The MELODIC family for simultaneous binary logistic regression in a reduced space

Mark de Rooij
Methodology & Statistics Unit, Institute of Psychology, Leiden University
Patrick J. F. Groenen
Econometric Institute, Erasmus School of Economics, Erasmus University

February 12, 2021

Please address correspondence to Mark de Rooij, Methodology and Statistics department, Institute of Psychology, Leiden University. PO Box 9555, 2300 RB Leiden, The Netherlands. Email: rooijm at fsw.leidenuniv.nl

Abstract

Logistic regression is a commonly used method for binary classification. Researchers often have more than a single binary response variable and simultaneous analysis is beneficial because it provides insight into the dependencies among response variables as well as between the predictor variables and the responses. Moreover, in such a simultaneous analysis the equations can lend each other strength, which might increase predictive accuracy. In this paper, we propose the MELODIC family for simultaneous binary logistic regression modeling. In this family, the regression models are defined in a Euclidean space of reduced dimension, based on a distance rule. The model may be interpreted in terms of logistic regression coefficients or in terms of a biplot. We discuss a fast iterative majorization (or MM) algorithm for parameter estimation. Two applications are shown in detail: one relating personality characteristics to drug consumption profiles and one relating personality characteristics to depressive and anxiety disorders. We present a thorough comparison of our MELODIC family with alternative approaches for multivariate binary data.

KEYWORDS: Multivariate logistic regression; Euclidean distance; Multidimensional Unfolding; MM algorithm.

1 Introduction

Logistic regression (Berkson, 1944; Cox, 1958; Agresti, 2003) is one of the most commonly used tools for binary classification. Although the logistic function has been known since the early 19th century, the logistic regression model was developed in the second half of the 20th century (Cramer, 2002). Adaptions of logistic regression models have been developed to make it more flexible, through basis expansion, or less flexible, by means of regularization. For an overview, see Friedman et al. (2001).

Researchers regularly have more than a single response variable. That is, there are applications where several response variables can be predicted from a common set of predictor variables (Breiman and Friedman, 1997). Examples include:

- The analysis of depressive and anxiety disorders. Mental disorders are highly prevalent in modern western societies and a high degree of comorbidity can often be observed among these disorders. In the Netherlands Study for Depression and Anxiety (Penninx et al., 2008) data were collected on a large number of subjects about their personality and about mental disorders (Spinhoven et al., 2009).

- The analysis of drug consumption profiles. Fehrman et al. (2017) are interested in subjects' drug consumption profiles and how these relate to personality characteristics such as sensation seeking and impulsivity. They collected data about the consumption of 18 different drugs.

- In clinical trials, the effect of treatments is established where the outcome can be dichotomous, cured or not cured. Treatments, however, come with side effects and these can be coded as present or absent. It is important

to study treatment and side effects together in order to obtain the whole picture. For an empirical example, see Molenberghs and Verbeke (2006).

- Psychosocial problems frequently occur in young adults. To screen for these problems in community settings, for example during large-scale general health check-ups, the Strengths and Difficulties Questionnaire (SDQ) can be used as it is a relatively short instrument. The SDQ has two parts: a self-report and a parent-report. To be useful as a screening tool it should have good validity properties, that is, it should be able to predict certain psychosocial problems. Vugteveen et al. (2018) investigated the validity of the SDQ with respect to four diagnoses.

With multiple binary outcomes it is possible to fit a logistic regression model separately for each outcome, but it is often wise to build a single multivariate model. In such a multivariate model the dependencies between the various outcomes can be better understood and strength can be borrowed between the different outcomes. Second, such a multivariate model is more parsimonious in the sense that less parameters have to be estimated. Furthermore, estimated regression weights may be better in terms of mean squared error. Stein et al. (1956), for example, showed that simple averages of a multivariate normal distribution are inferior to shrunken averages in terms of mean squared error, where the shrinkage tends toward the average of averages. Breiman and Friedman (1997) show that shrinkage of coefficients of several multiple regression models toward each other is beneficial in terms of predictive accuracy. In a similar vein, building several logistic regressions in a reduced space might provide better estimates of the regression coefficients in terms of mean squared error.

There are basically two broad ways of analyzing multivariate data and performing dimension reduction: The first is based on inner products from which principal component analysis (Pearson, 1901; Hotelling, 1936; Jolliffe, 2002) and

reduced rank regression (Izenman, 1975; Ter Braak and Looman, 1994) are derived; the second is based on distances which have led to multidimensional scaling (Torgerson, 1952, 1958; Gower, 1966; Guttman, 1968) and multidimensional unfolding (Coombs, 1950; Roskam, 1968; Heiser, 1981; Busing, 2010). This distance framework is conceptually easier than the inner product framework and leads to more straightforward interpretation (De Rooij and Heiser, 2005). Distances, especially Euclidean and Manhattan, are all around us and can already be understood by very young children.

In multidimensional unfolding, we generally have a dissimilarity matrix between two sets of objects. The goal is to find a low-dimensional mapping including points for the row objects and the column objects such that the distances between the points of the two sets are as close as possible (often in the "least squares" sense) to the observed dissimilarities. We will develop a family of models based on similar ideas.

In this paper, we will develop a family of logistic models within a distance framework. We call it the MELODIC family, written out, the MultivariatE LOgistic DIstance to Categories family. More specifically, we will develop a framework of models in which both participants and the categories of the different response variables have a position in low-dimensional Euclidean space. The distances between the position of a participant and the positions of the two categories of a single response variable determine the probabilities for these two response options. The position of a participant will be parameterized as a linear combination of the predictor variables.

The family extends the recently proposed multivariate logistic distance models (Worku and De Rooij, 2018) which built on earlier logistic distance models (Takane et al., 1987; Takane, 1987; De Rooij, 2009) and can be considered as examples in the "Gifi goes logistic" framework as laid out by De Leeuw (2005)

and Evans (2014).

In the next section, we will develop the general model and two constrained variants. We will discuss properties of the models and provide interpretational rules. Two types of these rules can be distinguished: the numerical and the graphical. These two modes of interpretation for a single model are beneficial because there are those people who say that "a graph is worth a thousand words", the so-called *graph people* (Friendly and Kwan, 2011), while others (the *table people*) firmly disagree (Gelman, 2011). In Section 3, we develop an Iterative Majorization or MM algorithm (De Leeuw and Heiser, 1977; Groenen, 1993; Heiser, 1995; Hunter and Lange, 2004) for estimating the parameters of our models by minimizing a deviance function. Section 4 describes two illustrative applications. In Section 5, we discuss related statistical models and provide some comparisons. We conclude, in Section 6, with a general discussion of our developments and some possibilities for further investigation.

2 MELODIC family

2.1 Data and notation

We consider a system with P explanatory, predictor, or independent variables X and R outcome, response, or dependent binary variables Y. That is, we have a sample of observations $\{\mathbf{x}_i, \mathbf{y}_i\}_1^n$ with $\mathbf{x}_i \in \mathbb{R}^P$ and $\mathbf{y}_i \in \{0,1\}^R$. The response variables will be recoded into indicator vectors $\mathbf{g}_{ir}$ of length two, where the first element equals 1 if $y_{ir} = 0$ and the second element equals 1 if $y_{ir} = 1$.

We will use the following notation.

- $i = 1, \ldots, n$ for individuals (participants, subjects, objects).
- $p = 1, \ldots, P$ for predictor variables (explanatory or independent variables).

- $r = 1, \ldots, R$ for response variables (outcome or dependent variables).
- $m = 1, \ldots, M$ an indicator for the dimensions.
- There is a set of predictor variables $X = \{X_p\}_{p=1}^{P}$. Observed values of the predictor variables are collected in the $n \times P$ matrix $\mathbf{X}$ with elements x_{ip}. We assume, without loss of generality, that the predictor variables are centered, that is $\mathbf{1}^\top \mathbf{X} = \mathbf{0}$.

- There is a set of response variables $Y = \{Y_r\}_{r=1}^{R}$. Observed values of the response variables are collected in the $n \times R$ matrix $\mathbf{Y}$. The matrix has elements $y_{ir} \in \{0, 1\}$. We will code the responses in a super indicator matrix $\mathbf{G}$ having $C = 2R$ categories, that is,

$$\mathbf{G} = [\mathbf{G}_1 | \mathbf{G}_2 | \ldots | \mathbf{G}_R].$$

- $\mathbf{B}$ represents a $P \times M$ matrix with regression weights for the predictor variables.
- $\mathbf{u}_i$ is an M vector with coordinates for person i in M-dimensional Euclidean space. These coordinates will be collected in the $n \times M$ matrix $\mathbf{U}$ with elements u_{im}.
- $\mathbf{V}_r$ is a $2 \times M$ matrix having the coordinates of category 0 (i.e., $\mathbf{v}_{r0}$) in the first row and in the second row the coordinates of category 1 (i.e., $\mathbf{v}_{r1}$), both for response variable r. These matrices will be collected in the $2R \times M$ matrix $\mathbf{V} = \left[\mathbf{V}_1^\top, \ldots, \mathbf{V}_R^\top\right]^\top$ with elements v_{rcm}.
- The observations are $\{\mathbf{x}_i, \mathbf{y}_i\}_1^n$.
- We define a block diagonal matrix $\mathbf{J}$ with 2×2 diagonal blocks $\mathbf{I}_2 - \frac{1}{2}\mathbf{1}\mathbf{1}^\top$.

- We use tildes for current estimates in the iterative process, that is, $\widetilde{\mathbf{B}}$ represents the matrix with estimates in a given cycle of the algorithm.
- Diag() denotes the operator that takes the diagonal values of a matrix and places them in a vector.

2.2 General model

We define the conditional probability that person i is in class c ($c = \{0, 1\}$) of response variable r, $\pi_{rc}(\mathbf{x}_i) = P(Y_{ir} = c|\mathbf{x}_i)$ as

$$\pi_{rc}(\mathbf{x}_i) = \frac{\exp(-\delta(\mathbf{u}_i, \mathbf{v}_{rc}))}{\exp(-\delta(\mathbf{u}_i, \mathbf{v}_{r0})) + \exp(-\delta(\mathbf{u}_i, \mathbf{v}_{r1}))}, \tag{1}$$

where $\delta(\cdot, \cdot)$ denotes half the squared Euclidean distance

$$\delta(\mathbf{u}_i, \mathbf{v}_{rc}) = \frac{1}{2} \sum_{m=1}^{M} (u_{im} - v_{rcm})^2 = \frac{1}{2} \sum_{m=1}^{M} \left(u_{im}^2 + v_{rcm}^2 - 2u_{im}v_{rcm}\right), \tag{2}$$

in M-dimensional Euclidean space. The dimensionality M has to be chosen by the researcher with possible values being between 1 and $\min(P, R)$. The coordinates of the subjects ($\mathbf{u}_i$) are assumed to be a linear combination of the predictor variables, that is, $\mathbf{u}_i = \mathbf{x}_i^\top \mathbf{B}$, where $\mathbf{B}$ is a $P \times M$ matrix with regression weights. The coordinates of category c of response variable r on dimension m are denoted by v_{rcm} and collected in the M-vector $\mathbf{v}_{rc}$.

Every subject i is thus represented in an M-dimensional Euclidean space. Moreover, this subject has a distance to a point representing category 1 of response variable r and to a point representing category 0 of response variable r. These two distances determine the probability for the subject to answer with either of these categories; the smaller the distance, the larger the probability. In other words, a subject is most likely to be in the closest class.

The log odds in favor of the 1 category and against category 0 for response variable r given the subject's position is given by

$$\log \frac{\pi_{r1}(\mathbf{x}_i)}{1-\pi_{r1}(\mathbf{x}_i)} = \log \frac{\pi_{r1}(\mathbf{x}_i)}{\pi_{r0}(\mathbf{x}_i)} = \delta(\mathbf{u}_i, \mathbf{v}_{r0}) - \delta(\mathbf{u}_i, \mathbf{v}_{r1}), \tag{3}$$

a simple difference of squared Euclidean distances. This log odds can be further worked out as

$$\log \frac{\pi_{r1}(\mathbf{x}_i)}{\pi_{r0}(\mathbf{x}_i)} = \sum_{m=1}^{M} \left[\frac{1}{2}(v_{r0m}^2 - v_{r1m}^2) + \mathbf{x}_i^\top \mathbf{b}_m (v_{r1m} - v_{r0m}) \right], \tag{4}$$

where we see that the effect of predictor variable p on response variable r is determined by the regression coefficients b_{pm} and the distance between the two categories. In general, the further apart the two categories are, the better they are discriminated by the predictor variables. If the two categories fall on the same position in the Euclidean space, they are *indistinguishable* (Anderson, 1984) based on this set of predictor variables.

Let us define $a_r^* = \frac{1}{2}\sum_{m=1}^{M}(v_{r0m}^2 - v_{r1m}^2)$ and $\mathbf{b}_r^* = \sum_{m=1}^{M} \mathbf{b}_m(v_{r1m} - v_{r0m})$. Then the log odds can be written as

$$\log \frac{\pi_{r1}(\mathbf{x}_i)}{\pi_{r0}(\mathbf{x}_i)} = a_r^* + \mathbf{x}_i^\top \mathbf{b}_r^*, \tag{5}$$

showing that the model can be interpreted as standard binary logistic regression models. We call the b^* the model *implied coefficients*.

2.3 Constrained models

In the general model described above, the categories of the response variables lie freely somewhere in the M-dimensional space. Sometimes, however, researchers already have an idea about the underlying structure of the response

variables. In the literature about depressive and anxiety disorders, for example, one theory says that fear and distress are its underlying dimensions, where each dimension comprises a subset of the disorders. In terms of our models, this means that the categories of the response variables pertaining to the distress dimension lie on a single dimension (i.e., the coordinates of the categories of these response for the other dimensions equal zero). Similarly, for the categories of the response variables pertaining to the fear dimension, the coordinates on the distress dimension all equal zero.

If a specific response variable pertains to, say, dimension 1, the class coordinates on all other dimensions are set to zero, that is $v_{r1m} = v_{r0m} = 0, \forall m \neq 1$. Such a structure simplifies the model and its interpretation, because in the log odds definition (see Equation 4) the last term becomes zero for several dimensions and only the regression weights of the dimension to which the response variable pertains are important for the discrimination of the categories of that response variable.

One further constraint is to let all response variables have the same discriminatory ability. In that case, $(v_{r1m} - v_{r0m}) = 1$ for the dimensions to which response variable r pertains. For this constrained model, Worku and De Rooij (2018) showed that the parameters can be estimated using standard software for logistic regression by using a structured design matrix for the predictors. This paper describes them as members of a larger family of models, the MELODIC family.

2.4 Graphical representation

When the dimensionality equals two ($M = 2$), the model can be easily represented graphically. This representation shows 1) the categories of the response variables as points, 2) a decision line for every response variable designating

the predicted class at a specific point, 3) variable axes for the predictor variables, and 4) the subjects' positions as points. Many aspects of the interpretation of these graphical representations follow the theory of biplots as discussed in Gower and Hand (1995) and Gower et al. (2011).

Let us first look at a graphical representation for a single response variable and a set of subjects, of which three of them are highlighted. Figure 1 gives such a graph where A0 and A1 are the two categories of a response variables named A, and i, j and k present three subjects. The line halfway between classes A0 and A1 represents the decision line, in other words the line represents the points for which the odds are even. The log odds that subject i chooses A0 instead of A1 are clearly in favor of class A1, because that is the closest class.

The squared distances from Subject i to categories A0 and A1 additively decompose into one part toward the line through A0 and A1 (i.e., the A01 line) and one part along this line. Equation 3 shows that the log odds are defined in terms of a difference in squared distances, and therefore the part toward the A01 line drops out of the equation. In more detail, for this example we have

$$\log \frac{\pi_{A0}(\mathbf{x}_i)}{\pi_{A1}(\mathbf{x}_i)} \quad = \quad \delta(\mathbf{u}_i, \mathbf{v}_{A1}) - \delta(\mathbf{u}_i, \mathbf{v}_{A0}).$$

According to the Pythagorean theorem, the squared distance $\delta(\mathbf{u}_i, \mathbf{v}_{A1})$ can be decomposed into $\delta(\mathbf{u}_i, \mathbf{v}_{A01}) + \delta(\mathbf{v}_{A01}, \mathbf{v}_{A1})$, where $\mathbf{v}_{A01}$ is the coordinate of the projection of $\mathbf{u}_i$ on the A01 line. Using this decomposition for both terms we obtain

$$\begin{aligned} \log \frac{\pi_{A0}(\mathbf{x}_i)}{\pi_{A1}(\mathbf{x}_i)} &= (\delta(\mathbf{u}_i, \mathbf{v}_{A01}) + \delta(\mathbf{v}_{A01}, \mathbf{v}_{A1})) - (\delta(\mathbf{u}_i, \mathbf{v}_{A01}) + \delta(\mathbf{v}_{A01}, \mathbf{v}_{A0})) \\ &= \delta(\mathbf{v}_{A01}, \mathbf{v}_{A1}) - \delta(\mathbf{v}_{A01}, \mathbf{v}_{A0}). \end{aligned}$$

For person j or k, we can use the same decomposition. The projections for the

three subjects are however equal, and therefore the log odds of category A0 against A1 for persons i, j, and k are equal (and for all three subjects in favor of category A1). As noted above, the decision line represents the set of positions where the odds are even, that is, the log odds are equal to zero. More generally, iso log odds curves, which are curves where the log odds equal any constant, are straight lines parallel to these decision lines and orthogonal to the A01 line. An example of such an iso log odds line is the one through the points representing the three subjects (blue dotted).

The variable axes can be understood as representations of subjects with varying scores on the corresponding predictor variables and an average score on all other predictor variables. In this way we can interpret the variable axis by moving along the variable axes and computing the log odds for each response variable. More formally, let us denote by $\mathbf{d}_r$ the M-vector with differences $(v_{r1m} - v_{r0m})$ and let predictor variable p be presented by its regression weights $\mathbf{b}_p$ (with $\mathbf{b}_p^\top$ row p of matrix $\mathbf{B}$), then we can write the effect of predictor variable p on response variable r as

$$\mathbf{b}_p^\top \mathbf{d}_r = \|\mathbf{b}_p\| \cdot \|\mathbf{d}_r\| \cdot \cos(\mathbf{b}_p, \mathbf{d}_r),$$

showing that the log odds are largest when the direction of the variable axis for predictor variable p is parallel to the line connecting the two categories of response variable r, while it is zero if the variable axis is orthogonal to this line. Figure 2 illustrates this property. In Figure 2, the variable axes are represented for four predictor variables. We use the convention that the labels attached to the variable axes are placed on the positive side of the variable. The two categories of a response variable (A0 and A1) are depicted by points. The variable axis for predictor variable X_3 is parallel to the A01 line, indicating that this variable discriminates this response variable well, while the variable axis for predictor

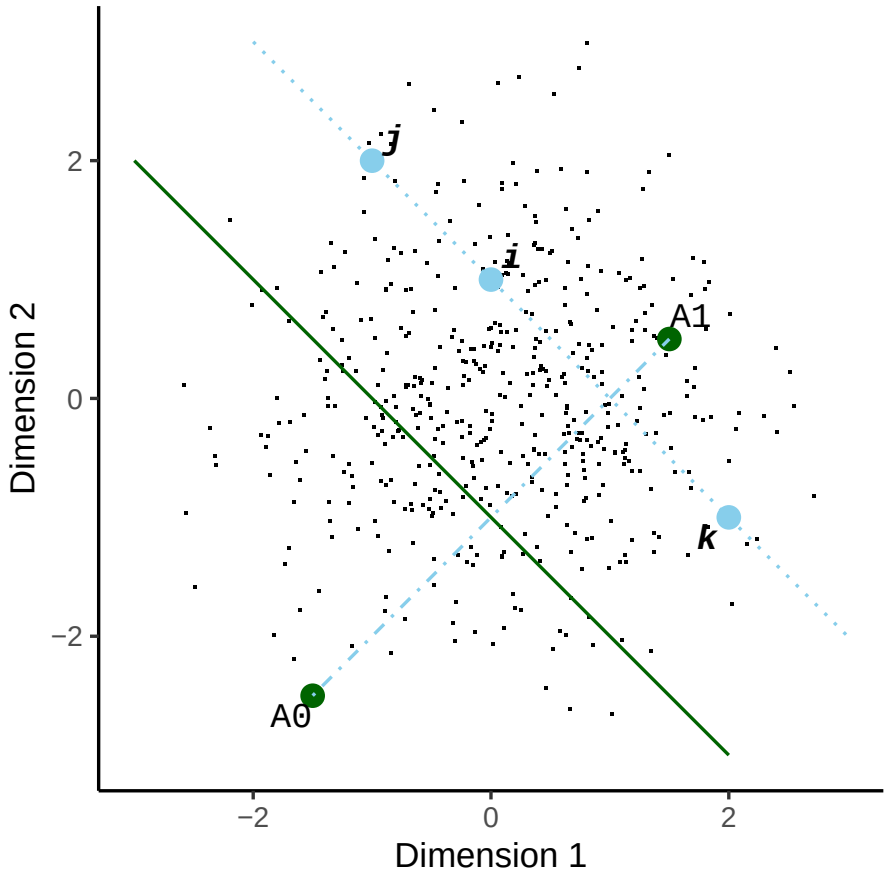

Figure 1: Graphical representation with a single dichotomous response variable A (with categories A0 and A1) and three participants (i, j, and k). The solid green line represents the decision line where the probabilities for A0 and A1 are equal. The blue dotted line projects the points representing the three subjects onto the A01 line (blue dashed-dotted line). All points on this dotted line represent observations with the same log odds.

X_2 is almost orthogonal to the A01 line, indicating that X_2 does not discriminate between these two classes. We could draw the projections of A0 and A1 on each of the variable axes to see the discriminatory power: the further apart these projections are, the higher the power. For example, the projections of the two points onto variable X_2 are very close to each other, indicating that X_2 does not discriminate these two classes well.

The discriminative power depends not only on the distance between the projections but also on the estimated value of the regression coefficients. Larger regression weights indicate more general discriminative power for the complete set of response variables. We will indicate the value of the regression weight by using markers along the variable axis in steps of 1 standard deviation. The further apart these markers are, the larger the regression weights and the higher the discriminative power will be.

With R binary response variables, the number of different possible response profiles is 2^R. When $M = R$, each of these profiles can be perfectly represented. In lower dimensional space ($M < R$), however, not all response profiles find a place in the solution. For example, in a one-dimensional space, only $R + 1$ different response profiles are represented. In two-dimensional space, the number of represented profiles is

$$\sum_{m=1}^{2} \binom{R}{m}$$

(Coombs and Kao, 1955).

In the constrained models, the number of represented response profiles is lower than in the general model because all decision lines are either horizontal or vertical.[1] With five response variables, of which three pertain to the first dimension and two to the second, the model represents $4 \times 3 = 12$ response profiles, which is even smaller than the 16 in the general unconstrained model,

[1] Assuming we do not have a response variable pertaining to multiple dimensions.

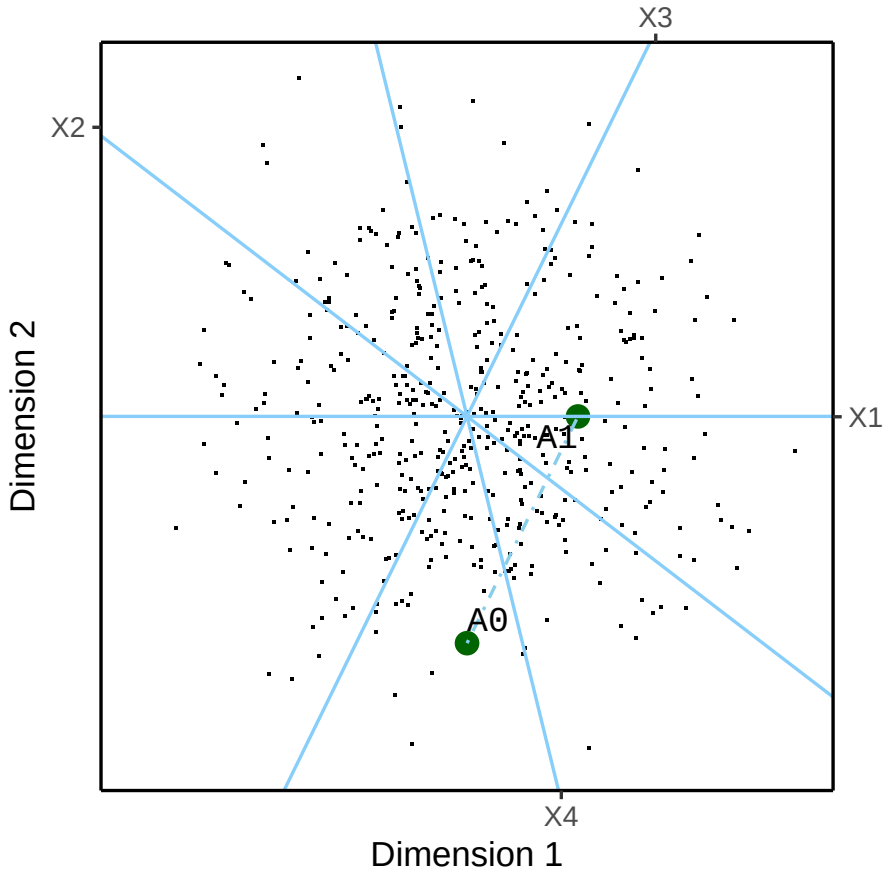

Figure 2: Graphical representation with the two class points of a single response variable named A (the classes are represented by the points labeled A0 and A1), variable axes for four predictor variables, and subject points (small dots). The dashed-dotted line connecting A0 and A1 represents the vector $\mathbf{d}_A$. Predictor variable X_3 discriminates well between the two classes because the direction is parallel to the line joining the two classes, whereas predictor variable X_2 hardly discriminates between the two classes because the variable axis is almost orthogonal to the line joining the two classes.

and much smaller than the $2^R = 32$ possible response profiles.

3 An MM Algorithm

In this section, we develop an MM algorithm, where the first M stands for "majorize" and the second M for "minimize". Such algorithms are also known as iterative majorization (IM) algorithms. MM algorithms have the property of guaranteed descent and in MM algorithms it is easy to use low rank restrictions. The global idea of MM algorithms is that, instead of minimizing the original loss function, we seek an auxiliary function that 1) touches the original function at the current estimates, 2) lies above the original function, and 3) is easy to minimize. For a detailed treatment of the general principles of IM or MM, we refer to Heiser (1995) and Hunter and Lange (2004). In the following subsections, we will majorize a deviance function with a least squares function. This results in a fast algorithm. Before developing the algorithm, we will discuss identification of model parameters.

3.1 Admissible Transformations

Before we develop an algorithm for the estimation of model parameters, we must discuss indeterminacies, that is, admissible transformations that change neither the estimated probabilities nor the loss value.

1. Multidimensional scaling and unfolding models in general have *translational freedom*. We center $\mathbf{X}$ so that the origin of the Euclidean space is fixed at the average value of the predictor variables.

2. The model has *rotational freedom*: any map can be rotated without changing the distances or the probabilities. We will require that $\frac{1}{n}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B} = \mathbf{I}$ so that the rotational indeterminacy is removed. Reflection can be removed

by requiring that the regression weights for the first predictor variable are positive.

3. The model

$$\pi_{rc}(\mathbf{x}_i) = \frac{\exp(\theta_{irc})}{\exp(\theta_{ir0}) + \exp(\theta_{ir1})}$$

is invariant under an additive constant in the "linear predictor" (θ_{irc}) , that is

$$\frac{\exp(\theta_{irc})}{\exp(\theta_{ir0}) + \exp(\theta_{ir1})} = \frac{\exp(\theta_{irc} + \zeta_r)}{\exp(\theta_{ir0} + \zeta_r) + \exp(\theta_{ir1} + \zeta_r)}. \tag{6}$$

For our distance model, $\theta_{irc} = -\delta(\mathbf{u}_i, \mathbf{v}_{rc})$, we can add a constant to the squared distances *per response variable*, implying that the term $\sum_{m=1}^{M} u_{im}^2$ can be removed from the distance formulation. We will see a further simplification below.

3.2 Algorithm for the unconstrained model

The deviance function to be minimized is

$$\begin{aligned} L(\mathbf{B}, \mathbf{V}) &= -2 \sum_{i=1}^{n} \sum_{r=1}^{R} \sum_{c=0}^{1} g_{irc} \log \pi_{rc}(\mathbf{x}_i) \\ &= -2 \sum_{i=1}^{n} \sum_{r=1}^{R} \sum_{c=0}^{1} g_{irc} \log \left(\frac{\exp(\theta_{irc})}{\exp(\theta_{ir0}) + \exp(\theta_{ir1})} \right), \end{aligned} \tag{7}$$

where for our model $\theta_{irc} = -\frac{1}{2}\delta(\mathbf{u}_i, \mathbf{v}_{rc})$. Groenen et al. (2003), De Leeuw (2006), and Groenen and Josse (2016) show that the function

$$f_{ir}(\boldsymbol{\theta}_i) = -2 \sum_{c=0}^{1} g_{irc} \log \frac{\exp(\theta_{irc})}{\exp(\theta_{ir0}) + \exp(\theta_{ir1})},$$

is majorized by

$$g_{ir}(\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}_i) = f_{ir}(\tilde{\boldsymbol{\theta}}_i) + (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^\top \nabla f_{ir}(\tilde{\boldsymbol{\theta}}_i) + \frac{1}{4}\|\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i\|^2,$$

with

$$\nabla f_{ir}(\tilde{\boldsymbol{\theta}}_i) = -(\mathbf{g}_{ir} - \tilde{\boldsymbol{\pi}}_{ir}),$$

and where the tilde means that it is evaluated using the current estimates. A proof of the majorizing property is given in the appendix of Groenen and Josse (2016).

By analogy to Groenen and Josse, we therefore have that

$$\begin{aligned} L(\mathbf{B}, \mathbf{V}) &\leq \frac{1}{4}\|\mathbf{Z} - \boldsymbol{\Theta}\|^2 + L(\widetilde{\mathbf{B}}, \widetilde{\mathbf{V}}) - \frac{1}{4}\|\mathbf{Z}\|^2 + \left\|\mathbf{G} - \widetilde{\boldsymbol{\Pi}}\right\|^2 \\ &= \frac{1}{4}\|\mathbf{Z} - \boldsymbol{\Theta}\|^2 + \text{constant} = g(\mathbf{B}, \mathbf{V}) + \text{constant}, \end{aligned} \tag{8}$$

where $\mathbf{Z} = \{z_{irc}\}$ with $z_{irc} = \tilde{\theta}_{irc} + 2(g_{irc} - \tilde{\pi}_{irc})$ and $\boldsymbol{\Theta} = \{\theta_{irc}\}$ with $\theta_{irc} = -\frac{1}{2}\sum_{m=1}^{M}(v_{rcm}^2 - 2u_{im}v_{rcm})$. In matrix terms we can write $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta} = -\frac{1}{2}\left(\mathbf{1}\mathbf{d}_v^\top - 2\mathbf{X}\mathbf{B}\mathbf{V}^\top\right),$$

and $\mathbf{Z}$ as

$$\mathbf{Z} = \widetilde{\boldsymbol{\Theta}} + 2(\mathbf{G} - \widetilde{\boldsymbol{\Pi}}).$$

Therefore, the objective function to be minimized in every iteration is

$$g(\mathbf{B}, \mathbf{V}) = \left\|\mathbf{Z} + \mathbf{1}\mathbf{d}_v^\top/2 - \mathbf{X}\mathbf{B}\mathbf{V}^\top\right\|^2.$$

The third indeterminacy, outlined above, allows us to rewrite the minimiza-

tion function as

$$g(\mathbf{B}, \mathbf{V}) = \left\| \mathbf{ZJ} + \frac{1}{2}\mathbf{1}\mathbf{d}_v^\top \mathbf{J} - \mathbf{XBV}^\top \mathbf{J} \right\|^2, \tag{9}$$

with $\mathbf{J}$ a symmetric block diagonal matrix with 2×2 diagonal blocks $\mathbf{J}_r$, the usual centering matrices, $\mathbf{J}_r = \mathbf{I}_2 - \frac{1}{2}\mathbf{1}\mathbf{1}^\top$.

Let us define the matrices $\mathbf{A}_l = \mathbf{I}_R \otimes [1, 1]^\top$ and $\mathbf{A}_k = \mathbf{I}_R \otimes [1, -1]^\top$ with $\otimes$ the Kronecker product, such that $\mathbf{V}$ can be reparametrized as

$$\mathbf{V} = \mathbf{A}_l\mathbf{L} + \mathbf{A}_k\mathbf{K}$$

with $\mathbf{L}$ the $R \times M$ matrix with response variable *locations* and $\mathbf{K}$ the $R \times M$ matrix representing the *discriminatory power* for the response variables. As a numerical example with two response variables, consider

$$\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 3 & 2 \\ 0 & 0 \\ 4 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -2 & -3 \end{bmatrix},$$

where the second matrix on the right-hand side of the equation represents $\mathbf{L}$ (the locations or midpoints of the class coordinates) and the last matrix on the right-hand side of the equation represents $\mathbf{K}$. The larger the absolute values in row r of matrix $\mathbf{K}$, the better the two categories from this response variable (r) can be discriminated by the predictor variables; if the values equal zero, the categories cannot be discriminated and the positions of the two categories of a response variable fall in the same place. In the numerical example, the second response variable is better discriminated (that is, the class points are further apart).

Using this reparametrization in our loss function, the last term $\mathbf{XBV}^\top\mathbf{J}$ simplifies to $\mathbf{XBV}^\top\mathbf{J} = \mathbf{XBK}^\top\mathbf{A}_k^\top$ because the term $\mathbf{XBL}^\top\mathbf{A}_l^\top\mathbf{J} = \mathbf{0}$ as $\mathbf{A}_l^\top\mathbf{J} = \mathbf{0}$. Let us now have a closer look at the second term of the loss function (Equation 9), that is, $\mathbf{1}\mathbf{d}_v^\top\mathbf{J} = \mathbf{J}\mathbf{d}_v$. This term can be rewritten as

$$\begin{aligned}\mathbf{d}_v = \mathrm{Diag}(\mathbf{V}\mathbf{V}^\top) &= \mathrm{Diag}\left((\mathbf{A}_l\mathbf{L} + \mathbf{A}_k\mathbf{K})(\mathbf{A}_l\mathbf{L} + \mathbf{A}_k\mathbf{K})^\top\right) \\ &= \mathrm{Diag}\left(\mathbf{A}_l\mathbf{L}\mathbf{L}^\top\mathbf{A}_l^\top + \mathbf{A}_k\mathbf{K}\mathbf{K}^\top\mathbf{A}_k^\top + \mathbf{A}_l\mathbf{L}\mathbf{K}^\top\mathbf{A}_k^\top + \mathbf{A}_k\mathbf{K}\mathbf{L}^\top\mathbf{A}_l^\top\right),\end{aligned}$$

where $\mathrm{Diag}(\mathbf{X})$ creates a column vector of the diagonal elements of $\mathbf{X}$. It can be verified that the terms $\mathbf{J}\mathrm{Diag}\left(\mathbf{A}_l\mathbf{L}\mathbf{L}^\top\mathbf{A}_l^\top\right)$ and $\mathbf{J}\mathrm{Diag}\left(\mathbf{A}_k\mathbf{K}\mathbf{K}^\top\mathbf{A}_k^\top\right)$ are both equal to zero. Therefore

$$\mathrm{Diag}(\mathbf{V}\mathbf{V}^\top) = \mathbf{J}\mathrm{Diag}\left(2\mathbf{A}_k\mathbf{K}\mathbf{L}^\top\mathbf{A}_l^\top\right).$$

Focusing on a single response variable r, we can rewrite the corresponding part of the previous equation as

$$\mathbf{k}_r^\top\mathbf{l}_r\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix},$$

from which it follows that $\frac{1}{2}\mathbf{1}\mathbf{d}_v\mathbf{J}$ can be written as

$$\frac{1}{2}\mathbf{1}\mathbf{d}_v^\top\mathbf{J} = \mathbf{1}\mathrm{Diag}(\mathbf{K}\mathbf{L}^\top)^\top\mathbf{A}_k^\top.$$

Going back to our loss function and using the decomposition into $\mathbf{K}$ and $\mathbf{L}$, it becomes

$$g(\mathbf{B}, \mathbf{K}, \mathbf{L}) = \left\|\mathbf{Z}\mathbf{J} + \mathbf{1}\mathrm{Diag}(\mathbf{K}\mathbf{L}^\top)^\top\mathbf{A}_k^\top - \mathbf{X}\mathbf{B}\mathbf{K}^\top\mathbf{A}_k^\top\right\|^2,$$

that equals

$$g(\mathbf{B},\mathbf{K},\mathbf{L}) = 2\left\|\frac{1}{2}\mathbf{ZJA}_k + \mathbf{1}\mathrm{Diag}(\mathbf{KL}^\top)^\top - \mathbf{XBK}^\top\right\|^2$$

and can be rewritten as

$$g(\mathbf{B},\mathbf{K},\mathbf{L}) = 2\left\|\frac{1}{2}\mathbf{ZJA}_k + \mathbf{1a}^\top - \mathbf{XBK}^\top\right\|^2,$$

where $\mathbf{a}$ is a vector with elements $a_r = \sum_{m=1}^{M} k_{rm}l_{rm}$. The elements of $\mathbf{a}$ can be estimated independently of $\mathbf{K}$, because values of $\mathbf{L}$ always exist that together with the k_{rm} can reconstruct a_r (see below). Therefore, this latter loss function can be solved separately for 1) $\mathbf{a}$ and 2) $\mathbf{B},\mathbf{K}$.

To update $\mathbf{a}$, define

$$\tilde{\mathbf{Z}}_1 = \frac{1}{2}\mathbf{ZJA}_k,$$

then the update is $\mathbf{a}^+ = -\tilde{\mathbf{Z}}_1^\top\mathbf{1}/n$.

To find the update for $\mathbf{B}$ and $\mathbf{K}$, let us define

$$\tilde{\mathbf{Z}}_2 = \frac{1}{2}\mathbf{ZJA}_k + \mathbf{1a}^\top,$$

so that we have to minimize

$$\left\|\tilde{\mathbf{Z}}_2 - \mathbf{XBK}^\top\right\|^2,$$

under the restriction that $n^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{XB} = \mathbf{I}$. As

$$\left\|\tilde{\mathbf{Z}}_2 - \mathbf{XBK}^\top\right\|^2 = \left\|\tilde{\mathbf{Z}}_2 - \mathbf{XN}\right\|^2 + \left\|\mathbf{N} - \mathbf{BK}^\top\right\|^2_{\mathbf{X}^\top\mathbf{X}}$$

with $\mathbf{N} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\tilde{\mathbf{Z}}_2$, the unconstrained update, an update of $\mathbf{B}$ and $\mathbf{K}$ is found by the generalized singular value decomposition of $\mathbf{N}$ (Takane, 2013).

These two steps can be combined, that is,

$$\mathbf{R}_x^{-1}\mathbf{X}^\top\tilde{\mathbf{Z}}_2 = \mathbf{P}\mathbf{\Phi}\mathbf{Q}^\top,$$

where $\mathbf{R}_x$ is the matrix square root of the matrix $\mathbf{X}^\top\mathbf{X}$, that is $\mathbf{X}^\top\mathbf{X} = \mathbf{R}_x\mathbf{R}_x^\top$. The updates for $\mathbf{B}$ and $\mathbf{K}$ can be obtained as

$$\mathbf{B}^+ = \sqrt{n}\mathbf{R}_x^{-1}\mathbf{P}_M,$$

where $\mathbf{P}_M$ denotes the M columns of $\mathbf{P}$ corresponding to the M largest singular values and

$$\mathbf{K}^+ = \frac{1}{\sqrt{n}}\mathbf{Q}_M\mathbf{\Phi}_M.$$

Finally, an update for $\mathbf{L}$ can be obtained from $\mathbf{a}$ and $\mathbf{K}$. For every response variable we have

$$a_r = \sum_{m=1}^{M} k_{rm}l_{rm},$$

which is an unidentified system, that is, there are many choices of l_{rm} that provide a solution. These solutions correspond to any position on the decision line (or plane or hyperplane in higher-dimensional spaces), as discussed in relation to Figure 1. We find the position on this hyperplane that is closest to the origin of the Euclidean space, that is,

$$l_{rm}^+ = \frac{a_r k_{rm}}{\sum_m k_{rm}^2}.$$

A summary of the algorithm can be found in Algorithm 1.

The number of parameters of the model is:

- $PM - M(M+1)/2$ for the regression weights $\mathbf{B}$;
- RM parameters in the matrix $\mathbf{K}$;

- R parameters in $\mathbf{L}$;

which sum to $(P+R)M + R - M(M+1)/2$.

Data: $\mathbf{X}$, $\mathbf{G}$, M

Result: $\mathbf{B}$, $\mathbf{K}$, $\mathbf{L}$

Compute: $\mathbf{R}_x^{-1}\mathbf{X}^\top\mathbf{G} = \mathbf{P}\mathbf{\Phi}\mathbf{Q}^\top$;

Initialize: $\mathbf{B}^{(0)} = \sqrt{n}\mathbf{R}_x^{-1}\mathbf{P}_M$;

Compute: $\mathbf{V} = \frac{1}{\sqrt{n}}\mathbf{Q}_M\mathbf{\Phi}_M$;

Initialize: $\mathbf{K}^{(0)}$ by taking the uneven rows of $\mathbf{JV}$;

Initialize: $\mathbf{L}^{(0)}$ by taking the uneven rows of $(\mathbf{I}-\mathbf{J})\mathbf{V}$;

Compute: $\tilde{\mathbf{\Pi}}$ and $\tilde{\mathbf{\Theta}}$;

while $t = 0$ *or* $(L^t - L^{(t-1)})/L^t > 10^{-8}$ **do**

  $t = t + 1$;

  Compute: $\mathbf{Z} = \tilde{\mathbf{\Theta}} + 2(\mathbf{G} - \tilde{\mathbf{\Pi}})$;

  Compute: $\mathbf{Z}_1 = \frac{1}{2}\mathbf{ZJA}_k$;

  Compute: $\mathbf{a} = -\mathbf{Z}_1^\top\mathbf{1}/n$;

  Compute: $\mathbf{Z}_2 = \frac{1}{2}\mathbf{ZJA}_k + \mathbf{1a}^\top$;

  Compute: $\mathbf{R}_x^{-1}\mathbf{X}^\top\mathbf{Z}_2 = \mathbf{P}\mathbf{\Phi}\mathbf{Q}^\top$;

  Update: $\mathbf{B}^{(t)} = \sqrt{n}\mathbf{R}_x^{-1}\mathbf{P}_M$;

  Update: $\mathbf{K}^{(t)} = \frac{1}{\sqrt{n}}\mathbf{Q}_M\mathbf{\Phi}_M$;

  Update: $l_{rm}^{(t)} = a_r k_{rm}/(\sum_m k_{rm}^2), \forall r$;

  Compute $\tilde{\mathbf{\Pi}}$, $\tilde{\mathbf{\Theta}}$, and L^t;

end

Algorithm 1: MELODIC Algorithm

3.3 Algorithm for constrained model

Sometimes researchers have an idea in advance of which responses belong together in which dimensions. Let us denote the set of response variables that

pertains to dimension m by $\mathcal{D}_m$. Furthermore, let us denote the set of dimensions to which response variable r pertains as $\mathcal{S}_r$. Then, for $m \notin \mathcal{S}_r$, we restrict $l_{rm} = k_{rm} = 0$.

Much of the unconstrained MM algorithm can be used, except for two aspects: (i) de orthonormality restriction on $\mathbf{XB}$ needs to be relaxed and (2) the updates will be done dimension wise. We still need a scale restriction per dimension on $\mathbf{XB}$. The equivalent of (9) can be written

$$g_c(\mathbf{B}, \mathbf{K}, \mathbf{L}) = \left\| \frac{1}{2}\mathbf{ZJA}_k + \mathbf{1a}^\top - \sum_{m=1}^{M} \mathbf{Xb}_m\mathbf{k}_m^\top \right\|^2,$$

which shows that the last term can be decomposed in dimensional terms. To update for dimension s, we first define

$$\tilde{\mathbf{Z}}_3 = \frac{1}{2}\mathbf{ZJA}_k + \mathbf{1a}^\top - \sum_{m \neq s} \mathbf{Xb}_m\mathbf{k}_m^\top.$$

Next, we define the matrix $\tilde{\mathbf{Z}}_s$ to be the subset of the matrix $\tilde{\mathbf{Z}}_3$ consisting of the columns for which $r \in \mathcal{D}_s$.

Similar to the unconstrained model, a generalized singular value decomposition of $\tilde{\mathbf{Z}}_s$ gives updates for $\mathbf{b}_s$ and $\mathbf{k}_s$ by taking the highest singular value and the corresponding vector. The update for $\mathbf{a}$ is the same as in the unconstrained model. The update for $\mathbf{L}$ is similar to the unconstrained model, but we only give non-zero values to the dimensions to which response variable r pertains. A summary of the algorithm is given in Algorithm 2.

The total number of parameters for the constrained model depends on the specific constraints. Nevertheless, we have

- $(P-1)M$ parameters for the regression weights $\mathbf{B}$;
- We define an indicator matrix of size $R \times M$ indicating which response

belongs to which dimension. The number of parameters in the matrix $\mathbf{K}$ equals the number of ones in that indicator matrix (see the next section for an example);

- R parameters in $\mathbf{L}$.

Data: $\mathbf{X}$, $\mathbf{G}$, M

Result: $\mathbf{B}$, $\mathbf{K}$, $\mathbf{L}$

Compute: $\mathbf{R}_x^{-1}\mathbf{X}^\top\mathbf{G} = \mathbf{P}\mathbf{\Phi}\mathbf{Q}^\top$;

Initialize: $\mathbf{B}^{(0)} = \sqrt{n}\mathbf{R}_x^{-1}\mathbf{P}_M$;

Compute: $\mathbf{V} = \frac{1}{\sqrt{n}}\mathbf{Q}_M\mathbf{\Phi}_M$;

Initialize: $\mathbf{K}^{(0)}$ by taking the uneven rows of $\mathbf{JV}$;

Initialize: $\mathbf{L}^{(0)}$ by taking the uneven rows of $(\mathbf{I}-\mathbf{J})\mathbf{V}$;

Set elements in $\mathbf{K}^{(0)}$ and $\mathbf{L}^{(0)}$ to zero, following the constraints;

Compute: $\tilde{\mathbf{\Pi}}$ and $\tilde{\mathbf{\Theta}}$;

while $t = 0$ *or* $(L^t - L^{(t-1)})/L^t > 10^{-8}$ **do**

  $t = t + 1$;

  Compute: $\mathbf{Z} = \tilde{\mathbf{\Theta}} + 2(\mathbf{G} - \tilde{\mathbf{\Pi}})$;

  Compute: $\tilde{\mathbf{Z}}_1 = \frac{1}{2}\mathbf{ZJA}_k$;

  Compute: $\mathbf{a}^+ = -\tilde{\mathbf{Z}}_1^\top\mathbf{1}/n$;

  for $s = 1, \ldots, M$ **do**

    Compute: $\tilde{\mathbf{Z}}_s$;

    Compute: $\mathbf{R}_x^{-1}\mathbf{X}^\top\tilde{\mathbf{Z}}_s = \mathbf{P}\mathbf{\Phi}\mathbf{Q}^\top$;

    Update: $\mathbf{b}_s^{(t)} = \sqrt{n}\mathbf{R}_x^{-1}\mathbf{P}_1$;

    Update: $\mathbf{k}_s^{(t)} = \frac{1}{\sqrt{n}}\mathbf{Q}_1\mathbf{\Phi}_1$;

  end

  Update: $l_{rm}^{(t)} = a_r / \sum_{m=1}^{M} k_{rm}$, for $m \in \mathcal{S}_r$ and $\forall r$;

  Compute $\tilde{\mathbf{\Pi}}$, $\tilde{\mathbf{\Theta}}$, and L^t;

end

Algorithm 2: MELODIC Algorithm for Constrained Model

4 Two empirical applications

In this section we discuss two empirical applications of the model. The first data set considers profiles of drug consumption; the second, profiles of mental disorders. For the first data set we use the unconstrained model. For the second data set we start with a set of constrained models representing different theories.

4.1 Drug Consumption Data

The drug consumption data (Fehrman et al., 2017) has records for 1885 respondents. For each respondent, nine attributes are measured. We have personality measurements based on the big five personality traits, neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A), and conscientiousness (C), and two other personality characteristics, namely impulsivity (I) and sensation seeking (S). Data were also collected on age and gender.[2]

In addition, participants were questioned concerning their use of 18 legal and illegal drugs. For each drug, participants were asked whether they never used the drug, used it over a decade ago, in the last decade, in the last year, month, week, or day. In our analysis we coded whether participants used the particular drug in the last year (yes or no). Furthermore, in our analysis we focused on the drugs that had a minimum percentage of 10% and a maximum of 90%, which are Amphetamine, Benzodiazepine, Cannabis, Cocaine, Ecstasy, Ketamine, legal highs, LSD, Methadone, Mushrooms, and Nicotine ($R = 11$).

The first step in the analysis is to select the dimensionality. We fit models in one to seven dimensions and compute information criteria statistics for comparison. The results are given in Table 1, where we can see that either the two- or three-dimensional solution is optimal according to the AIC and BIC statistics.

[2] Also level of education, ethnicity, and country of origin are available in the original data base. We omitted these from the analysis.

Dimensionality	Deviance	#param	AIC	BIC
1	18311	30	18371	18538
2	18117	48	18213	**18479**
3	18030	65	**18160**	18520
4	17998	81	18160	18609
5	17987	96	18179	18711
6	17980	110	18200	18810
7	17975	123	18221	18903

Table 1: AIC and BIC statistics for models in 1 to 7 dimensions for the drug consumption data.

Left Out	Deviance	#param	AIC	BIC
Age	19303	46	19395	19650
Gender	18417	46	18509	18764
Neuroticism	18181	46	18273	18528
Extraversion	18134	46	18226	18481
Openess	18449	46	18541	18796
Agreeablenes	18137	46	18229	18484
Conscientiousness	18172	46	18264	18519
Impulsivity	18121	46	18213	**18468**
Sensation seeking	18409	46	18501	18756

Table 2: AIC and BIC statistics for two-dimensional models with 1 predictor left out of the model.

We should also check the influence of the predictor variables. Each of the predictor variables is left out of the two-dimensional model. The AIC and BIC statistics are shown in Table 2, where it can be seen that only impulsivity can be considered for being left out of the model; all other predictors would lead to a substantial loss in fit. We decided to further interpret the model using all predictor variables except impulsivity.

The graphical representation of the two-dimensional model is shown in Figure 3.[3] The first thing that catches the eye in Figure 3 is that the categories for "yes" (labels ending with 1) are all to the right-hand side of the categories for

[3]We left out the decision lines to avoid clutter.

"no" (labels ending with 0). Therefore, participants who use drugs lie on the positive side of the first dimension. It can be seen that cannabis is furthest to the left, while Ketamine, LSD and Methadone are on the extreme right-hand side. Apparently, if participants start using drugs, they start with cannabis and only add other drugs later.

Considering the predictor side, we see that five predictor variables run from left to right: openness to experience (O), sensation seeking (S), extraversion (E), age, and gender. With respect to gender and age, boys use more drugs than girls and younger people use more drugs than the elderly. Furthermore, participants who are more open to experience (O) and less extravert (E) use more drugs. Finally, participants scoring high on sensation seeking (S) use more drugs than participants who score low on sensation seeking.

The vertical dimension is harder to interpret because the differences between the yes and no points are often small. The largest differences are for benzodiazepine and methadone (yes category has a higher coordinate) and for LSD (yes category has a lower coordinate). The predictor variable neuroticism (N) points strongly in this direction, indicating that neurotic participants tend to use benzodiazepine and methadone more frequently but LSD less frequently. The variable axis for agreeableness (A) is almost parallel to the vertical dimension, but in the opposite direction to neuroticism, indicating opposite effects.

To get a more detailed interpretation, let us look at the estimated implied logistic regression coefficients (Equation 5) in Table 3). Since the predictor variables are standardized to have zero mean and standard deviation one, these are changes in log odds for one standard deviation increases in the predictors. The numbers in each column can be interpreted as the standardized coefficients in a single logistic regression model.

We can verify how well each response variable is represented in the low-

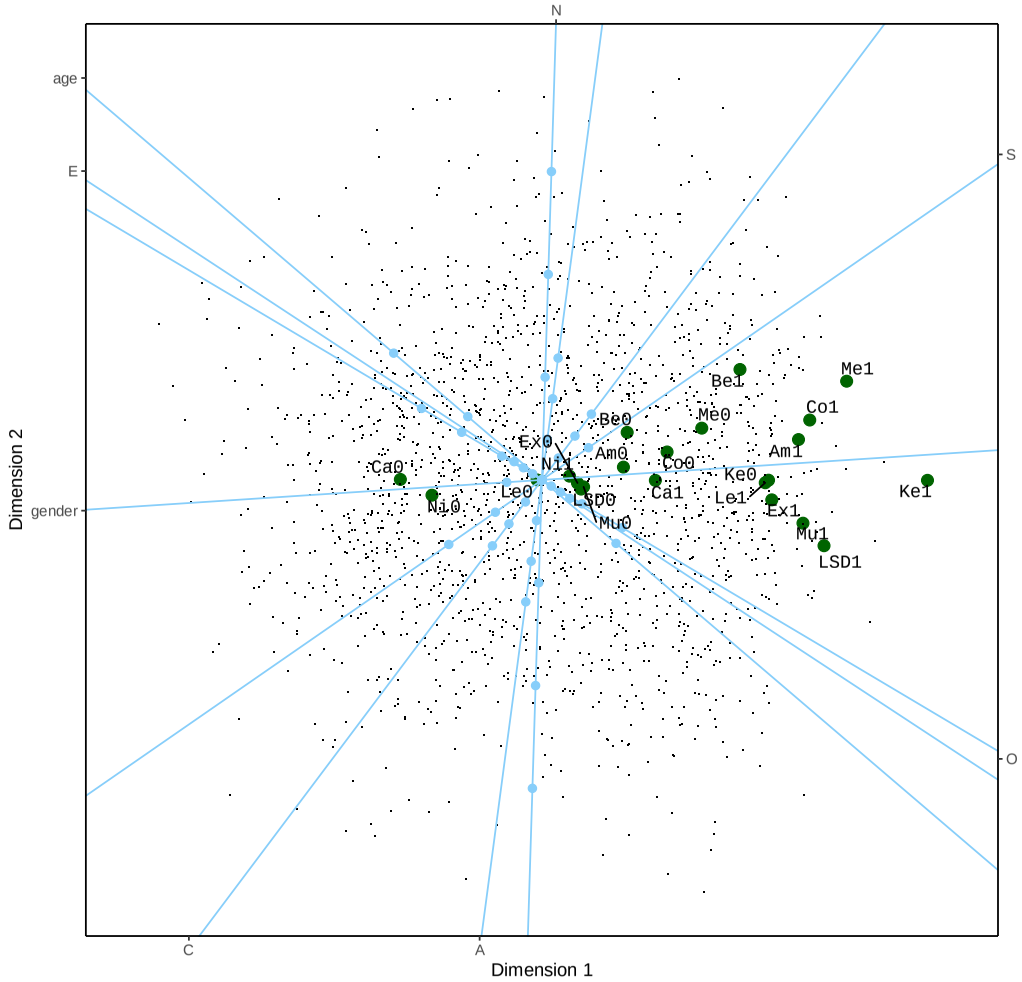

Figure 3: Two-dimensional solution for the drug consumption data. Predictor variable labels are printed on the border of the graph where N = Neuroticism; E = Extraversion; O = Openness to experience; A = Agreeableness; C = Conscientiousness; S = Sensation seeking. We use the convention that labels are printed at the positive side of the variable. Markers on the variable axes indicate standard deviation increases/decreases from the mean. Category points are labeled by the name of the drug together with a 1 (yes) or 0 (no), that is, `Am1` denotes the class point for use of Amphetamine and `Am0` for no use of Amphetamine. Am = Amphetamine; Be = Benzodiazepine; Ca = Cannabis; Co = Cocaine; Ex = Ecstasy; Ke = Ketamine; Le = legal highs; LSD = LSD; Me = Methadone; Mu = Mushrooms; Ni = Nicotine.

Predictor variables	Response Variables										
	Am	Be	Ca	Co	Ex	Ke	Le	LSD	Me	Mu	Ni
age	-0.59	-0.23	-0.99	-0.45	-0.81	-0.62	-0.90	-1.13	-0.41	-0.97	-0.47
gender	-0.33	-0.22	-0.47	-0.27	-0.36	-0.29	-0.42	-0.44	-0.27	-0.40	-0.26
Neuroticism	0.18	0.36	0.04	0.20	-0.06	0.03	0.02	-0.26	0.28	-0.16	0.13
Extraversion	-0.08	-0.03	-0.12	-0.06	-0.10	-0.08	-0.11	-0.14	-0.06	-0.12	-0.06
Openess	0.33	0.16	0.54	0.26	0.43	0.33	0.48	0.58	0.25	0.51	0.27
Agreeableness	-0.11	-0.17	-0.07	-0.11	-0.02	-0.04	-0.06	0.05	-0.14	0.02	-0.08
Conscientiousness	-0.18	-0.17	-0.22	-0.16	-0.15	-0.14	-0.19	-0.14	-0.18	-0.15	-0.14
Sensation seeking	0.47	0.38	0.62	0.40	0.45	0.39	0.55	0.50	0.43	0.47	0.37
Quality	1.00	0.96	0.97	0.90	0.96	0.95	1.00	0.98	0.94	0.99	0.98

Table 3: Estimated implied regression coefficients (equation 5) of each of the predictor variables for each of the response variables. In the columns Am = Amphetamine; Be = Benzodiazepine; Ca =Cannabis; Co = Cocaine; Ex = Ecstasy; Ke = Ketamine; LE = legal highs; LSD = LSD; Me = Methadone; Mu = Mushrooms; Ni = Nicotine. Last line shows the quality of representation (Q_r).

dimensional space. To do this, we define a measure called Quality of Representation, Q_r, which is defined by

$$Q_r = (L_{(0,r)} - L_r)/(L_{(0,r)} - L_{lr}),$$

where $L_{(0,r)}$ is the deviance of the intercept-only logistic regression model for response variable r, L_r is the part of our loss function for response variable r, and L_{lr} is the deviance from a logistic regression with the same predictor variables. Thus, $Q_r$can be interpreted as the proportion of loss in deviance imposed by the Melodic model compared to an unconstrained logistic regression for response variable r. The quality of representation for the response variables in this analysis are given in the last row of Table 3, where it can be seen that most response variables are well represented. The response variable "cocaine" (Co) is worst represented, although still with 89.8% recovered.

4.2 Depression and Anxiety data

Depression and anxiety disorders are common at all ages. Approximately one out of three people in the Netherlands will be faced with them at some time during their lives. It is not clear why some people recover quickly and why others suffer for long periods of time. The Netherlands Study of Depression and Anxiety (NESDA) was therefore designed to investigate the course of depression and anxiety disorders over a period of several years. For more information about the study design, see Penninx et al. (2008). In our application, we will analyze data from the first wave, focusing on the relationship between personality and depression and anxiety disorders. The data were previously analyzed by Spinhoven et al. (2009). Data were collected from three different populations: from primary health care; from generalized mental health care; and from the general population. Our analysis will focus on the population of generalized health care.

We have data for 786 participants. The diagnoses Dysthymia (D), Major Depressive Disorder (MDD), Generalized Anxiety Disorder (GAD), Social Phobia (SP), and Panic Disorder (PD) were established with the Composite Interview Diagnostic Instrument (CIDI) psychiatric interview. Personality was operationalized using the 60-item NEO Five-Factor Inventory (NEO-FFI). The NEO-FFI questionnaire measures the following five personality domains: Neuroticism, Extraversion, Agreeableness, Conscientiousness and Openness to Experience. In addition to these five predictors, three background variables were measured: age, gender, and education in years.

The prevalences in the data are 21.25% for dysthymia, 76.21% for major depressive disorder, 30.41% for generalized anxiety disorder, 41.6% for social phobia, and 52.8% for panic disorder. Of the 786 participants, 272 have a single disorder, the others all have multiple disorders. There are 235 participants with

two disorders, 147 with three, 96 with four, and 36 participants with five disorders.

Due to the high comorbidity among disorders, the scientific field of psychiatry developed three different theories:

1. a unidimensional structure where all the disorders are represented by a single dimension;

2. a two-dimensional structure with one dimension representing distress (D, MDD, GAD) and the other fear (SP, PD);

3. a two-dimensional structure with one dimension representing depression (D, MDD) and the other anxiety (GAD, SP, PD).

We can of course define another two-dimensional structure (Theory 4) in which dysthymia and major depressive disorder pertain to the first dimension, social phobia and panic disorder to the second, and generalized anxiety disorder to both dimensions.

Each of the three two-dimensional theories gives rise to a different response variable by dimension indicator matrix:

$$\mathbf{D}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{D}_3 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{D}_4 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

We fitted the four models reflecting the four theories to the data. The fit statistics can be found in Table 4, where it can be seen that all four theories give about the same fit, but the distress-fear hypothesis (corresponding to $\mathbf{D}_2$) has

a slightly lower AIC and the unidimensional model a lower BIC value than the other theories.

Theory/Model	Deviance	#param	AIC	BIC
1	4553.34	17	4587	4667
2	4531.17	24	4579	4691
3	4533.71	24	4582	4694
4	4530.35	25	4580	4697

Table 4: AIC and BIC statistics for the models reflecting the four theories.

The graphical display for the distress-fear model (Theory 2) is given in Figure 4. We can see that three response variables pertain to the horizontal dimension, while two pertain to the vertical dimension. The class points for dysthymia, major depressive disorder, and generalized anxiety disorder fall on the horizontal dimension (and thus have vertical decision lines), while the class points for social phobia and panic disorder fall on the vertical dimension (and therefore have horizontal decision lines). The decision lines partition the two-dimensional space into rectangular regions in which a certain response profile is most probable. We further see that, on the horizontal dimension, dysthymia is best discriminated as the two points lie farthest apart (distance 0.80), while the distance for major depressive disorder is 0.64 and for generalized anxiety disorder 0.56. On the vertical axis, social phobia is well discriminated (distance 0.76) while panic disorder is hardly discriminated (distance 0.16). The latter means that, using the three background variables and the five personality variables together with the imposed model structure, we have hardly any information to distinguish participants with and without panic disorder. We will come back to this issue later.

The implied logistic regression coefficients are given in Table 5. As in our previous analysis we standardized the predictor variables such that the coefficients give changes in log odds for one standard deviation changes in the predic-

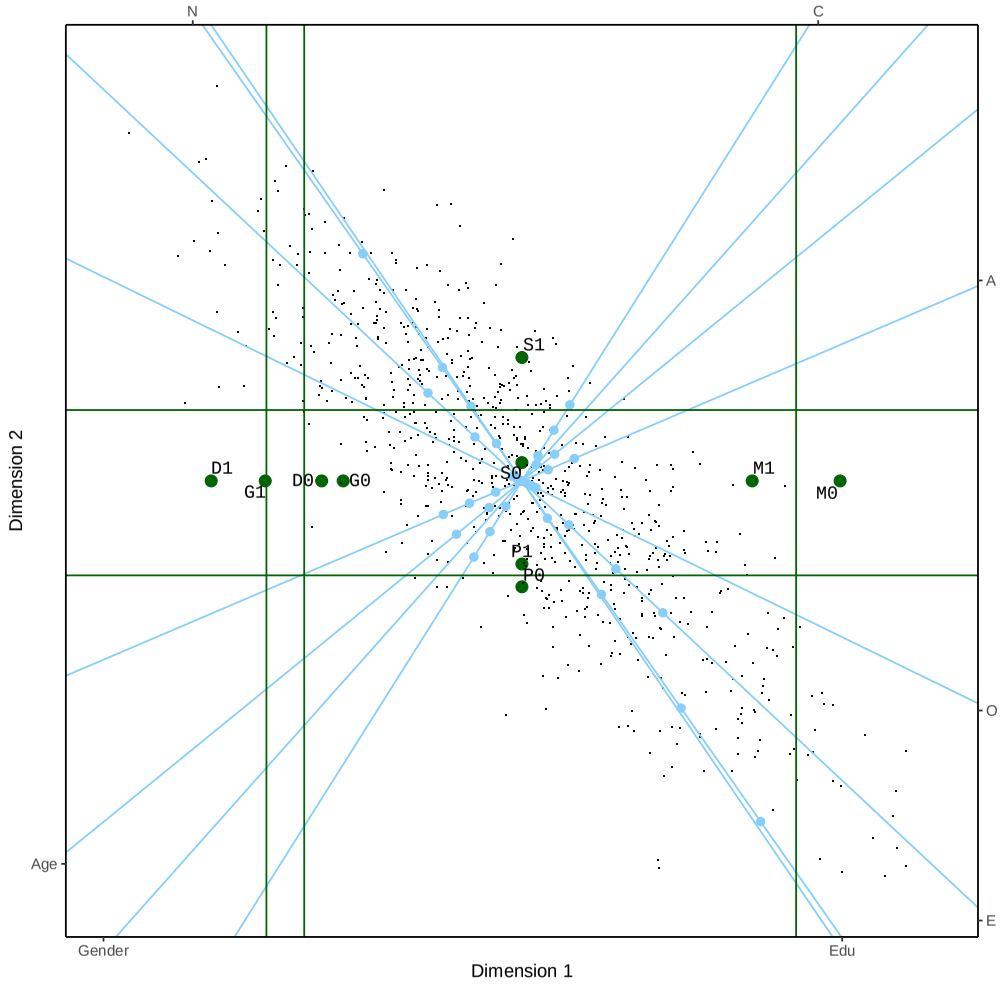

Figure 4: Two-dimensional solution following Theory 2 for the depression and anxiety data. Predictor variable labels are printed on the border of the graph where N = Neuroticism; E = Extraversion; O = Openness to experience; A = Agreeableness; C = Conscientiousness and Edu = Education. We use the convention that labels are printed at the positive side of the variable. Markers on the variable axes indicate standard deviation increases/decreases from the mean. Category points are labeled by the name of the drug together with a 1 (yes) or 0 (no) with D = Dysthymia, M = Major Depressive Disorder, G = Generalized Anxiety Disorder, S = Social Phobia (SP), and P = Panic disorder (PD).

	Response Variable				
Predictor Variable	D	M	G	S	P
Gender	0.08	0.07	0.06	-0.09	-0.02
Age	0.19	0.15	0.13	-0.15	-0.03
Education	-0.15	-0.12	-0.10	-0.21	-0.04
Neuroticism	0.46	0.37	0.32	0.62	0.14
Extraversion	-0.27	-0.22	-0.19	-0.24	-0.05
Openness	-0.03	-0.02	-0.02	-0.01	-0.00
Agreeableness	-0.15	-0.12	-0.11	0.06	0.01
Conscientiousness	-0.09	-0.07	-0.07	0.14	0.03
Quality	0.98	0.90	0.85	0.99	0.17

Table 5: Implied logistic regression coefficients for the Depression and Anxiety data. D = Dysthymia, M = Major Depressive Disorder, G = Generalized Anxiety Disorder, S = Social Phobia (SP), and P = Panic Disorder (PD). The last row represents the quality of representation for the five response variables.

tors. The most important predictor for mental disorders is neuroticism, which concurs with the conclusion in Spinhoven et al. (2009)

The quality of representation for the five response variables is given in the last row of Table 5, where we see that the response variable panic disorder is poorly represented in this model. This could already be inferred from the graphical representation (the two points almost coincide) and the implied coefficients table, where most coefficients for the response variable panic disorder are very small. Apparently, when we use a standard logistic regression model for this response variable, the predictor variables discriminate the two categories much better.

Because one response variable is poorly represented in the best fitting model, we also fitted an unconstrained model in two dimensions. Such a model has 28 parameters; the value of the loss function (deviance) is 4521.81 (AIC = 4578; BIC = 4708). The AIC indicates a better fit than the previous constrained models. The quality of representation of the response variables in this model is 0.97, 0.88, 0.85, 0.86, and 0.90, no longer indicating any poorly fitting response variables

anymore.

The biplot for this two-dimensional unconstrained model is given in Figure 5, where we can see that the decision lines for major depressive disorder, dysthymia, generalized anxiety disorder, and social phobia run more or less parallel to the vertical dimension and the decision line for panic disorder is more or less horizontal. This exploratory finding suggests a new theory: that major depressive disorder, dysthymia, generalized anxiety disorder, and social phobia pertain to a single underlying dimension, but that panic disorder behaves differently.

5 Related and competing approaches

The MELODIC family is a statistical toolbox for simultaneous logistic regressions in a reduced dimensional space for the analysis of multivariate binary data. Other statistical models have been proposed for such data; in this section we show some relationships and comparisons. The related approaches can be divided in two types of models: marginal models and conditional models.

Marginal models are like standard regression models, dealing in some way with the dependency among responses. In generalized estimating equations (GEE; Liang and Zeger, 1986; Zeger and Liang, 1986), a working correlation structure is adopted and estimation and inference are adjusted based on this structure using a sandwich estimator (White, 1980). GEE has been mainly developed in the longitudinal context but can also be applied to multivariate responses. Maximum likelihood estimation is also possible for marginal models -see for example Bergsma et al. (2009)- but this is computationally more demanding.

Our MELODIC family can be seen as a member of the GEE family, where implicitly we adopt an independent working correlation structure. Moreover,

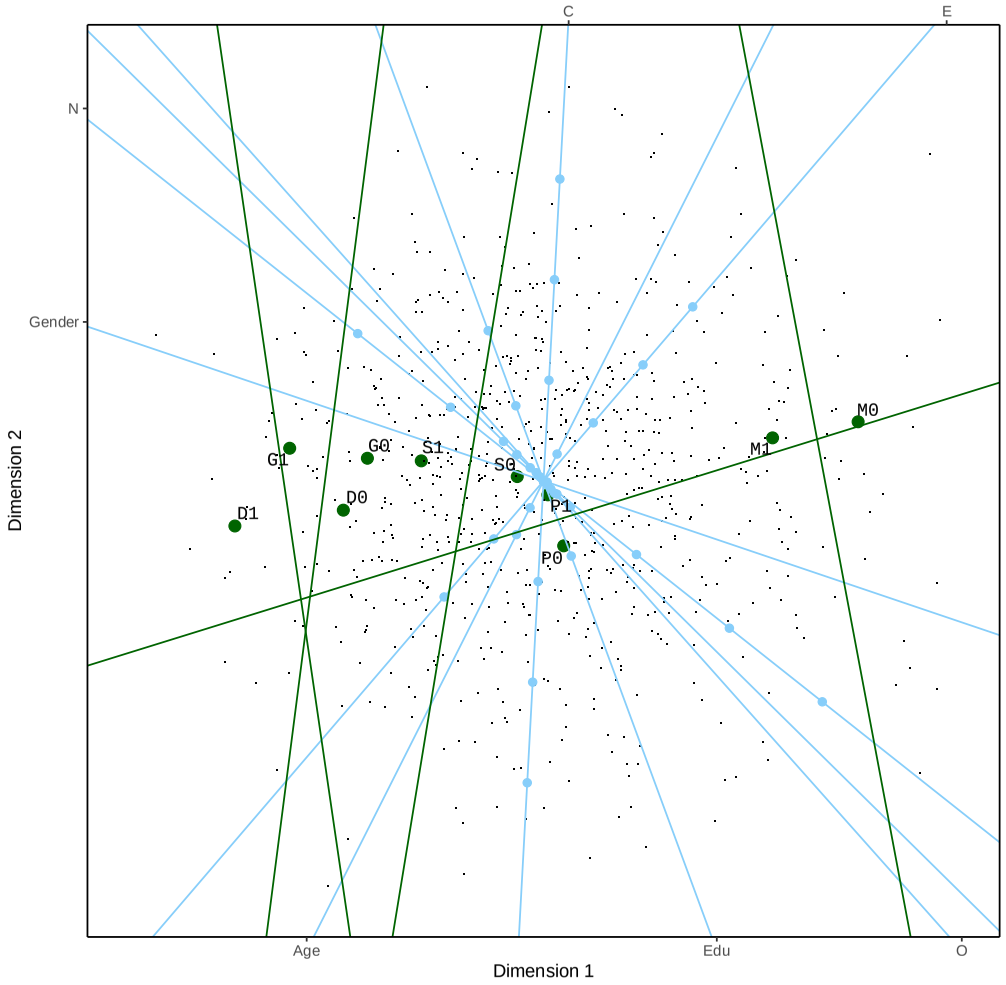

Figure 5: Graphical representation of the two dimensional unconstrained solution for the depression and anxiety data. Predictor variable labels are printed on the border of the graph where N = Neuroticism; E = Extraversion; O = Openness to experience; A = Agreeableness; C = Conscientiousness and Edu = Education. We use the convention that labels are printed at the positive side of the variable. Markers on the variable axes indicate standard deviation increases/decreases from the mean. Category points are labeled by the name of the drug together with a 1 (yes) or 0 (no) with D = Dysthymia, M = Major Depressive Disorder, G = Generalized Anxiety Disorder, S = Social Phobia (SP), and P = Panic Disorder (PD).

we lay a dimensional structure on the response space. A standard GEE model would be equal to our one-dimensional model, with the constraint that all responses are equally well discriminated (see Worku and De Rooij, 2018). Ziegler et al. (1998) discuss a set-up where the predictor variables have different effects on each of the response variables. This set-up would correspond to our model in maximum dimensionality (R), where each response pertains to a single dimension. Asar and İlk (2014) propose a method in which selected predictor variables have the same effect on selected response variables. This is similar to our constrained model, where predictors have a similar effect on response variables pertaining to a given dimension.

Reduced rank vector generalized linear models (RR-VGLMs; Yee and Hastie, 2003) is a family of models closely related to our own. Whereas we started our development from a distance perspective, these RR-VGLMs start from an inner-product perspective. De Rooij and Heiser (2005) give an extensive discussion of the interpretational differences of the two perspectives. Moreover, they show conditions under which the inner product rule and the distance rule are equivalent. If we had defined our model as

$$\pi_{rc}(\mathbf{x}_i) = \frac{\alpha_{rc} \exp(-\delta(\mathbf{u}_i, \mathbf{v}_{rc}))}{\alpha_{r0} \exp(-\delta(\mathbf{u}_i, \mathbf{v}_{r0})) + \alpha_{r1} \exp(-\delta(\mathbf{u}_i, \mathbf{v}_{r1}))},$$

then it would be equivalent to a RR-VGLM for binary data. However, the α-parameters would problematize the interpretation of the graphical representation, as the distance rule outlined in Section 2.4 would no longer be valid and the decision lines would not be equidistant from the two class points.

In *conditional models*, latent variables are included in order to model the dependency among the responses. The main examples are generalized linear mixed models and latent class models. The family of GLMMs includes item response models and factor analysis models (Skrondal and Rabe-Hesketh, 2004). When

these are expanded with predictor variables, explanatory item response models (De Boeck and Wilson, 2004) and structural equation models are obtained. Explanatory item response models have mainly been developed for unidimensional latent variables. Some progress has been made with multidimensional item response models, but the underlying structure should be known a priori (as in our constrained models). Explanatory multidimensional item response models still need further development, partly because estimation is often quite troublesome in such models due to the intractable integral in the likelihood function (Tuerlinckx et al., 2006).

Latent class models (Lazarsfeld and Henry, 1968; McCutcheon, 1987) have been developed for multivariate binary data including predictors (Vermunt, 2010). In latent class models, as in GLMMs, the dependency among the response variables is modeled using a latent variable, which in this case is categorical. No dimensional structure is imposed underlying the outcomes, only a choice of the number of categories of the categorical latent variable. These latent class models often require a large sample size in order to obtain stable and reliable results (Gudicha et al., 2016).

Hubbard et al. (2010), when comparing generalized estimating equations and generalized linear mixed models, noted that "mixed models involve unverifiable assumptions on the data-generating distribution, which lead to potentially misleading estimates and biased inference". More specifically, although the distribution of the random effects cannot be identified from the data, the estimates and inference change according to different choices of the distribution of these random effects. This makes the application of conditional models problematic. Another issue for conditional models is the number of indicator or response variables. In our example on depressive and anxiety disorders, for example, there are only five response variables. Distributing these five response

variables over two underlying dimensions results in a dimension with only two dichotomous indicators. It is generally acknowledged that this number is much too low for valid inference. This small number is less of an issue in our family, because we do not assume a particular distribution for the underlying dimension.

A final practical problem in these conditional models is that researchers often first try to find the dimensional structure and in a second step include the predictor variables. The measurement model (step 1), however, might change substantially when the predictor variables are included, leading to a completely different interpretation. To solve this problem, researchers have developed three-step (Bolck et al., 2004, sometimes called the BCH approach) and two-step estimators (Bakk and Kuha, 2018) within the context of latent class models which were recently adapted for other conditional models. In our family of models we have no division in measurement and structural model; the two go hand in hand.

6 Conclusion and Discussion

In this study, we presented distance models for simultaneous logistic regression analysis of multiple binary response variables based on ideas of multidimensional unfolding. Row objects (participants in our examples) are presented together with the two categories of the response variables in a low-dimensional Euclidean space, where the relative distance between a point representing a participant and the points representing the classes of a response variable determines the probability for each class. The model is estimated by minimizing a deviance function. These models take into account the dependency among the response variables by using a low-dimensional Euclidean space: with the increase in value of a predictor variable, the probabilities of all response vari-

ables change simultaneously. We christened the models the MELODIC family, that is the MultivariatE LOgistic DIstance to Categories family. We presented versions of the model both for cases in which we have an *a priori* theory about the dimensional structure of the response variables and for cases when we do not have such a theory. Two empirical applications are shown, one with and one without such an *a priori* structure.

In the case of a two-dimensional model, the result can be interpreted using a biplot. In the case of a higher-dimensional solution similar biplots can be constructed for pairs of dimensions. A coherent interpretation of the complete model from such bi-dimensional plots, might, however, more difficult. Alternatively, the model can be interpreted by the implied logistic regression coefficients which have a change in log odds interpretation similar to ordinary logistic regression models. We illustrated both methods of interpretation in the empirical examples. The fact that the model can be interpreted using a graph and using tables is beneficial, because applicants of statistical models can be divided into two groups: those who prefer visualizations and those who prefer numbers. With the MELODIC family, an applicant can choose which mode of interpretation is most suitable.

We developed a fast iterative majorization algorithm to estimate the parameters of the model. The algorithm converges monotonically to the global optimum of the deviance function. The algorithm alternates between 1) updating an auxiliary vector $\mathbf{a}$, which is simply obtained by taking an average, and 2) updating the regression weights ($\mathbf{B}$) and item discriminations ($\mathbf{K}$), which can be obtained from a generalized singular value decomposition. All model parameters can be obtained from these updates.

A measure of quality of representation for each response variable was also proposed, ranging between 0 and 1. A higher value implies only a small loss

of fit with regard to a univariate logistic regression with the specific response variable and the same set of predictor variables. This measure can be used as a diagnostic tool to assess whether response variables are well represented by the model. In our second application, we saw that for one response variable the quality of representation was low. Further exploratory analysis suggested a different substantial theory.

In applications, we need to select the predictor variables as well as the dimensionality of the model. We used information criteria such as the AIC and BIC in the empirical applications. Alternatively, cross validation or other model selection criteria can be used. We did not discuss uncertainty estimation for our model. Assuming the model is true, we can compute the Hessian matrix and derive standard errors for the parameter estimates from this matrix. Following the GEE set-up, we could develop a sandwich estimator for the covariance matrix of the parameters. Alternatively, the bootstrap can be used (Efron and Tibshirani, 1986). The two latter approaches acknowledge the fact that the model is an approximation (Buja et al., 2019a,b) and probably not a completely accurate representation of a population model. In that sense, the sandwich estimator and the bootstrap can estimate uncertainty with respect to a target model in the population. Focusing on predictive accuracy instead of explanatory value (cf. Shmueli, 2010), the performance of our model in comparison to that of independently fitted logistic regression models is expected to be higher (Breiman and Friedman, 1997).

In this manuscript we only focused on linear effects of the predictor variables. Non-linear effects of predictor variables on the responses can easily be incorporated as long as they can be translated into a design matrix $\mathbf{X}$, such as with the use of quadratic and cubic effects or with splines defined in terms of a truncated power basis (Friedman et al., 2001). Non-linear variable axes can be

presented in the graphical representation as smooth curves, where effects are still additive. Interactions can also be included in the model. In the graphical representation, *conditional variable axes* need to be represented. In the case of an interaction between variables X_1 and X_2, the graphical representation has a variable axis for X_1 for *each* value of X_2 (or the other way around). For an example of biplots with such interactions among predictors, see De Rooij (2011). Note that both the nonlinearity and the interactions are effects with respect to all response variables.

The past two decades have seen a rise in penalized estimation methods, which are methods that impose penalties on the parameters of the model. For the MELODIC family, these could be penalties on the regression weights to generalize the model to the case where $P >> n$, such as L_1 (Lasso penalty; Tibshirani (1996)) or L_2 penalties (Ridge penalty; Hoerl and Kennard (1970)) . To implement these in the MELODIC family, we would need to alter the identification constraints. In the current algorithm we used $n^{-1}\mathbf{B}^\top\mathbf{X}^\top\mathbf{X}\mathbf{B} = \mathbf{I}$ to identify the model, but this scaling does not seem to be in line with a penalty on $\mathbf{B}$. Therefore, the fixed scaling normalization should be placed on the discrimination values ($\mathbf{K}$). Another potential type of penalty is a nuclear norm penalty (Fazel, 2002). In the outlined algorithm (see Algorithm 1) we use a singular value decomposition. If we apply an L_1 penalty to the singular values, the discrimination ($\mathbf{K}$) between the categories of response variables slowly diminishes. Because the singular values are ordered, the discrimination in the higher dimensions becomes zero first and later the discrimination of the first dimensions also becomes zero. If we choose the optimal value of the penalty parameter by cross validation, such a penalty could be used for dimension selection

We are currently building an R-package that enables empirical researchers to

apply the models to their own data. For the moment, the R-code of the examples presented can be obtained from the first author.

References

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(1):1–22.

Asar, Ö. and İlk, Ö. (2014). Flexible multivariate marginal models for analyzing multivariate longitudinal data, with applications in r. *Computer methods and programs in biomedicine*, 115(3):135–146.

Bakk, Z. and Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, 83(4):871–892.

Bergsma, W., Croon, M., and Hagenaars, J. (2009). *Marginal Models: For Dependent, Clustered, and Longitudinal Categorical Data*. Statistics for Social and Behavioral Sciences. Springer New York.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365.

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1):3–27.

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., Zhao, L., et al. (2019a). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., Zhao, L., et al. (2019b). Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565.

Busing, F. M. T. A. (2010). *Advances in multidimensional unfolding*. Doctoral thesis, Leiden University.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological review*, 57(3):145.

Coombs, C. H. and Kao, R. (1955). Nonmetric factor analysis. *University of Michigan. Department of Engineering Research. Bulletin*.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Discussion Paper*, 02-119(4).

De Boeck, P. and Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.

De Leeuw, J. (2005). Gifi goes logistic: Scasa keynote.

De Leeuw, J. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational statistics & data analysis*, 50(1):21–39.

De Leeuw, J. and Heiser, W. J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. *Geometric representations of relational data*, pages 735–752.

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

De Rooij, M. (2011). Transitional ideal point models for longitudinal multinomial outcomes. *Statistical Modelling*, 11(2):115–135.

De Rooij, M. and Heiser, W. J. (2005). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *psychometrika*, 70(1):99–122.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

Evans, G. W. (2014). *Logistic Gifi: A Logistic Distance Association Model for Exploratory Analysis of Categorical Data*. PhD thesis, UCLA.

Fazel, M. (2002). *Matrix rank minimization with applications*. Doctoral thesis, Stanford University.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N. (2017). The five factor model of personality and evaluation of drug consumption risk. In *Data Science*, pages 231–242. Springer.

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

Friendly, M. and Kwan, E. (2011). Comment-why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, 20(1):18.

Gelman, A. (2011). Why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, 20(1):3–7.

Gower, J. and Hand, D. (1995). *Biplots*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Gower, J., Lubbe, S., and Roux, N. (2011). *Understanding Biplots*. Wiley.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Groenen, P. J. F. (1993). *The majorization approach to multidimensional scaling*. DSWO press Leiden.

Groenen, P. J. F., Giaquinto, P., and Kiers, H. A. L. (2003). Weighted majorization algorithms for weighted least squares decomposition models. Econometric Institute Research Papers EI 2003-09, Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute.

Groenen, P. J. F. and Josse, J. (2016). Multinomial multiple correspondence analysis. *arXiv preprint arXiv:1603.03174*.

Gudicha, D. W., Tekle, F. B., and Vermunt, J. K. (2016). Power and sample size computation for wald tests in latent class models. *Journal of Classification*, 33(1):30–51.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33(4):469–506.

Heiser, W. J. (1981). *Unfolding analysis of proximity data*. Doctoral dissertation, Leiden University.

Heiser, W. J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In Krzanowski, W. J., editor, *Recent advances in descriptive multivariate analysis*, pages 157–189. Clarendon Press.

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

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, 1(1):27–35.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Satariano, S. A., Jewell, N., Bruckner, T., and Satariano, W. A. (2010). To gee or not to gee: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, pages 467–474.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin Co.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

Molenberghs, G. and Verbeke, G. (2006). *Models for discrete longitudinal data*. Springer Science & Business Media.

Pearson, K. (1901). Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559.

Penninx, B. W., Beekman, A. T., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., Cuijpers, P., De Jong, P. J., Van Marwijk, H. W., Assendelft, W. J., et al. (2008). The Netherlands study of depression and anxiety (NESDA): rationale, objectives and methods. *International journal of methods in psychiatric research*, 17(3):121–140.

Roskam, E. E. (1968). *Metric Analysis Or Ordinal Data in Psychology*. Voorschoten, The Netherlands: Vam.

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

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

Spinhoven, P., De Rooij, M., Heiser, W., Smit, J. H., and 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.

Stein, C. et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika*, 52(4):493–513.

Takane, Y. (2013). *Constrained principal component analysis and related techniques*. CRC Press.

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

Ter Braak, C. J. and Looman, C. W. (1994). Biplots in reduced-rank regression. *Biometrical journal*, 36(8):983–1003.

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

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

Torgerson, W. S. (1958). *Theory and methods of scaling.* Wiley.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, pages 450–469.

Vugteveen, J., De Bildt, A., Hartman, C., and Timmerman, M. (2018). Using the dutch multi-informant strengths and difficulties questionnaire (SDQ) to predict adolescent psychiatric diagnoses. *European child & adolescent psychiatry*, 27(10):1347–1359.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838.

Worku, H. M. and De Rooij, M. (2018). A multivariate logistic distance model for the analysis of multiple binary responses. *Journal of Classification*, 35(1):124–146.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical modelling*, 3(1):15–41.

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

Ziegler, A., Kastner, C., and Blettner, M. (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 40(2):115–139.