well as regularization in order to enhance interpretability and to improve the prediction accuracy of a regression model. LASSO regression shrinks the regression coefficients, by constraining the L_1 norm of the coefficients. The LASSO estimator was originally formulated for OLS, but can be extended for the GLM situation. Given the outcome vector $Y = (y_1, \ldots, y_n)^T$, the $n \times p$ matrix of predictor variables $\mathbf{X}$, and tuning parameter $\lambda \geq 0$, the original LASSO estimate can be defined as

$$\hat{\boldsymbol{\beta}}_{LASSO} \in \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}||Y - \beta\mathbf{X}||_2^2 + \lambda||\boldsymbol{\beta}||_1. \tag{4.8}$$

Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, $||.||_2$ is the Euclidean norm and $||.||_1$ equals the L_1 norm. Note the element notation in 4.8, the LASSO only has a unique solution when rank($\mathbf{X}$) = p, because only then the criterion is strictly convex. Because of the nature of the L_1 norm, making λ sufficiently large, will cause some of the coefficients in the solution, to be shrunken exactly to zero. When $\lambda = 0$, no constraints are imposed and an OLS fit will be obtained. Moreover, when $\lambda = \infty$, the constraint penalizes all curvature, thereby setting all coefficients to zero except for the intercept.

Ridge regression is a shrinkage method like LASSO, introduced by Hoerl and Kennard (1970). The Ridge estimate is defined by

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}||Y - \beta\mathbf{X}||_2^2 + \lambda||\boldsymbol{\beta}||_2. \tag{4.9}$$

Because of the nature of the constraint, coefficients are also shrunken towards zero. Yet, contrary to LASSO, Ridge does not set coefficients in the solution exactly to zero. Instead of setting one of the regression terms of two co-linear predictor variables to zero, Ridge will share the regression weight between them. As λ goes to ∞ the Ridge estimator approaches zero but never equals zero. Again when $\lambda = 0$, no constraints are imposed and we will obtain a OLS model.

Although overfitting can be controlled by both methods, the use of an L_1 norm enables us to obtain a sparse solution. Both LASSO and Ridge are not scale invariant, so we assume that our input matrix $\mathbf{X}$ is standardized to have a mean of zero and a variance of one.

We will now illustrate regularisation with LASSO and Ridge in the context of splines by means of an example: A simulation study has been conducted to show the fit of three different models: an ordinary least squares fit, the LASSO estimator and the Ridge estimator. Data are simulated as follows: The X_i are taken uniformly over the interval [0, 1] and Y_i are simulated using $f(X_i) + \epsilon_i$, where $f(.)$ is a known non-linear function and $\epsilon \sim N(0, 2)$. The following function is used:

$$f(X) = 0.2X^{11} \times (10\,(1 - X))^6 + 10\,(10X)^3 \times (1 - X)^{10}. \tag{4.10}$$

A function to obtain the TP basis was implemented in `R`, furthermore the `bs` function from the package `splines` (R Development Core Team, 2008) was employed to obtain a B-spline basis. First we confirmed that the two different bases yield the same estimates, when an OLS model is fitted on the obtained design matrices. Hereafter the LASSO estimator and the Ridge estimator were obtained by implementing the model in the package `glmnet` (Friedman, Hastie, & Tibshirani, 2010), using the maximum number of knots, that is the number of unique values of the predictor variable minus one. The hyper parameter of both models, λ, was selected by performing ten-fold cross-validation, a technique which will be elaborated on further in the next chapter.

Figure 4.6 shows the models fitted, using a TP basis. As can be seen from the plot, the OLS fit nearly interpolates all data points. Both the LASSO estimator and the Ridge estimator yield a smooth function, while globally following the trend of the data. This is due to the fact that the TP basis is globally defined as long as the first three coefficients are not set to zero. From the plot we can see that the fit obtained by the LASSO estimator is closer to the true function than the one obtained by the Ridge estimator. In this example only 50 coefficients were retained with the LASSO estimator, these are based on 46 knots (50 coefficients min the intercept and the first three global terms). The initial model had 502 basis functions, based on 499 knots.

Figure 4.7 shows the model fitted, using a B-spline basis. Contrary to the unpenalized model, the two bases do not give an equivalent fit when the LASSO estimator and the Ridge estimator are employed. As can be seen from the plot, the OLS fit interpolates the data points. The LASSO estimator yields not a smooth function as with the TP basis. This is due to the fact that the B-spline basis is only defined locally; the curve equals the intercept on pieces of the range of X, where the coefficients are shrunken towards zero. In this example only 41 coefficients where retained with the LASSO estimator, where the initial model had again 502 basis functions. Although none of the coefficients of the Ridge estimator equal zero, we do not obtain a smooth function with the Ridge estimator either.

Figures 4.7 and 4.6 show that both penalized solutions approximate the true function $f(.)$ quite poorly. A better solution is obtained when a spline with less knots is fitted. Figures 4.8

and 4.9 show the same models fitted on cubic spline bases with only five equidistant knots, instead of the maximum number of knots. When a truncated power basis is used (see Figure 4.8), the OLS fit, as well as the fit obtained by the LASSO estimator, describe the underlying trend in the data much better than the models in Figure 4.6. However, the Ridge estimator still behaves poorly. When fitting the same models with a B-spline basis (as in Figure 4.9), all three models approximate the true function $f(.)$ well.

The question why to use a penalized solution at all remains, given that the OLS fit describes the pattern in the data as adequate as the penalized solutions when only five equidistant knots are used (see Figure 4.8 and 4.9). One of the reasons we are often not satisfies with the OLS fit is that although the fit often has low bias, it can be high in variance. Especially when our model is more complex, it is prone to overfitting (as we already saw in Figures 4.6 and 4.7). By introducing a little bias into the model, by means of a penalty term, we can reduce variance and can sometimes improve prediction accuracy. We will elaborate further on the prediction accuracy and the trade-off between the bias and the variance in the next chapter. Furthermore, when we have a large number of predictor variables, we often want to perform model selection which can be done with LASSO. As mentioned before the LASSO estimator can cause some of the coefficients to be biased exactly to zero. Fitting a non-linear model with a small number of equidistant knots (for example five) and a B-spline basis is favourable. Note that although the fit obtained with unpenalized spline regression is equivalent for both bases, the B-spline basis is favourable because of the attractive numerical properties of the basis.

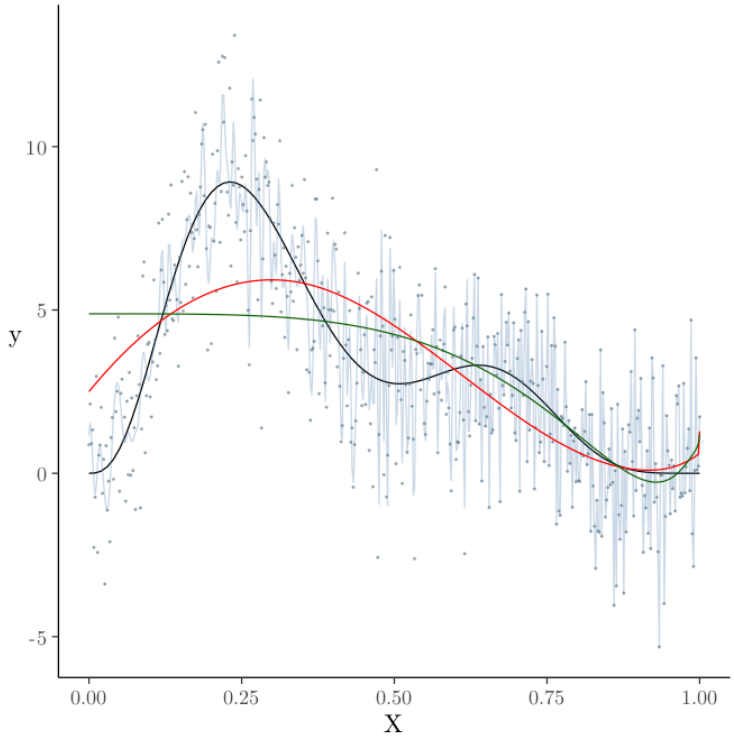


Figure 4.6 LASSO estimator (red), Ridge estimator (green) and an OLS fit (blue) of a cubic spline with a truncated power basis and the maximum number of knots.The black curve represents the true function.

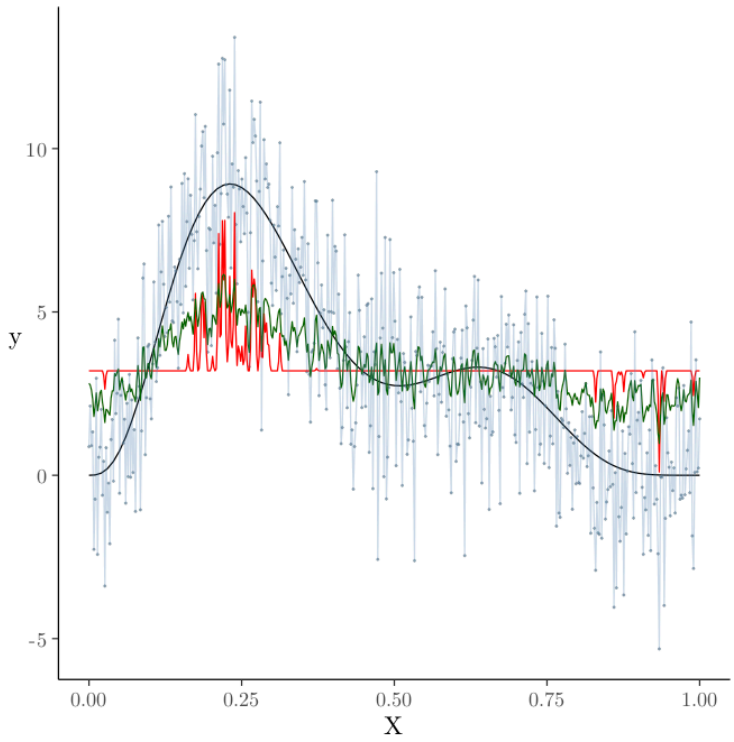


Figure 4.7 LASSO estimator (red), Ridge estimator (green) and an OLS fit (blue) of a cubic spline with a B-spline basis and the maximum number of knots.The black curve represents the true function.

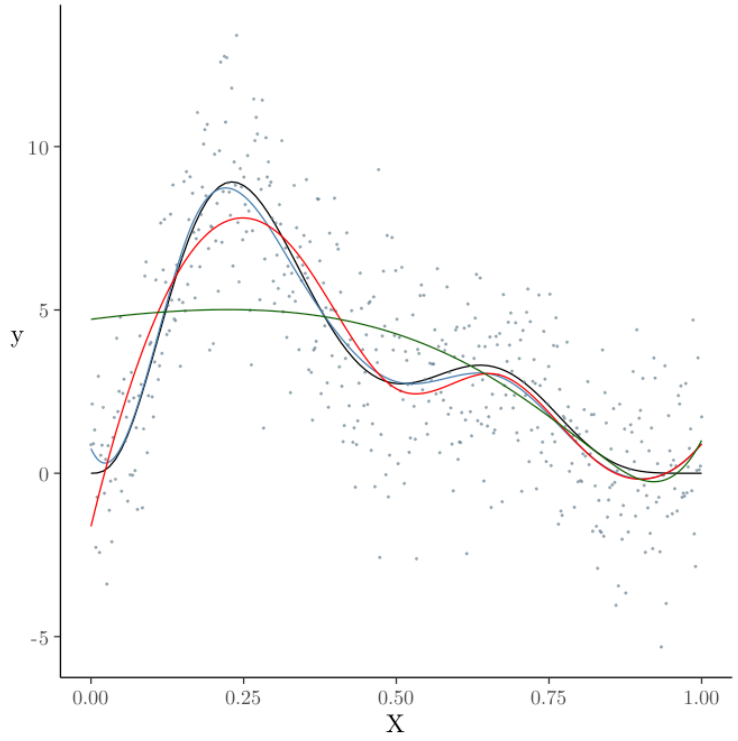


Figure 4.8 LASSO estimator (red), Ridge estimator (green) and an OLS fit (blue) of a cubic spline with a truncated power basis and five knots.The black curve represents the true function.

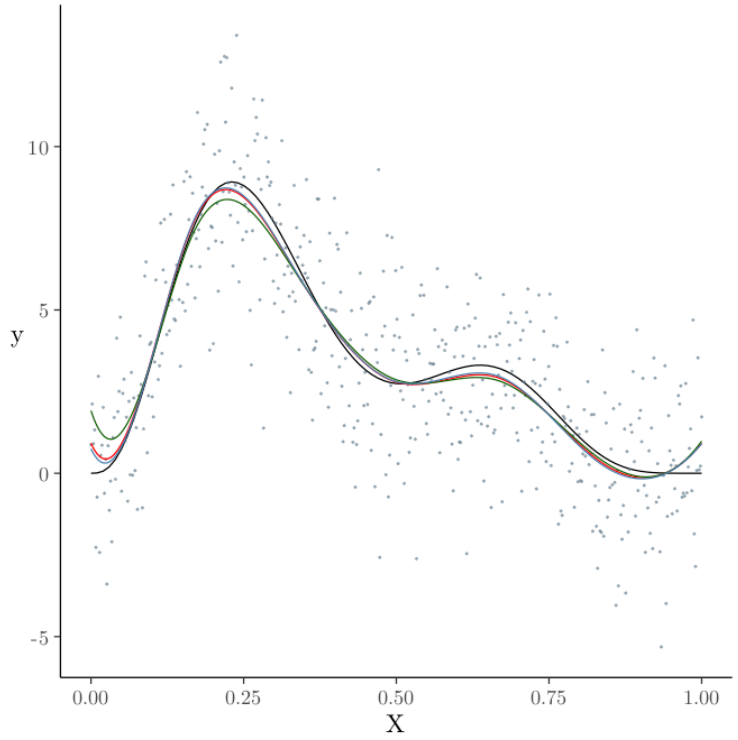


Figure 4.9 LASSO estimator (red), Ridge estimator (green) and an OLS fit (blue) of a cubic spline with a B-spline basis and five knots.The black curve represents the true function.

4.6 Biplot model with non-linear penalized terms

To illustrate the impact of the use of non-linear relationships in the MLD model on the interpretation of the corresponding biplot, let us again consider the NESDA data set. In the paper of Worku and de Rooij (2018) the predictor variables *openness to experience, agreeableness* and *conscientiousness* were excluded from the final model because of their performance in the linear model. We fitted a model with a two-dimensional 'distress-fear' structure for which the predictor matrix $\mathbf{X}$ was augmented. This is done by replacing the columns of predictor variables *openness to experience, agreeableness* and *conscientiousness* by a cubic B-spline basis of these variables, all with three equidistant knots. The augmented predictor matrix is used to fit a non-linear MLD model, in which we penalized the estimators obtained by the spline bases by means of the LASSO estimator. The model is fitted, using the `glmnet` package (Friedman et al., 2010). Because only the basis functions of the spline terms are penalized, we could potentially obtain solely linear terms. Figure 4.10 shows a biplot of the model with a two-dimensional 'depression-fear' structure. The positions of the subjects and the categories of the response variables are obtained in a similar fashion as in the original model (see Section 2.6). The trajectories of the non-linear terms are obtained in the following way: We first create a cubic B-spline basis with three knots of a vector with scores ranging from -3 to 3 with increments of .03. Thereafter, we multiply the obtained B-spline basis with the corresponding coefficients of the basis functions of the fitted model for both dimensions. We then connect the obtained coordinates to form a non-linear variable trajectory. The coordinates that are used to form the variable trajectory can be interpreted as the coordinates of different subjects with scores ranging from -3 to 3 on this variable and a score of zero on the other predictor variables. All predictor variables are included in the plot. Additionally labels are included at the positive side of the variable axes of the variable axes that are clearly visible, i.e. *education, neuroticism, gender, extraversion* and *conscientiousness*, are shown. From the plot we can see that we did not only obtain linear terms. However, the effect of *openness to experience* and *agreeableness* on both dimensions is very small and not clearly visible in the plot. This is due to the fact that the B-spline basis is locally defined and the LASSO estimator, with $\lambda = 0.0031$, shrunk some of the coefficients of the basis functions to zero; the variable trajectories for some of the variables only deviate from zero on a small part of their range.

Figure 4.11 shows a small part of the biplot visualized in Figure 4.10 in more detail. Only the variable trajectory of the predictor variable *conscientiousness* is vizualized in the figure. On the variable trajectory markers are placed at values ranging from -2 to 3, with increments of one. The figure indicates that, up until a score of minus one, a higher level of *conscientiousness* is associated with a positive effect on the log-odds of disorders positioned on the first dimension. The effect seems to reverse around a score of approximately minus one. However, a score of one and higher seems to have no effect on the log-odds of disorders belonging to the first dimension.

The effect of *conscientiousness* on the second dimension seems to vary over the range of scores as wel: The log-odds of membership to category 1 versus the baseline category 0 decreases with higher scores on the personality trait *conscientiousness*. This effect seems to reverse at a score of approximately minus two: With a score of minus two and higher, a higher score on the predictor variable *conscientiousness* is associated with a higher probability of having a disorder belonging to this dimension.

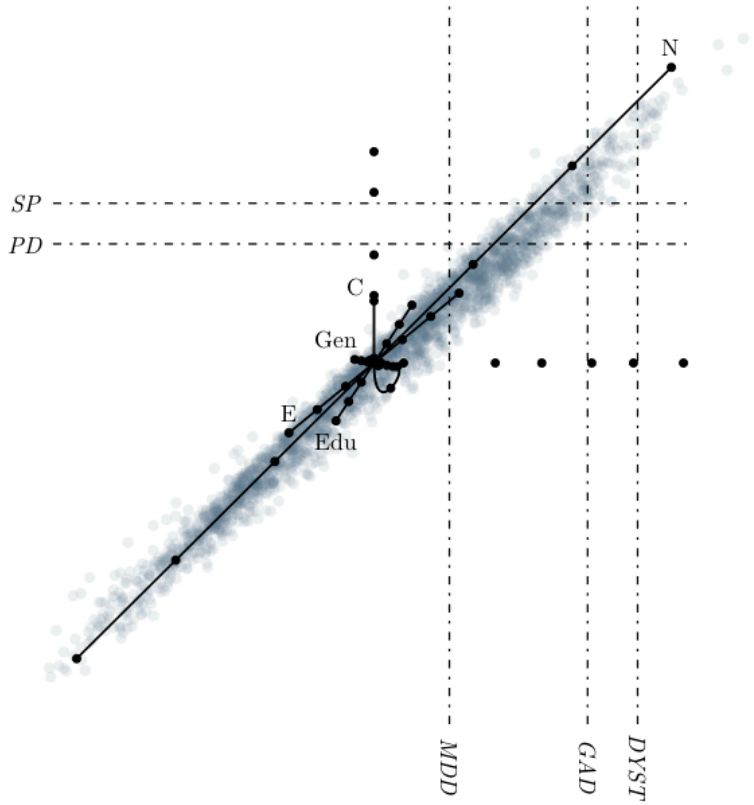


Figure 4.10 Biplot of a penalized non-linear MLD model fitted on the NESDA data with a two-dimensional 'distress-fear' structure.

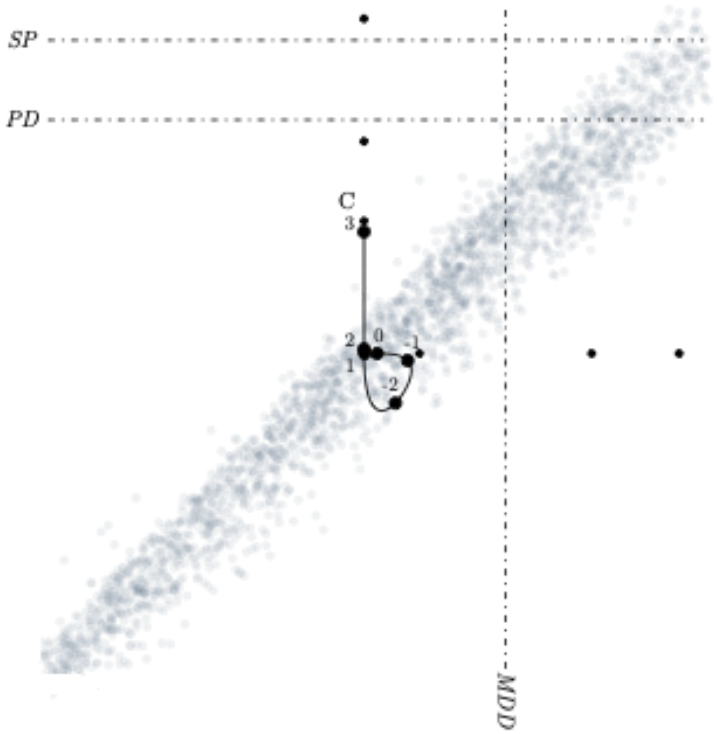


Figure 4.11 Enlargement of the variable trajectory of the predictor variable *conscientiousness* as shown in Figure 4.10.

5

Model Selection

5.1 Model selection original model

In the MLD model, the predictor variables for the final model as well as the dimensionality structure of the model need to be determined. We desire to select a model that is parsimonious, but still able to capture the underlying structure as well as the comorbidity patterns of the data.

By setting the distance between the two categories of a response variable in the MLD model to one, as in the restricted model, it becomes equivalent to a marginal model for multivariate binary data. The model can therefore be estimated by using marginal quasi likelihood methods. Because of the restriction on the class points, we presume that the predictor variables discriminate equally well for all response variables belonging to the same dimension. When this assumption is not justified, the mean structure of the model is not correctly specified. Moreover, when the mean structure of the model is not correctly specified, the obtained model will be biased. The bias of an estimator is a measure on how good our estimator is estimating the real value of our parameter. The bias of an estimator is defined as:

$$bias(\hat{\boldsymbol{\theta}}) = \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}, \tag{5.1}$$

where $\mathbb{E}(\hat{\boldsymbol{\theta}})$ is the expectation of the parameter estimate and $\boldsymbol{\theta}$ is the true underlying parameter. The estimator is biased when the expected value of the estimated parameter differs from the true underlying parameter. Biased estimators tend to systematically over- or underestimate the true parameters.

Variable selection in the original MLD model is based on a Wald test performed on the regression parameters with standards errors, obtained with the sandwich estimation method or the clustered bootstrap method. If the MLD model is biased, the estimated variances of the regression coefficients, based on the sandwich estimation method, will be biased as well because the GEE method assumes the mean structure of the model to be correctly specified. The violation of this assumption invalidates null-hypothesis significance tests, like the Wald test (Fox, 2015). Therefore, when the predictor variables do not discriminate equally well, i.e. our model is biased, we need an alternative to select predictors in our model.

Lately, there has been an increased interest in predictive power, in contrast with explanatory data analysis. When accessing the prediction capability of a model, the interest lies in the predictions generated by the model, not in the predictor variables itself (Shmueli, 2010). Instead of trying to determine if a coefficient equals zero, as in null-hypothesis significance testing, we would like to focus on a predictors relevance to the response variable when selecting a predictor in the model. Therefore we propose to use the prediction capability of a model on independent test data to validate the model. We can use the ability to predict classes for independent test data to select between different competing models.

In the original MLD model, the QIC_u norm (Pan, 2001) was used to determine the dimensionality structure of the model. When comparing different models on their prediction capability, we provide a unified framework in which we can both examine the predictive power of the variables in the model, as well as the dimensionality structure of the model.

5.2 Bias-variance trade off

Within the context of *machine learning*, the performance of a model relates to its prediction capability on independent data (Friedman et al., 2001). When employing maximum likelihood, the optimal model is determined by choosing parameters that make the observed data 'most likely'. However, such a model does not necessarily provide the best prediction. Because we evaluate the model using the data we used for training, we are prone to overfit the model: The model stores information that is specific to the training data used to obtain the model and is not part of the general trend of the data generating process, rendering the model non-generic. A model that is overfitting the data is said to be high in variance, i.e. the estimators obtained by the model vary heavily when different data sets would have been used to fit the model. The opposite is called underfitting, and refers to an overly simple model that can neither model the training data, nor capture the important trends of the data generating process. Although a model underfitting the data is more stable, and therefore low in variance, it yields a model that is high in bias. In general we desire to select a model that is not too strongly tailored to the particularities of the training set, but still is able to capture the general trends of the data generating process, i.e. a model with a good predictive performance on independent new data.

In order to clarify the relationship between the two mechanisms that are a source of error, bias and variance, let us show the decomposition of the expected prediction error, that is, the expected error on new input, also known as test or generalization error. The expected prediction error using squared-error loss, can be decomposed as:

$$EPE = \underbrace{\text{Var}(y)}_{\text{irreducible error}} + \underbrace{\text{Var}(\hat{f}(\mathbf{x})) + \text{Bias}^2(\hat{f}(\mathbf{x}))}_{\text{reducible error}}. \tag{5.2}$$

The Expected Prediction Error of using the fitted model, $\hat{f}(.)$, to predict y is the sum of the variances of $\hat{f}(\mathbf{x})$ and y plus the squared bias of $\hat{f}(\mathbf{x})$. The the variance and the squared bias of our estimator $\hat{f}(\mathbf{x})$ form the mean squared error (MSE) between the true function and the predictions. Unfortunately we cannot influence the irreducible noise, the variance of y, but the MSE is a function of our estimator it can possibly be reduced (Goodfellow, Bengio, & Courville, 2016; Matloff, 2017).

As mentioned before, by wrongfully making the assumption that a predictor variable discriminates equally well for all response variables pertaining to one dimension, a biased model is obtained. When the assumption is not justified, we know in advance that our model is not capable of capturing the true underlying structure in the data, that is, the model is not complex enough to capture this structure and the estimators obtained by the MLD model will therefore structurally underfit our data. Yet, the obtained estimators are low in variability, i.e. they are more stable, compared to estimators obtained by a more complex model like the unrestricted model. Thus, although there is an increase in bias, compared to an unrestricted model, there will be a decrease in variance.

5.3 Cross-validation

Ideally we would evaluate the predictive performance of a model by randomly dividing the data into two parts: A *training set* on which we fit the model and a *validation set* which is used to estimate the prediction error of the model. A disadvantage of this method is that it is not very efficient, i.e. only part of the data is used to fit the model. We therefore desire to employ a method to access the predictive performance of the model, while utilizing the data more efficiently. Cross-validation (Allen, 1974; Stone, 1974) is one of the most elegant and commonly used methods to evaluate predictive performance of a model in fields as machine learning and pattern recognition. When performing cross-validation the data are randomly partitioned into V equally sized folds. Iteratively each fold v is retained as the validation set, while the other $V-1$ folds combined form the training set. A model is trained on the training set and its predictive error is evaluated on the validation set. We do this for $v = 1, \ldots, V$ and average the V estimates of prediction error to obtain a final estimate of the prediction error. Optionally a standard error of this estimate can be calculated as well. Note that this approach is often referred to as K-fold cross-validation. However, using the letter K in this regard would lead to confusion, as it is already used to indicate the number of knots. In the field of *unsupervised learning* the method is often called V-fold cross-validation, therefore we choose to use this notation instead.

How to select an adequate value for V has been substantively studied in the literature. This choice is again based on a bias-variance trade-off: When V is large the obtained estimator is likely to have a low bias with regard to the true prediction error. Yet, the expected prediction errors of

the trained models are highly (positively) correlated with each other because the training sets are so similar. The expected prediction error of highly correlated folds has higher variance compared to the expected prediction error of folds that are not as highly correlated (James, Witten, Hastie, & Tibshirani, 2013). Contrary, when V is small, the expected prediction error obtained by cross-validation has lower variance, but the bias could potentially be higher. Hence, we want the size V to be a good compromise between the bias and the variance of our estimate. Typically five- or tenfold cross-validation is recommended to balance between the bias and the variance (Breiman & Spector, 1992; Friedman et al., 2001; Kohavi, 1995).

When employing cross-validation to examine the prediction error of the MLD model, we need to retain the correlation structure between the multivariate responses (see Equation (2.14)). For this reason clustered cross-validation can be utilized. This is a cross-validation procedure in which the data is partitioned in folds on a subject level. In this fashion, the generated folds retain the same dependence structure as the original data (Roberts et al., 2017).

5.3.1 Nested Cross-validation

Apart from model selection, cross-validation can be used to tune the parameters of a model (for example λ when performing Ridge regression or LASSO). We desire to select the hyper parameter with a value that minimizes the loss function. Yet, when cross-validation is simultaneously used to evaluate the model and to select the hyper parameters, we need to be careful: When the same test set is used to both select the values of the parameter and evaluate the model, we are prone to underestimate the prediction error and therefore overfit our model, as pointed out in the paper of Cawley and Talbot (2010).

To overcome this problem nested cross-validation is required. As in normal cross-validation each fold v is retained once as the validation set, while the other $V-1$ folds combined form the training set. This training set is used to employ inner cross-validation, that is, we randomly partition our training set again in folds, used to perform cross-validation to tune the parameters. Hereafter, the model selected by inner cross-validation is evaluated on our validation set v.

5.4 Loss function

When evaluating the predictive performance of a model, we compute the prediction error by means of a loss function. Since the decomposition of the expected prediction error of the squared loss is quite trivial (as shown in Section 5.2), the mean square error (MSE) is a popular choice of loss function. When $\hat{f}(\mathbf{x})$ is a probability of belonging to a certain class, as in the MLD model, the MSE equals the mean squared error of the prediction and is generally referred to as the *Brier Score* (Brier, 1950) and is defined in the multivariate setting as

$$Brier = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J}(y_{ij} - \pi_j(\mathbf{x}_i))^2. \tag{5.3}$$

Note that the original definition of the Brier Score accounted for the number of classes within a multinomial setting, i.e. the mean squared loss multiplied by the number of classes. Despite this, the formulation without the multiplication factor is most commonly used (Jolliffe & Stephenson, 2012) and will henceforth be the formulation in this thesis.

Other choices of loss functions can be made to evaluate our model. For example, one might be interested in utilizing the miss-classification rate or the cross entropy error. However, the use of alternative loss functions can result in different conclusions about the 'optimal' model. We choose to use the Brier Score because of its mathematical properties (e.g. we can easily decompose the Brier Score into the squared bias and the variance) and because of the popularity to use the (mean) squared loss function within a regression setting.

5.5 Model validation

To illustrate and compare the different models proposed in this thesis, 16 different models were fitted on the NESDA data set. All data were centered and scaled before fitting the models. The dimensionality structure of the proposed models is

- an 'uni-dimensional' structure with all five disorders represented on a single dimension;
- a two-dimensional 'distress-fear' (d/f) structure with one dimension representing distress [MDD, GAD & DYST] and the other dimension representing fear [PD, & SP];
- a two-dimensional 'depression-anxiety' (d/a) structure with one dimension representing depression [MDD & DYST] and the other dimension representing anxiety [PD, SP & GAD];
- a two-dimensional structure where the response variable GAD is represented on both dimensions (GAD 2d): One dimension represents [GAD, MDD & DYST] and the other dimension represents anxiety [PD, SP & GAD];

The different dimensionality structures were used to fit eight linear models: Four unpenalized linear MLD models were fitted utilizing the `glm()` function in `R`. Furthermore, four penalized linear MLD models were fitted, in which the LASSO estimator was used. This was done by means of the `glmnet` package (Friedman et al., 2010).

Additionally, eight the same but non-linear models were fitted employing the different dimensionality structures. In the original paper of Worku and de Rooij (2018), the predictor variables *openness to experience, agreeableness* and *conscientiousness* were excluded from their final model

because of their performance in the linear model. We augmented the predictor matrix $\mathbf{X}$ by replacing the columns of these variables by a cubic B-spline basis of these variables, all with three equidistant knots. Thereafter we used the augmented predictor matrix to fit four unpenalized MLD models function, and four penalized MLD models employing the L_1 norm. The models were fitted respectively by means of the `glm()` function and the `glmnet` package in `R`. Note that, in the panalized solution, we only penalized the estimators obtained by the spline bases. Because LASSO is not scale invariant, our data needs to have the same scale, but the values of our obtained spline basis are bounded between zero and one (see Section 4.2.1) and their scale therefore differs from the other variables (Denison, Mallick, & Smith, 1998; Osborne, Presnell, & Turlach, 1998).

Ten-fold clustered cross-validation was performed to examine the prediction error of all models. This was done in order to preserve the correlation structure between the multivariate responses. In addition, we estimated penalty parameter λ for the penalized models by means of nested cross-validation (see Section 5.3.1). Again ten folds were used for the inner cross-validation to select the penalty parameter.

One could consider to select the number of knots and order of the spline bases through the use of nested cross validation as well. Because it is claimed that a cubic spline, a spline of order four, is the lowest order spline for which knot discontinuity is not noticable to the human eye, it is the most commonly used spline in practice. Therefore, we choose not to select the degree of our spline bases through the use of nested cross-validation, but to use a cubic spline instead. Moreover, we choose not to employ nested cross-validation to select the number of knots in our spline bases because of the computational burden it entails.

We evaluated the predictive performance of the models by means of cross validation. The cross validation errors of all models are shown in Table 5.1 accompanied by their corresponding standard error, obtained over the ten folds. The Brier Score was employed as a loss-function to evaluate the different models and equals the cross-validation error. The cross-validation error of the different linear and non-linear unpenalized models tend to be negligible.

By introducing non-linear terms in the unpenalized model, the model is more flexible. Therefore a model with less bias could be obtained. Yet, there tends to be no difference in the expected prediction error of the linear models compared to the non-linear model. Suggesting that the decrease in bias is compensated by a increase in variance, leading to a model with no better prediction performance. Likewise, there tends to be no difference in the cross-validation error of the different linear and non-linear penalized models. Yet, there appears to be a substantial difference in the cross-validation error of the unpenalized models compared to the penalized models; the predictive performance of the penalized models is favourable over the unpenalized models. Thus, by introducing extra bias into the model through the use of a penalty term, the variance drops. This results in a sparse solution with better prediction capability compared to the unpenalized models.

When models do not substantially differ in their predictive performance, it is custom to select the most parsimonious, and therefore least complex, model within one standard deviation of the model with the lowest prediction error. For this reason we do not choose to select a non-linear model, as it does not improve the predictive performance, but makes the understanding of the model more complex. The most parsimonious model is the unidimensional penalized model. The LASSO estimator, with $\lambda = 0.0034$, shrunk the coefficient of some of the predictor variables to zero. The selected variables in the final model are *education, neuroticism, extraversion* and *agreeableness*.

Table 5.1 Cross-validation error of sixteen different MLD models.

	Unpenalized		**Penalized**	
	CV error	*S.E.*	*CV* error	*S.E.*
Linear models				
1 dimensional	.1925	.0044	.1379	.0031
2 dimensional (d/f)	.1929	.0044	.1378	.0031
2 dimensional (d/a)	.1933	.0044	.1377	.0031
2 dimensional (GAD 2d)	.1903	.0043	.1391	.0030
Non-linear models				
1 dimensional	.1931	.0045	.1382	.0031
2 dimensional (d/f)	.1934	.0044	.1377	.0030
2 dimensional (d/a)	.1938	.0045	.1377	.0030
2 dimensional (GAD 2d)	.1910	.0043	.1390	.0029

CV error = cross-validation error based on the Brier Score,
S.E. = standard error.

6

Discussion

Worku and de Rooij (2018) proposed a marginal model for the analysis multivariate binary data in a distance framework, the multivariate logistic distance (MLD) model. Two different models were introduced by Worku and de Rooij, both a *restricted* and an *unrestricted* MLD model. The former model imposes a restriction on the class points of the response variables. In this work we have extended the work of Worku and de Rooij (2018) by proposing three extensions to the *restricted* MLD model. The extended model may be estimated using software for univariate models (such as GLM and GEE). The current chapter provides a concise summary of the extensions of the MLD model as presented in Chapters three to five. Furthermore, suggestions for future research will be given.

6.1 Modification dimensional structure

In the original MLD model, different theories about the dimensional structure could be studied to access comorbity patterns in the data. However, the model only allowed for the assessment of dimensional structures in which each response variable relates to a single dimension. As a result of this restriction, other dimensional structures that could, for example, be accessed with *Structural Equation Modelling* (SEM) or *Confirmatory Factor Analysis* (CFA), can not be studied.

In this thesis, we examined the possibility for a response variable to belong to multiple dimensions. We showed that, by imposing an extra constraint on the class points, the model with response variables on multiple dimensions could be defined. The constraint dictates that, if a response variable belongs to multiple dimensions, the coordinates for the different class points are equal for all dimensions. That is, $\gamma_{0j,m} = \gamma_{0j,1}$ and $\gamma_{1j,1} = \gamma_{1j,1} \ \forall \ m$, for response variables pertaining to multiple dimensions.

The amendment of the model entails changes in the interpretation of the corresponding biplot, which is illustrated in two dimensional space in Section 3.3.1. The class points of a variable belonging to both dimensions are positioned on a projection line. This projection line goes through the origin and has a 45 °angle with both dimensions. The scores of subjects can be

projected perpendicular to the projection line. The scaled distance between the projections is associated with the difference in the log-odds of membership to the different classes of disorders positioned on this projection line.

An advantage of the original MLD model of Worku and de Rooij over existing marginal models for multivariate data, is the possibility for dimension reduction as a form of regularization. The model is less complex compared to a standard marginal model for multivariate data, because less parameters are estimated. We have shown the application of the proposed extension employing an empirical data set, the NESDA data set. The biplot in Figure 3.4 is partitioned in different regions indicating the most probable response profiles. From the figure we can see that the proposed extension has a crucial impact on which regions occur: Allowing a response variable to belong to multiple dimensions leads to more response profiles. Therefore, the extended model is less restricted compared to the original model, while the same number of parameters are estimated.

The extended model is related to *multivariate marginal models* as proposed by Asar and İlk (2013). When employing the method of Asar and İlk, parameter estimations are obtained by the GEE approach. Compared to our extended model, each response belongs to a unique dimension resulting in a J-dimensional model. However, equality restrictions can be incorporated, resulting in shared parameters between some of the response variables for certain predictor variables. Yet, in our model all parameters are shared for response variables pertaining to the same dimension. For response variables that belong to multiple dimensions the parameters are obtained by taking the sum of the parameters per predictor variable over the different dimensions.

In empirical sciences latent variable models are often used to study comorbidity patterns in the data (see for example Spinhoven et al., 2013; Beesdo-baum et al., 2009). Yet, these models often make unverifiable distributional assumptions about the response variables and the latent variables (see Worku, 2018, Chapter 2). We have shown that the extended MLD model can be used for comparing theories about the comorbity patterns, without making these assumptions. Moreover, the extended model is not restricted in the dimensional structures, as in the original model.

6.2 Incorporating non-linear relationships

In the original model of Worku and de Rooij (2018), a strong linearity assumption was made, i.e. we could solely examine the linear effect of a predictor variable on the logit transformation of the probabilities of different classes of response variables. However, it is unlikely that this effect is always linear. We showed that we could extend the model by allowing for non-linear relationships. Different approaches to incorporate non-linear relationships into the model were presented. All approaches include augmentation of the predictor matrix, before fitting the model.

In addition, we showed the implications for the interpretation of the biplot when incorporating a non-linear variable axis into the model through the use of a spline basis (see Figures 4.4 and 4.5).

The biplot accompanying the non-linear MLD model is related to *spline-based nonlinear biplots*, as proposed by Groenen, Le Roux, and Gardner-Lubbe (2015). This is a visualisation technique to show the relationship between subjects and variables in a single plot using B-splines. The main difference between the biplot of Groenen et al. and the biplot accompanying our model, is the interpretation of the plot: In the plot of Groenen et al., subjects and curves representing the variables are visualized in low dimensional space in such a way that the point on the curve nearest to the subject relates to the predicted value for that subject on the corresponding variable. In the biplot accompanying our model, the position of the subjects is computed as a linear combination of their scores on the (transformed) predictor variables. Furthermore, we make a distinction between response variables, represented as points, and predictor variables, represented as curves.

When employing spline bases, the number and placement of the knots could potentially have a substantial effect on the fit. Therefore, we conducted a simulation study in which penalized regression was used to prevent overfitting, while utilizing the maximum number of knots (see Section 4.5). We compared two different spline bases, the Truncated Power basis and the B-spline basis, and tested if they are equivalent when fitting penalized regression both with an L_1 and an L_2 constraint. It was shown that the bases do not result in an equivalent fit when used for penalized regression. Only the TP basis resulted in a smooth fit when using the maximum number of knots (see Figure 4.6). This is due to the fact that the function is globally defined. In contrast, the B-spline basis is locally defined. Therefore on pieces of the range of the predictor variable, X, where all coefficients are shrunken towards zero, the curve is equal to the intercept (see Figure 4.7).

Employing a spline basis with substantially less knots yields a model with a considerably better fit compared to a model with the maximum number of knots (see Figure 4.6 and 4.8). Compared to the TP basis, the B-spline basis shows a better fit, presumable because of the numerical properties of the basis. We conclude that, fitting a non-linear model with a small number of equidistant knots and a B-spline basis is favourable for the MLD model. It should be noted that more research is needed to study the performance of penalized spline regression in the MLD model. Moreover, other penalization approaches could be explored like *Smoothing splines*, in which the wiggliness of the curve is controlled by penalizing the integrated squared second order derivative (see for example Friedman et al, 2001); *P-splines*, in which the finite order differences of the coefficients is penalized (Eilers & Marx, 1996); or *Adaptive Ridge*, a technique comparable with P-splines with the noticeable difference that automatic knot selection can be obtained by this method (Goepp, Bouaziz, & Nuel, 2018). We suggest future research to study the use of these different penalization approaches. The exploration of these techniques is beyond the scope of

this thesis, we solely accessed penalization approaches most frequently used in practice (Friedman et al., 2001).

Besides exploring the effect of other penalization approaches we suggest experimenting with other empirical data sets to see if the number and the placement of the knots still has minor influence on the fit, especially when observations are not evenly distributed over the range of X.

Furthermore, we were only able to penalize estimators of the same order, that is, when working in a multivariate setting, we can only penalize the non-linear terms, provided they are of the same order, or the linear terms. This is due to the fact that both LASSO and Ridge are sensitive to scaling. We suggest for future research the use of group penalization in which a penalty term per group coefficients of the same order is used (Osborne et al., 1998). Alternatively one could explore the possibility of centering and scaling the variables without a spline basis such that they are bounded between zero and one. In this way, when employing a B-spline basis, all variables are of the same order.

6.3 Change of model selection criterion

In statistical analysis, we often want to select a model from a set of candidate models, with the optimum balance between complexity and model fit (D. Anderson & Burnham, 2004). In the MLD model the dimensionality structure as well as the final predictor variables need to be selected.

Traditionally in scientific fields like psychology, economics and epidemiology, statistics is focused on explanatory modelling. Often researchers mistakenly assume that models with high explanatory power are also high in predictive power. Lately, there has been an increased interest in predictive modelling. In predictive modelling the prime interest is in the predictions the model generates, not in causal explanation (Breiman, 2001; Hand, 1999; Shmueli, 2010). In the *restricted* MLD model of Worku and de Rooij (2018), it is assumed that response variables belonging to the same dimension have the same underlying relationship with the predictor variables. When this assumption is not justified, an obtained parameter is not a good reflection of the true relationship between a predictor variable and the individual response variables, but a measure of the average effect of the predictor variable over all response variables pertaining to a dimension. The obtained estimates of the model are therefore likely to be biased. Hence, performing null-hypothesis significance tests with these obtained estimates, as suggested by Worku and de Rooij, is not an adequate method for selecting variables in the model. Therefore, we propose to select the predictor variables based on prediction capability. One of the most elegant and commonly used methods to evaluate the predictive performance of a model is cross-validation. The proposed selection criterion is in line with the recent interest in predictive modelling and rests on the bias variance trade-off mechanism (as explained in Section 5.2). Contrary to the model selection criterion proposed by Worku and de Rooij, prediction error can be used to simultaneously

select the dimensionality structure and the final predictor variables of the model in one unified framework.

In order to show the application of employing prediction capability as a selection criterion for the MLD model, we compared 16 different models: Four different dimensionality structures, with and without non-linear terms and with and without L_1 penalization were fitted. Ten-fold clustered cross-validation was performed to examine the expected prediction error of all fitted models. The Brier Score was used as a loss-function to evaluate the different models. The cross-validation error of all models, which equals the Brier Score, can be found in Table 5.1. It should be noted that we can partition the data into ten folds in many ways, leading every time to a slightly different estimate of the expected prediction error of our models. Repeating the cross-validation procedure a number of times, as advocated by for example Harrell (2015), is therefore strictly more precise than non-repeated cross-validation. However, the computational burden is also considerable, requiring more applications of the model. For this reason we only repeated the procedure once per model.

It was shown that the cross-validation error of the different linear and non-linear unpenalized models vary little. Likewise, there tends to be no substantial difference in the cross-validation error of the different linear and non-linear penalized models. There seems to be a substantial difference in the prediction capability of the penalized models compared to the unpenalized models: The penalized models tend to outperform the unpenalized models. By introducing some extra bias the model, by means of the L_1 constraint, a sparse solution is obtained. This results in a model with better prediction capability.

Again, the main limitation of this thesis is that we only evaluated the models on one empirical data set. Therefore, we recommend experimenting with other empirical data sets to evaluate prediction error as a selection criterion for the MLD model.

In this thesis we used the Brier score to evaluate the prediction capability of the MLD model. however, other loss functions that can be utilized to evaluate the prediction capability of the model. For example, one might be interested in the use of the miss-classification rate or the cross entropy error as a loss function. An interesting subject for future research would be to study the implications of the choice of loss function for the MLD model. Here the question is how to choose the most suitable loss function for multivariate binary data.

Besides experimenting with different loss functions, we would suggest future research to study the use of the .632+ bootstrap estimator as proposed by Efron and Tibshirani (1997) to evaluate the expected prediction error of the model. The research of Efron and Tibshirani suggests that even a better estimate of the prediction error can be obtained by using the .632+ bootstrap estimator, yielding an estimate with low variance and only moderate bias.

In the original model of Worku and de Rooij the QIC_u norm is used to select between competing models. This is a measure of the penalized likelihood and the equivalent of the AIC, when

quasi likelihood is used. Cross-validation, in that sense, is a more versatile selection method not depending on a (quasi-) likelihood methodology. It should be noted that leave-one-out cross-validation (a variant of V-fold cross-validation in which V equals n) is asymptotically equivalent to AIC for ordinary linear regression models (Stone, 1977). However, there is no indication that this is true in the multivariate binary setting, especially not when other loss functions than the squared loss are used.

Appendices

A Verification truncated power basis

Let

$$h(x) = h_1(x) = \beta_{11}x + \beta_{12}x^2 + \beta_{13}x^3 \quad \text{if } x < \xi$$
$$h(x) = h_2(x) = \beta_{21}x + \beta_{22}x^2 + \beta_{23}x^3 \quad \text{if } x \geq \xi,$$

such that $h_1(\xi) = h_2(\xi), h_1'(\xi) = h_2'(\xi)$ and $h_1''(\xi) = h_2''(\xi)$.
Then,

$$\beta_{11}\,\xi + \beta_{12}\,\xi^2 + \beta_{13}\,\xi^3 = \beta_{21}\,\xi + \beta_{22}\,\xi^2 + \beta_{23}\,\xi^3$$
$$\beta_{11} + 2\beta_{12}\,\xi + 3\beta_{13}\,\xi^2 = \beta_{21} + 2\beta_{22}\,\xi + 3\beta_{23}\,\xi^2$$
$$2\beta_{12} + 6\beta_{13}\,\xi = 2\beta_{22} + 6\beta_{23}\,\xi$$

and from this follows:

$$\begin{aligned}\beta_{22} - \beta_{12} &= -3(\beta_{23} - \beta_{13})\xi \\ \beta_{21} - \beta_{11} &= -2(\beta_{22} - \beta_{12})\xi - 3(\beta_{23} - \beta_{13})\xi^2 \\ &= 6(\beta_{23} - \beta_{13})\xi^2 - 3(\beta_{23} - \beta_{13})\xi^2 \\ &= 3(\beta_{23} - \beta_{13})\xi^2\end{aligned}$$

and

$$\begin{aligned}(\beta_{23} - \beta_{13})\xi^3 &= -(\beta_{22} - \beta_{12})\xi^2 - (\beta_{21} - \beta_{11}\xi \\ &= 3(\beta_{23} - \beta_{13})\xi^3 - 3(\beta_{23} - \beta_{13})\xi^3 \\ &= 0.\end{aligned}$$

When we define $h(x)$ as $h_1(x) + (\beta_{23} - \beta_{13})(x - \xi)_+^3$ it can be shown using the above that:

$$h(x) = h_1(x) \quad \text{when } x < \xi$$
$$h(x) = h_2(x) \quad \text{when } x \geq \xi$$

Verification when $x \geq \xi$:

$$\begin{aligned}
h(x) &= h_1(x) + (\beta_{23} - \beta_{13})(x - \xi)_+^3 \\
&= \beta_{11}x + \beta_{12}x^2 + \beta_{13}x^3 + (\beta_{23} - \beta_{13})x^3 - 3(\beta_{23} - \beta_{13})\xi x^2 \\
&\quad + 3(\beta_{23} - \beta_{13})\xi^2 x - (\beta_{23} - \beta_{13})\xi^3 \\
&= \beta_{11}x + \beta_{12}x^2 + \beta_{13}x^3 + (\beta_{23} - \beta_{13})x^3 + (\beta_{22} - \beta_{12})x^2 \\
&\quad + (\beta_{21} - \beta_{11})x - (\beta_{23} - \beta_{13})\xi^3 \\
&= \beta_{21}x + \beta_{22}x^2 + \beta_{23}x^3 \\
&= h_2(x)
\end{aligned}$$

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike* (pp. 215–222). Springer.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, *16*(1), 125–127.

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63.

Anderson, D. A., & Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society: Series B (Methodological)*, *47*(2), 203–210.

Asar, Ö., & İlk, Ö. (2013). mmm: an R package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine*, *112*(3), 649–654.

Beesdo-baum, K., Höfler, M., Gloster, A. T., Klotsche, J., Lieb, R., Beauducel, A., … Wittchen, H. U. (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*(4), 204–220.

Blasius, J., Eilers, P. H., & Gower, J. (2009). Better biplots. *Computational Statistics & Data Analysis*, *53*(8), 3145–3158.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review*, *60*(3), 291–319.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, *78*(1), 1–3.

Brown, T. A., Campbell, L. A., Lehman, C. L., Grisham, J. R., & Mancill, R. B. (2001). Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *Journal of abnormal psychology*, *110*(4), 585.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chiou, J. M., & Müller, H. G. (2005). Estimated estimating equations: semiparametric inference for clustered and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(4), 531–553.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Curry, H. B., & Schoenberg, I. J. (1947). On spline distributions and their limits: the polya distributions, abstr. *Bulletin of the American Mathematical Society*, *53*(11), 1114–1114.

de Rooij, M. (2009). Ideal point discriminant analysis revisited with a special emphasis on visualization. *Psychometrika*, *74*(2), 317–330.

de Boor, C. (1978). *A practical guide to splines.* New York: Springer.

Deen, M., & de Rooij, M. (2019). Clusterbootstrap: An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap. *Behavior research methods*, 1–19.

Denison, D., Mallick, B., & Smith, A. (1998). Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 333–350.

Drost, J., Van der Does, W., van Hemert, A. M., Penninx, B. W., & Spinhoven, P. (2014). Repetitive negative thinking as a transdiagnostic factor in depression and anxiety: A conceptual replication. *Behaviour Research and Therapy*, *63*, 177–183.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, *92*(438), 548–560.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

Eilers, P. H., & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(6), 637–653.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: models, methods and applications*. Springer.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics, New York.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453–467.

Goepp, V., Bouaziz, O., & Nuel, G. (2018). Spline regression with automatic knot selection. *arXiv preprint arXiv:1808.01770*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gower, J. C., & Hand, D. J. (1995). *Biplots* (Vol. 54). CRC Press.

Groenen, P. J., Le Roux, N. J., & Gardner-Lubbe, S. (2015). Spline-based nonlinear biplots. *Advances in Data Analysis and Classification*, *9*(2), 219–238.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11.

Hand, D. J. (1999). Statistics and data mining: intersecting disciplines. *SIGKDD Explorations*, *1*(1), 16–19.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hastie, T., & Tibshirani, R. (1986, 08). Generalized additive models. *Statist. Sci.*, *1*(3), 297–310.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman and Hall, London.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman and Hall/CRC.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *I*, 221-233.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological Bulletin*, *136*(5), 768.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*(10), 921–926.

Matloff, N. (2017). *Statistical regression and classification: from linear models to machine learning*. Chapman and Hall/CRC.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer, New York.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Osborne, M. R., Presnell, B., & Turlach, B. A. (1998). Knot selection for regression splines via the lasso. *Computing Science and Statistics*, 44–49.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*,

57(1), 120–125.

Penninx, B. W., Beekman, A. T., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., … Assendelft, W. J. (2008). The netherlands study of depression and anxiety (nesda): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, *17*(3), 121–140.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Thuiller, W. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression* (No. 12). Cambridge university press.

Sherman, M., & Cessie, S. l. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics-Simulation and Computation*, *26*(3), 901–925.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Spinhoven, P., de Rooij, M., Heiser, W., Smit, J. H., & Penninx, B. W. (2009). The role of personality in comorbidity among anxiety and depressive disorders in primary care and specialty care: a cross-sectional analysis. *General Hospital Psychiatry*, *31*(5), 470–477.

Spinhoven, P., Penelo, E., De Rooij, M., Penninx, B., & Ormel, J. (2014). Reciprocal effects of stable and temporary components of neuroticism and affective disorders: results of a longitudinal cohort study. *Psychological Medicine*, *44*(2), 337–348.

Spinhoven, P., van der Does, W., Ormel, J., Zitman, F. G., & Penninx, B. W. (2013). Confounding of big five personality assessments in emotional disorders by comorbidity and current disorder. *European Journal of Personality*, *27*(4), 389–397.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, *40*(4), 961–971.

Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, *61*(3), 509–515.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 44–47.

Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika*, *52*(3), 371–392.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, *61*(3), 439–447.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Worku, H. M. (2018). *Distance models for analysis of multivariate binary data* (PhD thesis). Leiden University.

Worku, H. M., & de Rooij, M. (2018). A Multivariate Logistic Distance Model for the Analysis of Multiple Binary Responses. *Journal of Classification*, *146*, 1–23.

Wright, S. P. (1998). Multivariate analysis using the mixed procedure (paper 229–23). *in: Proceedings of the 23rd Annual SAS Users Group (SUGI) International Conference*.

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121–130.