Multinomial restricted uinfolding

Rooij, M.J. de; Busing, F.M.T.A.

Journal of Classification (2024) 41:190–213
https://doi.org/10.1007/s00357-024-09465-3

ORIGINAL RESEARCH

Multinomial Restricted Unfolding

Mark de Rooij[1] · **Frank Busing**[1]

Accepted: 12 February 2024 / Published online: 8 April 2024

Abstract
For supervised classification we propose to use restricted multidimensional unfolding in a multinomial logistic framework. Where previous research proposed similar models based on squared distances, we propose to use usual (i.e., not squared) Euclidean distances. This change in functional form results in several interpretational advantages of the resulting biplot, a graphical representation of the classification model. First, the conditional probability of any class peaks at the location of the class in the Euclidean space. Second, the interpretation of the biplot is in terms of distances towards the class points, whereas in the squared distance model the interpretation is in terms of the distance towards the decision boundary. Third, the distance between two class points represents an upper bound for the estimated log-odds of choosing one of these classes over the other. For our multinomial restricted unfolding, we develop and test a Majorization Minimization algorithm that monotonically decreases the negative log-likelihood. With two empirical applications we point out the advantages of the distance model and show how to apply multinomial restricted unfolding in practice, including model selection.

Keywords Multiclass · Euclidean distance · Classification · Visualization

1 Introduction

Multiclass classification problems occur in many scientific disciplines. In multiclass classification the data consist of a set of predictor variables and one categorical response variable, that is $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^P$ and $y_i \in \{1, .., c, .., C\}$. Traditional examples of classification techniques for such data are linear discriminant analysis and multinomial logistic regression.

Multidimensional unfolding (Coombs, 1964; Heiser, 1981; Busing, 2010) is a technique closely related to Multidimensional Scaling (MDS). In multidimensional unfolding, we have a proximity matrix of size $n \times C$, that we would like to map in a low dimensional Euclidean space

✉ Mark de Rooij
rooijm@fsw.leidenuniv.nl

Frank Busing
busing@fsw.leidenuniv.nl

[1] Methodology and Statistics Department, Leiden University, Leiden, The Netherlands

in which the distances between points representing the row objects and points representing the column objects approximate the proximities as closely as possible, often in a least squares sense. In *restricted* multidimensional unfolding, a set of points, say the points for the row objects, is restricted to be a (linear) function of external information.

Takane and colleagues (Takane et al., 1987; Takane, 1987) combined the ideas of multinomial logistic regression and restricted multidimensional unfolding in what they called Ideal Point Discriminant Analysis (IPDA). In this case the proximity matrix is defined by the matrix $\mathbf{G}$, which is the indicator matrix corresponding to the response variable $\mathbf{y}$, where the elements can be interpreted as similarity values. Takane (1998) gave a detailed discussion on the interpretation of the IPDA biplot and concluded that this interpretation is complicated mainly due to a set of so-called bias parameters. De Rooij (2009) further investigated IPDA and showed that these bias parameters can be removed from the model without hampering model fit and thus facilitating interpretation. De resulting model without bias parameters was called the Ideal Point Classification (IPC) model. Both IPDA and IPC use squared distances.

In this paper, we define a model without bias parameters and based on Euclidean distances, not squared Euclidean distances. When researchers interpret a visualization they inspect distances, not squared distances. Therefore, the use of distances might have interpretational advantages. For both models, based on distances and based on squared distances, the probabilities are inversely related to the relative distances. However, the exact functional form changes and so will the representation of the probabilities and functions of these probabilities, like the (log) odds. Therefore, we investigate in detail the properties of the distance model in terms of probabilities and the log-odds representation.

This paper is organized as follows. In Section 2, we introduce the model, discuss its graphical representation through a biplot, and investigate relations between distances, classification regions, probabilities, and log-odds. We also briefly discuss model selection. In Section 3, we develop an MM algorithm to estimate the parameters of the model. Section 4 describes two simulation studies. In the first, we evaluate the performance of the algorithm. In the second, we evaluate the predictive performance of the distance model against that of the squared distance model and multinomial logistic regression. In Section 5, the model is applied to two empirical data sets. With the first empirical data set, we contrast the distance model from the squared distance model. With the second empirical data set, we illustrate how to use multinomial restricted unfolding in practice. We conclude this manuscript with some discussion.

2 Model and Interpretation

2.1 Model Definition

In our multinomial restricted unfolding model, there are two sets of points in an S dimensional space, where S has to be chosen by the researcher. Note that S will often be equal to two. The first set of points represents the participants (objects or persons); the second set represents the classes. The coordinates of the points representing the participants are defined by the S-vector $\mathbf{u}_i$ for $i = 1, \ldots, n$ and they are collected in the $n \times S$ matrix $\mathbf{U}$. The coordinates of the class points are given by the S-vector $\mathbf{v}_c$ for $c = 1, \ldots, C$ and collected in the $C \times S$ matrix $\mathbf{V}$.

The conditional probability $\pi_{ic}(\mathbf{x}_i) = P(y_i = c|\mathbf{x}_i)$ for participant i to be in a class c is a function of the between set Euclidean distances in S dimensions, that is

$$\pi_{ic}(\mathbf{x}_i) = \frac{\exp(\theta_{ic})}{\sum_{c'=1}^{C} \exp(\theta_{ic'})}$$

with

$$\theta_{ic} = -d(\mathbf{u}_i, \mathbf{v}_c) = -\sqrt{\sum_{s=1}^{S}(u_{is} - v_{cs})^2}.$$

The coordinates of the participant are constrained to be a linear combination of the predictor variables, that is $\mathbf{u}_i = \mathbf{B}^\top \mathbf{x}_i$, with $\mathbf{B}$ a matrix of regression coefficients. The conditional probabilities are inversely related to the relative distances, such that the closer a class point is to a person in Euclidean space, the higher the conditional probability.

Without loss of generality, we will center the predictor variables so that they have mean zero. With this centering, the origin of the Euclidean space corresponds to an object for which the profile of the predictor variables equals the mean values. Besides centering, the variables might also be scaled to have unit standard deviations.

2.2 Biplot

A graphical representation of our multinomial restricted unfolding represents the participants, the classes, and the predictor variables. The interest of the model lies in the relationship between the predictor variables $\mathbf{X}$ and the outcome variable $\mathbf{y}$. In a biplot, the predictor variables can be represented as variable axes, as in traditional biplots (Gower & Hand, 1996; Gower et al., 2011). In our biplots, we use the convention that the variable label is printed at the end with the highest value of the variable axis. The variable axis can be further enhanced with variable markers, which are a set of ticks on the axis that represent characteristic values of the observed variable. Furthermore, the variable axes will consist of a solid part and a dotted part. The solid part indicates the observed range of the predictor variable, while the dotted part extends the variable axis to the border of the display. The length of the solid part indicates the discriminatory power of the corresponding predictor.

In Fig. 1, we have an exemplary biplot with two predictor variables each shown by a solid variable axis. The regression weights for the variables are $\mathbf{b}_1 = (1.0, 0.0)^\top$ and $\mathbf{b}_2 = (0.5, 1.0)^\top$. The ticks on the axes represent units (when the variables are standardized these are standard deviations) from the average. The crossing of the two variable axes is at the average values of the two predictor variables because we centered the variables, as well as at the origin of the Euclidean space.

Participants are represented by points in the Euclidean space. Positions of the participants can be obtained from the variable axes by completing parallelograms, that is the process that is called *interpolation* by Gower and Hand (1996). We illustrate this process for a participant with 2 on the first predictor variable and -1 on the second. For each predictor, a vector is drawn from the origin towards these points, and the parallelogram is finished (i.e., the vectors are added) to obtain the position of the object with $\mathbf{x}_i = (2, -1)^\top$. This process is illustrated in Fig. 1 with the red arrows and the dotted lines. When there are more than two variables, we proceed by adding more vectors. In the biplot, we will represent all participants with small dots.

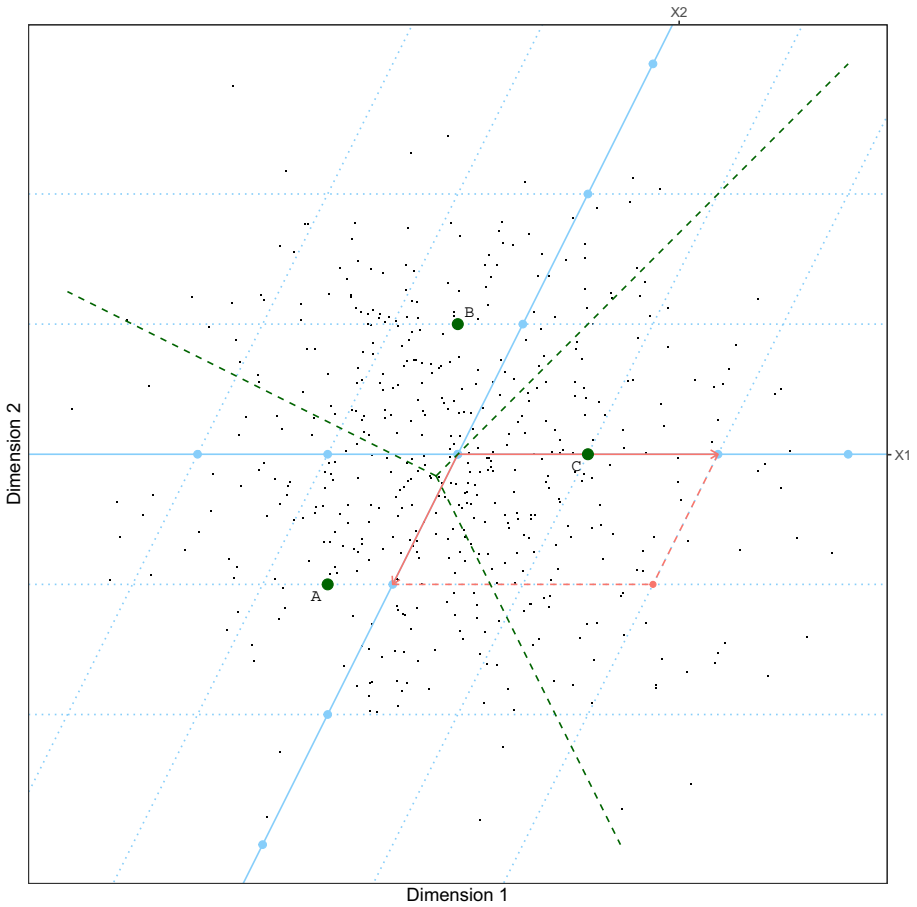

Fig. 1 Example configuration with two predictor variables and three classes. The predictor variables are represented with variable axis, the response classes by points. Object positions can be obtained by completing parallelograms, which is illustrated with the addition of the two red arrows. Also included are the decision lines which equal the Voronoi diagram, that are the green dashed lines equidistant from the class points

In the biplots, the classes are also represented as points. In the exemplary biplot (Fig. 1), three class points are included, where class A has position $(-1, -1)$, class B position $(0, 1)$ and class C position $(1, 0)$.

2.3 Classification Regions, Representation of Probabilities and of Log-Odds

From the biplot, we can determine the most probable class for a given participant. This is simply the class that is closest to the position of the participant. This smallest distance property partitions the Euclidean space in convex regions, where participants in each region have the highest probability for a given class. The partitioning of the Euclidean space is equal to a Voronoi diagram where the line segments are defined by the set of points that are equidistant to the two closest class points. In Fig. 1 these line segments are shown as green

dashed lines. The geometry of these classification regions is the same as for a model with squared Euclidean distances.

Logistic models, like our model, are often interpreted in terms of probabilities, (log) odds, and (log) odds ratios. For every position in the Euclidean space, the probability for each of the C classes can be computed, which is simply the result of our model equation. Because these probabilities depend on the number of classes and the configuration, we only make some general observations on these probabilities. In one-dimensional spaces, the probabilities remain constant (for all classes) outside the range of the class points. In higher dimensional spaces, however, such constant probabilities do not occur. For the distance model, the maximum probability for any class is attained at the position of that class. In contrast, for the squared distance model the maximum probability is not attained at the class point but more to the periphery of the graph, which makes interpretation more difficult.

In multinomial logistic models the (log) odds is of interpretational interest. The log-odds of category A against B is independent of (the position of) class C (or more generally other classes) and is defined by

$$\log\left(\frac{\pi_A(\mathbf{x})}{\pi_B(\mathbf{x})}\right) = d^p(\mathbf{B}^\top\mathbf{x}, \mathbf{v}_B) - d^p(\mathbf{B}^\top\mathbf{x}, \mathbf{v}_A),$$

where p is either 1 or 2, for the distance and squared distance model, respectively. In Fig. 2, we show iso-log-odds curves for the log-odds of class A against B as a function of $\mathbf{u} = \mathbf{B}^\top\mathbf{x}$. Again we contrast the distance model (left-hand side plot) with the squared distance model (right-hand side plot). Two things become apparent from these plots. First, the iso-log-odds curves for the distance model are defined by hyperbolas, whereas those for the squared distance model are defined by straight lines. The straight iso-log-odds curves in the squared distance models imply that any two participants positioned on a line perpendicular to the line AB have the same log-odds, no matter how distant from both points. Therefore, the log-odds for a person on top of one of the class points may have the same log-odds as another person at the periphery of the Euclidean space. Notice that the iso-log-odds line for the log-odds equal to zero is the same for both models, and equal to the line segment between A and B in the Voronoi diagram. Second, the range of log-odds values is much larger for the squared distance model than for the distance model. In fact, the upper bound for the (absolute value of the) log-odds in the distance model is given by the distance between the two points. In Fig. 2, for example, the log-odds cannot exceed $\sqrt{5} \approx 2.236$, which is attained on the line AB beyond the points A or B. For the squared distance model there is not such an implied upper bound.

We can also consider the log-odds as a function of the predictor variables, $\lambda(\mathbf{x}) = \log\left(\frac{\pi_A(\mathbf{x})}{\pi_B(\mathbf{x})}\right)$. As shown in Figs. 1 and 2 the predictor variables are represented by linear variable axes in the biplot. Moving over these variable axes and computing the log-odds value gives the estimated relationship between a predictor variable and the log-odds of two response classes.

Figure 3(a) and (c) show the iso-log-odds curves of A against B again. In Fig. 3(a) we added five horizontal lines representing variable x_1 conditional on five different values for x_2 (i.e., -2 (red), -1, 0, 1, 2 (blue)). These lines are *variable sub-axes*. In Fig. 3 (c) we added five lines representing variable x_2 conditional on five different values for x_1 (i.e., -2 (red), -1, 0, 1, 2 (blue)). These lines correspond to the grid displayed in Fig. 1. In Fig. 3(b), we show $\lambda(\mathbf{x})$ when we move from left to right over the five lines in (a). An increase in x_1 results for the negative values of x_2 (red curves) first in an increase followed by a decrease in log-odds. When $x_2 = 0$ the log-odds monotonically decrease (gray curve), whereas for the

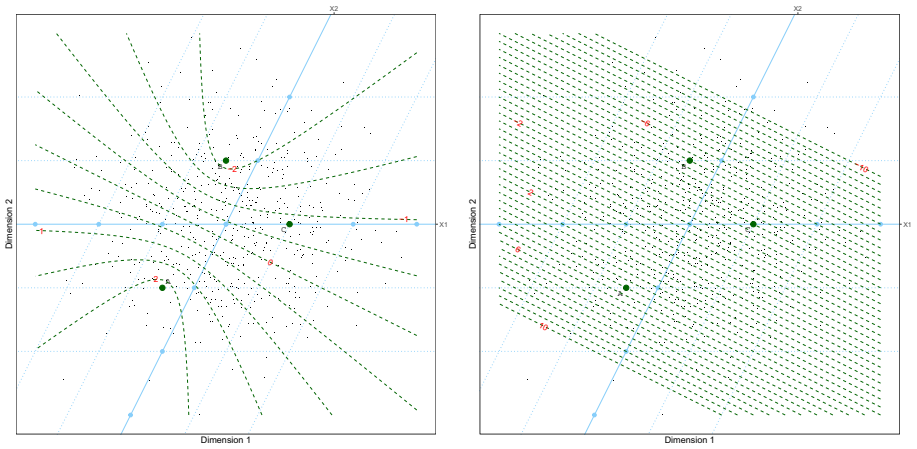

Fig. 2 Example configuration with two predictor variables and three classes. The predictor variables are represented with variable axis, the response classes by points. Also included are iso-log-odds lines for the log-odds of class A against B, that are, the green dashed lines. Left for the distance model, right for the squared distance model

positive values of x_2 (blue curves) there is first a decrease followed by an increase in log-odds. In Fig. 3(d), we show $\lambda(\mathbf{x})$ when we move over the five lines in (c). In this case, where the variable of interest (x_2) is parallel to the AB line, all log-odds values are decreasing, but not at the same rate. The steepest curves are found for a value where the variable sub-axes is cutting the two class points.

In contrast, for the squared distance model these conditional relationships between predictors and log-odds are linear and parallel, that is, the squared distance model is an additive linear model, whereas the distance model is a non-additive nonlinear model.

2.4 Optimization Criterion and Model Selection

We estimate our model by maximizing the likelihood, or, equivalently, minimizing the negative log-likelihood

$$\mathcal{L}(\mathbf{B}, \mathbf{V}) = -\sum_i \sum_c g_{ic} \log \left[\frac{\exp(\theta_{ic})}{\sum_{c'=1}^{C} \exp(\theta_{ic'})} \right],$$

where θ_{ic} is parameterized in terms of the Euclidean distance between a point representing observation i and a point representing class c. In the next section, we will derive an MM algorithm. For maximum likelihood methods, the matrix with second-order derivatives (i.e., the Hessian) gives information about the variance-covariance matrix and therefore the standard errors of the parameters. However, in our MM algorithm, we do not use the log-likelihood function itself, but instead minimize the majorization function. Therefore, our algorithm does not give this Hessian matrix as a by-product of the algorithm. To obtain confidence intervals for the model parameters we advise to use the bootstrap (Efron, 1979; Efron & Tibshirani, 1986).

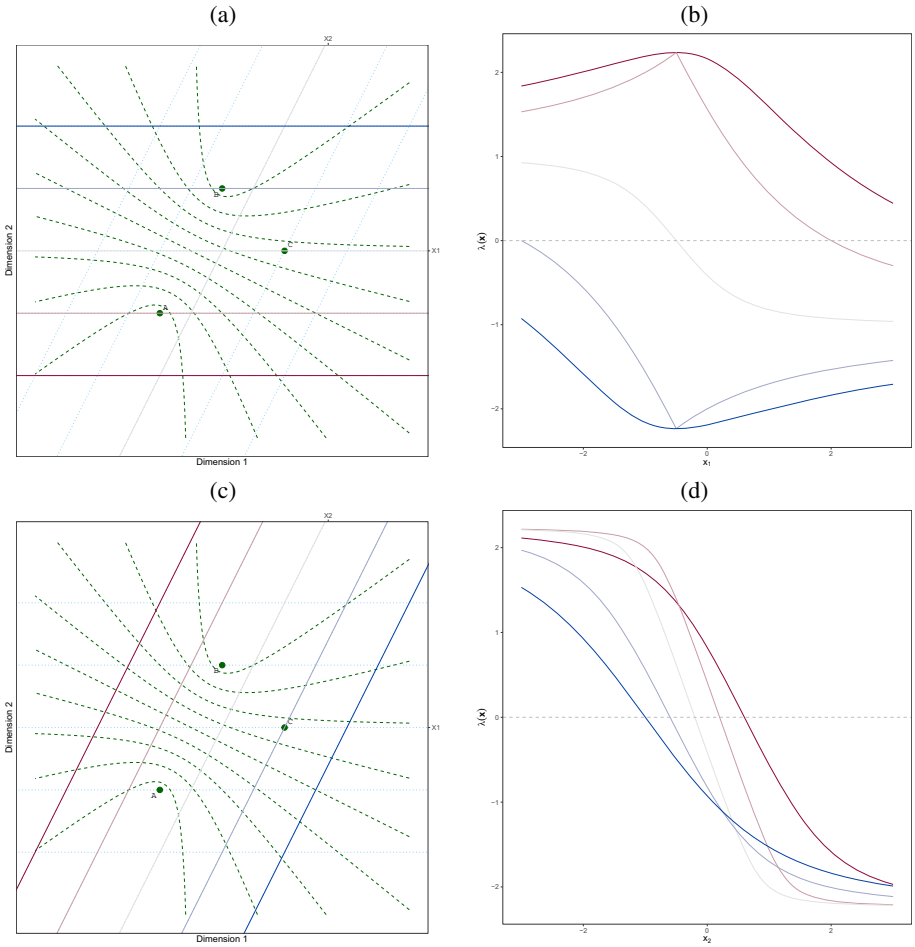

Fig. 3 The log-odds of A against B as a function of one predictor variable (x_1 at the top and x_2 at the bottom) given a fixed value of the other predictor variable. In the left-hand side figures, the variable sub-axes, variable axes for a given value of the other predictor, are displayed. Moving along the sub-axes the log-odds of A against B can be evaluated. These log-odds are displayed in the right-hand side plots. In each plot there are five lines representing different values for the other predictor variable, where red corresponds to negative values (-2, -1), gray to a neutral value (0) and blue to positive values (1, 2)

For finding the optimal dimensionality of the model, information criteria can be used. Akaike's information criterion is given by

$$\text{AIC} = 2\mathcal{L} + 2\text{npar},$$

where "npar" gives the number of estimated parameters. For a distance model with P predictor variables and C response classes in S-dimensional space $\text{npar} = S(P + C - (S - 1)/2)$. The Bayesian information criterion is given by

$$\text{BIC} = 2\mathcal{L} + \log(n)\text{npar},$$

with the same formula for the number of parameters. Note that likelihood ratio statistics cannot be used for dimensionality selection, as these statistics do not have an asymptotic chi-square distribution (Takane et al., 2003).

Besides finding an optimal dimensionality, researchers are often interested in which predictor variables describe the differences between the classes well. Therefore, models with and without a given predictor can be fitted and the fit of both models can be compared using either an information criterion or a likelihood ratio statistic. For the latter we take twice the difference in negative log-likelihood. This difference follows a chi-square distribution under the null hypothesis. For a continuous predictor variable, the degrees of freedom are equal to the dimensionality of the solution, whereas for a categorical predictor it is the dimensionality times the number of category levels minus one.

To verify the fit of the observations, we define the following residual for object i

$$r_i = 2 \sum_c g_{ic} \log \left(\frac{1}{\hat{\pi}_{ic}} \right),$$

such that $\sum_i r_i = 2\mathcal{L}$, the deviance of the model. These residuals show which observations fit good or bad in a specific solution and can be used for further diagnostics, such as finding outliers and influential points.

3 MM Algorithm

In this section, we present an MM algorithm (Heiser, 1995; Hunter & Lange, 2004) for the estimation of the model parameters for the distance model. The abbreviation MM stands for Majorization Minimization, or Minorization Maximization. In our case we want to minimize the negative log-likelihood, so MM has the first meaning. The derivation below follows the work of Groenen et al. (2003), Groenen and Josse (2016), and De Leeuw (2006). The idea of MM for finding a minimum of the function $\mathcal{L}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters, is to define an auxiliary function, called a *majorization function*, $\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta})$ with two characteristics

$$\mathcal{L}(\boldsymbol{\vartheta}) = \mathcal{M}(\boldsymbol{\vartheta}|\boldsymbol{\vartheta})$$

where $\boldsymbol{\vartheta}$ is a supporting point, and

$$\mathcal{L}(\boldsymbol{\theta}) \leq \mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta}).$$

The two equations tell us that $\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta})$ is a function that lies above (i.e., majorizes) the original function and touches the original function at the supporting point. The supporting point is usually defined by the current values of the parameter in the iterative scheme.

An algorithm can be constructed because

$$\mathcal{L}(\boldsymbol{\theta}^+) \leq \mathcal{M}(\boldsymbol{\theta}^+|\boldsymbol{\vartheta}) \leq \mathcal{M}(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}) = \mathcal{L}(\boldsymbol{\vartheta}),$$

where $\boldsymbol{\theta}^+$ is

$$\boldsymbol{\theta}^+ = \text{argmin}_{\boldsymbol{\theta}}\ \mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta}),$$

the updated parameters. Instead of finding the minimum of $\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta})$ it also suffices to find updated parameter estimates that decrease the majorization function. An advantage of MM algorithms is that they always converge monotonically to a (local) minimum. The challenge is to find a parametrized function family, $\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta})$, that can be used in every step and is easy to minimize.

In our case, the original function equals the negative log-likelihood. We will majorize this function with a least squares function. For the majorization of the negative log-likelihood, we use the following theorem.

Quadratic Majorization Theorem: Let $\mathcal{L} : \mathbb{R}^m \to \mathbb{R}$ be a twice differentiable function and suppose there is a positive definite matrix $\mathbf{H}$ such that

$$\mathbf{H} - \partial^2 \mathcal{L}(\boldsymbol{\theta}),$$

is positive semi-definite, then for each $\boldsymbol{\vartheta}$ the convex quadratic function

$$\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta}) = \mathcal{L}(\boldsymbol{\vartheta}) + \mathcal{L}'(\boldsymbol{\vartheta})(\boldsymbol{\theta} - \boldsymbol{\vartheta}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\vartheta})^\top \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\vartheta})$$

majorizes $\mathcal{L}(\boldsymbol{\theta})$ at $\boldsymbol{\vartheta}$ (Böhning & Lindsay, 1988; Böhning, 1992; Hunter & Lange, 2004).

Additionally, we will use the property that majorization is closed under summation, that is, if $\mathcal{M}_1$ majorizes $\mathcal{L}_1$ and $\mathcal{M}_2$ majorizes $\mathcal{L}_2$, then $\mathcal{M}_1 + \mathcal{M}_2$ majorizes $\mathcal{L}_1 + \mathcal{L}_2$.

3.1 MM for Generic Multinomial Model

For our multinomial models, we like to minimize the negative log-likelihood, that is,

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \mathcal{L}_i(\boldsymbol{\theta}_i) = -\sum_i \sum_c g_{ic} \log \left[\frac{\exp(\theta_{ic})}{\sum_{c'=1}^{C} \exp(\theta_{ic'})} \right],$$

where $\boldsymbol{\theta}_i = [\theta_{i1}, \ldots, \theta_{iC}]^\top$ and g_{ic} are the elements of the response indicator matrix ($\mathbf{G}$) introduced before, that is, $g_{ic} = 1$ if $y_i = c$ and zero otherwise. We use the summation property together with the quadratic majorization theorem to find a majorization function for the negative log-likelihood. The first derivative is

$$\frac{\partial \mathcal{L}_i(\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i} = -(\mathbf{g}_i - \boldsymbol{\pi}_i),$$

where $\mathbf{g}_i = [g_{i1}, \ldots, g_{iC}]^\top$ is the response indicator vector for observation i and $\boldsymbol{\pi}_i = [\pi_{i1}, \ldots, \pi_{iC}]^\top$ the corresponding vector with current estimated probabilities. The matrix of second derivatives is given by $\text{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^\top$. Evans (2014) showed that the matrix $\mathbf{H} = \frac{1}{4}\mathbf{I}$ defines a matrix such that $\mathbf{H} - \partial^2 \mathcal{L}_i(\boldsymbol{\theta}_i)$ is positive semi-definite. This matrix is extremely simple to use in an algorithm because of its fixed values.

Using the derivatives and the upper bound on the Hessian in the quadratic majorization theorem we have

$$\mathcal{L}_i(\boldsymbol{\theta}_i) \leq \mathcal{L}_i(\boldsymbol{\vartheta}_i) - (\boldsymbol{\theta}_i - \boldsymbol{\vartheta}_i)^\top (\mathbf{g}_i - \boldsymbol{\pi}_i) + \frac{1}{8}(\boldsymbol{\theta}_i - \boldsymbol{\vartheta}_i)^\top (\boldsymbol{\theta}_i - \boldsymbol{\vartheta}_i),$$

which can be rewritten as

$$\mathcal{L}_i(\boldsymbol{\theta}_i) \leq \mathcal{L}_i(\boldsymbol{\vartheta}_i) - \boldsymbol{\theta}_i^\top (\mathbf{g}_i - \boldsymbol{\pi}_i) + \boldsymbol{\vartheta}_i^\top (\mathbf{g}_i - \boldsymbol{\pi}_i) + \frac{1}{8}\left(\boldsymbol{\theta}_i^\top \boldsymbol{\theta}_i + \boldsymbol{\vartheta}_i^\top \boldsymbol{\vartheta}_i - 2\boldsymbol{\theta}_i^\top \boldsymbol{\vartheta}_i \right).$$

Rearranging the terms gives

$$\mathcal{L}_i(\boldsymbol{\theta}_i) \leq \mathcal{L}_i(\boldsymbol{\vartheta}_i) + \frac{1}{8}\boldsymbol{\theta}_i^\top \boldsymbol{\theta}_i - \frac{1}{8} \cdot 2\boldsymbol{\theta}_i^\top (\boldsymbol{\vartheta}_i + 4(\mathbf{g}_i - \tilde{\boldsymbol{\pi}}_i)) + \frac{1}{8}\boldsymbol{\vartheta}_i^\top \boldsymbol{\vartheta}_i + \boldsymbol{\vartheta}_i^\top (\mathbf{g}_i - \boldsymbol{\pi}_i),$$

where $\mathcal{L}_i(\boldsymbol{\vartheta}_i)$ and the last two terms are constant with respect to $\boldsymbol{\theta}$.

Let us now define $\boldsymbol{\delta}_i = -(\boldsymbol{\vartheta}_i + 4(\mathbf{g}_i - \boldsymbol{\pi}_i))$, so that

$$\mathcal{L}_i(\boldsymbol{\theta}_i) \leq \frac{1}{8}\boldsymbol{\theta}_i^\top \boldsymbol{\theta}_i + \frac{1}{8} \cdot 2\boldsymbol{\theta}_i^\top \boldsymbol{\delta}_i + \text{constant}.$$

Note that $\boldsymbol{\delta}_i$ is independent of $\boldsymbol{\theta}$, so we can simply add $\frac{1}{8}\boldsymbol{\delta}_i^\top \boldsymbol{\delta}_i$ to the equation, to obtain

$$\mathcal{L}_i(\boldsymbol{\theta}_i) \leq \frac{1}{8}\boldsymbol{\theta}_i^\top \boldsymbol{\theta}_i + \frac{1}{8} \cdot 2\boldsymbol{\theta}_i^\top \boldsymbol{\delta}_i + \frac{1}{8}\boldsymbol{\delta}_i^\top \boldsymbol{\delta}_i + \text{constant},$$

which can be rewritten as

$$\mathcal{L}_i(\boldsymbol{\theta}_i) \leq \frac{1}{8}||\boldsymbol{\delta}_i + \boldsymbol{\theta}_i||^2 + \text{constant} = \mathcal{M}_i(\boldsymbol{\theta}_i|\boldsymbol{\delta}_i) + \text{constant}.$$

By using the summation property, we therefore have the following majorization function

$$\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\delta}) = \sum_i \mathcal{M}_i(\boldsymbol{\theta}_i|\boldsymbol{\delta}_i) = \sum_i \sum_c (\delta_{ic} + \theta_{ic})^2,$$

where $\delta_{ic} = -(\vartheta_{ic} + 4(g_{ic} - \pi_{ic}))$, the *working dissimilarities*. In our model we define $\theta_{ic} = -d(\mathbf{u}_i, \mathbf{v}_c)$, so that

$$\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\delta}) = \sum_i \mathcal{M}_i(\boldsymbol{\theta}|\boldsymbol{\delta}_i) = \sum_i \sum_c (\delta_{ic} - d(\mathbf{u}_i, \mathbf{v}_c))^2,$$

which equals the well-known least squares function called "raw STRESS". This function is often used in multidimensional unfolding (see, for example, Heiser (1981)), except that in this case some δ_{ic} might be negative.

3.2 Minimizing Raw STRESS with Negative Dissimilarities

Let us define a matrix $n \times C$ matrix $\mathbf{W}$ with elements w_{ic} equal to one so that we can write our majorizing function as

$$\mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\delta}) = \sum_i \sum_c w_{ic}(\delta_{ic} - d(\mathbf{u}_i, \mathbf{v}_c))^2. \tag{1}$$

For minimization of this function De Leeuw (1977) and De Leeuw and Heiser (1977) proposed the SMACOF algorithm. The SMACOF algorithm is itself an MM algorithm. Convergence properties of this algorithm are described by De Leeuw (1988). Heiser (1981) and Heiser (1987) showed that multidimensional unfolding can be considered a special case of multidimensional scaling. Subsequently, he developed the SMACOF algorithm to deal with rectangular proximity matrices. Further developments are described in Busing (2010). An elementary treatment of the algorithm for multidimensional scaling can be found in Chapter 8 of Borg and Groenen (2005) and for multidimensional unfolding in Chapter 14.

In the SMACOF algorithm, a matrix $\mathbf{A}$, which is related to the Guttman transform (Guttman, 1968), is defined with elements

$$a_{ic} = \begin{cases} 0 & \text{if } d(\mathbf{u}_i, \mathbf{v}_c) = 0 \\ \frac{\delta_{ic}}{d(\mathbf{u}_i, \mathbf{v}_c)} & \text{otherwise} \end{cases}.$$

Following Busing (2010, pages 176 and 183–187) updates for regression coefficients $\mathbf{B}$ and class points $\mathbf{V}$ are given by

$$\mathbf{B}^+ = \left(\mathbf{X}^\top \mathbf{R} \mathbf{X}\right)^{-1} \left[\mathbf{X}^\top (\mathbf{P}\mathbf{U} - \mathbf{A}\mathbf{V}) + \mathbf{X}^\top \mathbf{W}\mathbf{V}\right] \tag{2}$$

and

$$\mathbf{V}^{+} = \mathbf{C}^{-1}\left(\mathbf{Q}\mathbf{V} - \mathbf{A}^{\top}\mathbf{U} + \mathbf{W}^{\top}\mathbf{U}\right), \tag{3}$$

where $\mathbf{R} = \text{diag}(w_{i+})$, $\mathbf{C} = \text{diag}(w_{+c})$, $\mathbf{P} = \text{diag}(a_{i+})$, and $\mathbf{Q} = \text{diag}(a_{+c})$.

This SMACOF algorithm assumes nonnegative dissimilarities. However, we cannot guarantee that the working dissimilarities are nonnegative in every iteration. Heiser (1991) describes a generalized MM algorithm for least squares multidimensional scaling where some of the dissimilarities are negative. His proposal is an adaption of the usual SMACOF algorithm. The line of thought of Heiser's contribution is that two majorizing functions are defined: one for the nonnegative dissimilarities and one for the negative dissimilarities. It turns out that the new algorithm is a simple modification of the standard SMACOF algorithm, where only some elements of the two intermediate matrices $\mathbf{A}$ and $\mathbf{W}$ differ, depending on the sign of the dissimilarity.

The first intermediate matrix, $\mathbf{A}$, has additional elements set to zero for negative dissimilarities. That is, the elements of $\mathbf{A}$ are as

$$a_{ic} = \begin{cases} 0 & \text{if } \delta_{ic} < 0 \text{ or } d(\mathbf{u}_i, \mathbf{v}_c) = 0, \\ \frac{\delta_{ic}}{d(\mathbf{u}_i, \mathbf{v}_c)} & \text{otherwise.} \end{cases}$$

Also the second intermediate matrix, the weight matrix $\mathbf{W}$, needs to be redefined as a function of the sign of the (working) dissimilarities. This redefinition clearly distinguishes Heiser's two majorization functions as for nonnegative dissimilarities $w_{ic} = 1$ and for negative dissimilarities

$$w_{ic} = \begin{cases} \frac{d(\mathbf{u}_i, \mathbf{v}_c) + |\delta_{ic}|}{d(\mathbf{u}_i, \mathbf{v}_c)} & \text{if } d(\mathbf{u}_i, \mathbf{v}_c) > 0, \\ \frac{\epsilon + \delta_{ic}^2}{\epsilon} & \text{otherwise,} \end{cases}$$

where ϵ is a small positive constant. Heiser (1991) showed that with these redefined matrices we can still use the standard updating formulas (2) and (3). Repeatedly updating $\mathbf{B}$ and $\mathbf{V}$ and setting $\mathbf{U} = \mathbf{X}\mathbf{B}$, adapting $\mathbf{A}$ and $\mathbf{W}$ in each step minimizes the negative log-likelihood.

3.3 Algorithm

The algorithm just described is summarized in Algorithm 1. To start the algorithm some initial values for the parameters are needed. With the current parameter values, fitted values and working dissimilarities can be computed. With these working dissimilarities, we start the least squares restricted multidimensional unfolding procedure in the inner loop. This inner loop iterates till convergence or till a maximum number of inner iterations is achieved, the default is 32. With the updated parameters the value of the negative log-likelihood is computed and convergence is monitored by computing the decrease in negative log-likelihood values. If convergence is not yet achieved and the maximum number of outer iterations is not yet achieved, another outer iteration is started where new working dissimilarities are computed. After convergence (or when the maximum number of outer iterations is achieved) we rotate the solution so that $\mathbf{U}$ is in principle orientation. As the algorithm uses majorization in computing the working dissimilarities as well as in updating $\mathbf{B}$ and $\mathbf{V}$ it is a *double MM* algorithm.

We need to remark that the loss function is consistently minimized by this algorithm. However, the algorithm does not guarantee that the obtained minimum is the global minimum. This so-called convergence to local minimum problem may be mitigated by either choosing

Algorithm 1 Multinomial restricted unfolding algorithm minimizing $\mathcal{L}(\cdot)$ using a double majorization algorithm.

procedure MRU($\mathbf{G}$, $\mathbf{X}$, $\mathbf{B}$, $\mathbf{V}$)
  predefine: maxouter, maxinner, ϵ_1, ϵ_2
  compute $\mathbf{U} \leftarrow \mathbf{XB}$
  compute $\mathbf{\Pi}$
  assess $\mathcal{L}^0(\mathbf{B}, \mathbf{V})$
  for $t_1 \leftarrow 1$, maxouter **do**
    compute $\mathbf{\Delta} \leftarrow d(\mathbf{U}, \mathbf{V}) - 4(\mathbf{G} - \mathbf{\Pi})$
    assess $\mathcal{M}^0(\mathbf{B}, \mathbf{V}|\mathbf{\Delta})$
    for $t_2 \leftarrow 1$, maxinner **do**
      compute $\mathbf{A}$ and $\mathbf{P}$, $\mathbf{Q}$
      compute $\mathbf{W}$ and $\mathbf{R}$, $\mathbf{C}$
      compute $\mathbf{B} \leftarrow \left(\mathbf{X}^\top \mathbf{RX}\right)^{-1} \left[\mathbf{X}^\top (\mathbf{PU} - \mathbf{AV}) + \mathbf{X}^\top \mathbf{WV}\right]$
      compute $\mathbf{U} \leftarrow \mathbf{XB}$
      compute $\mathbf{V} \leftarrow \mathbf{C}^{-1} \left(\mathbf{QV} - \mathbf{A}^\top \mathbf{U} + \mathbf{W}^\top \mathbf{U}\right)$
      assess $\mathcal{M}^{t_2}(\mathbf{B}, \mathbf{V}|\mathbf{\Delta})$
      if $\mathcal{M}^{t_2}(\mathbf{B}, \mathbf{V}|\mathbf{\Delta}) - \mathcal{M}^{t_2-1}(\mathbf{B}, \mathbf{V}|\mathbf{\Delta}) < \epsilon_1$: break
    end for
    compute Π
    if $\mathcal{L}^{t_1}(\mathbf{B}, \mathbf{V}) - \mathcal{L}^{t_1-1}(\mathbf{B}, \mathbf{V}) < \epsilon_2$: break
  end for
  eigenvalue decomposition $\mathbf{U}^\top \mathbf{U}$: $\mathbf{E\Phi E}^\top$
  rotate $\mathbf{B} \leftarrow \mathbf{BE}$ and $\mathbf{V} \leftarrow \mathbf{VE}$
  return($\mathbf{B}$, $\mathbf{V}$)
end procedure

good (or rational) initial parameter values or by using many random starts. Good rational starting values can be obtained by performing, for example, a linear discriminant analysis or a canonical correspondence analysis. However, even these good starts do not guarantee to find the global minimum.

4 Simulation Studies

In this section, we describe two simulation studies. The first is a relatively small simulation study to test whether the algorithm is able to recover a set of population parameters. The second simulation study compares the predictive performance of multinomial restricted unfolding to that of a similar model based on squared distances and to multinomial logistic regression. We used the parameter estimates of the analyses of the two empirical data sets described in the next section (the Liver data in Section 5.1 and the Dutch Election data in Section 5.2), to be the population parameters in these simulation studies. In the Liver data set, there are 4 response classes and three predictors, while in the Dutch Election data set there are 8 response classes and 5 predictors.

4.1 Parameter Recovery

We use the covariance matrix of the predictors and the estimated model parameters ($\mathbf{B}$ and $\mathbf{V}$) as population parameters, the values are given in the Appendix. With the covariance matrix we draw random variables from the multivariate normal distribution with mean zero.

We used sample sizes 100, 200, 500, 1000, and 2000. We multiplied the predictors with the population regression coefficients to obtain object positions in two-dimensional space. Using the distances between the object positions and the class positions, we computed probabilities for every observation and every class (see Section 2.1). With these probabilities, we sampled observed responses from the multinomial distribution. For each of the five sample sizes we draw 1000 data sets.

For both populations, we therefore generated 5000 data sets. On each of these generated data sets we fitted the multinomial restricted unfolding in two dimensions. As starting values for the algorithm we used the population parameters to avoid the local minimum problem. To rule out differences due to rotation, we used an orthogonal Procrustes analysis on the estimated solutions to optimally rotate them towards the population values.

In every replication, for every sample size, we obtain the estimated regression weights ($\hat{\mathbf{B}}$) and class coordinates ($\hat{\mathbf{V}}$). With these estimates, we define in every replication r the following root mean squared error statistics

$$\text{RMSE}_b(r) = \sqrt{\frac{1}{PS}\sum_{p=1}^{P}\sum_{s=1}^{S}(b_{ps} - \hat{b}_{ps})^2}$$

and

$$\text{RMSE}_v(r) = \sqrt{\frac{1}{CS}\sum_{c=1}^{C}\sum_{s=1}^{S}(v_{cs} - \hat{v}_{cs})^2}.$$

The average and standard deviation of these statistics over the 1000 replications are shown in Table 1. Overall, the results are as expected: The performance of the algorithm becomes better with larger sample sizes and the average RMSE and its standard deviation become smaller.

Comparing the results for the two populations, we see that for the second population the recovery is better, that is, smaller values for RMSE and its standard deviation. This second population is based on the Dutch Election data with 8 classes and 5 predictors. Because the number of classes and predictors is larger than in the first population there is more structure

Table 1 Results from the two simulation studies

Population	Sample size	$Mean(\text{RMSE}_b)$	$sd(\text{RMSE}_b)$	$Mean(\text{RMSE}_v)$	$sd(\text{RMSE}_v)$
Liver	100	2.418	2.563	2.437	2.209
	200	0.968	0.848	1.093	0.851
	500	0.395	0.281	0.491	0.322
	1000	0.225	0.136	0.298	0.159
	2000	0.141	0.081	0.187	0.093
Dutch	100	0.701	0.747	1.555	1.297
Election	200	0.345	0.277	0.822	0.517
	500	0.152	0.080	0.394	0.172
	1000	0.092	0.042	0.243	0.094
	2000	0.058	0.022	0.150	0.055

$Mean(\cdot)$ represents the mean over 1000 replications and $sd(\cdot)$ the standard deviation. RMSE is the root mean squared error

(i.e., less freedom) in the configuration and this is beneficial for recovery of the parameter values. Overall, we conclude from these simulation studies that the algorithm works well.

4.2 Predictive Performance

In the second simulation study we compare the predictive performance of multinomial restricted unfolding to that of a similar model with squared distances and to multinomial logistic regression. We defined six different populations based on the two empirical data sets to be discussed in the next Section. That is, on each data set we fitted a multinomial restricted unfolding model, an ideal point classification model (ipc; the model with squared distances) and a multinomial logistic regression model. The estimated parameters of these three analyses were subsequently taken as population parameters. Also the covariance matrix of the predictors for these empirical data sets was taken as the covariance matrix in the population. All population parameter values are shown in the Appendix.

For each of the six populations (three models; two data sets), we generated a training set of size 100, 200 or 500. Furthermore, a test set with 1000 observations is drawn from the same population distribution. The training data sets were analyzed by each of the three models using the population values as starting values for the algorithm. With the estimated parameters and the values of the predictor variables for the observations in the test set, predictions of the response variable in the test set were made. These predictions are the probabilities (π_{ic}) for each class. The overall quality of these predictions is defined by the following measure of prediction error:

$$-2 \sum_i \sum_c g_{ic} \log(\pi_{ic}).$$

For each of the six populations and for each of the sample sizes we generated 100 training and test sets.

The results of this simulation study are shown in Fig. 4 for the populations based on the Liver data and Fig. 5 for the populations based on the Dutch Election data.

In Fig. 4, we see that the predictive performances of the three models do not differ much. A general trend is that the model predicts best for the congruent populations, that is, multinomial logistic regression gives the best predictions when the population model is a multinomial logistic regression model and similarly for the other two models. The differences between predictive performance of the three models become smaller for larger sample sizes. Most noticeable to these results are the large outliers in predictive performance for multinomial logistic regression and the squared distance model (ipc), these are a result of overfitting in the training data sets.

In Fig. 5 we see the results for the second set of populations based on the Dutch Election data. We have to note that in 39 cases we have no measure of prediction error for multinomial logistic regression. In those cases the estimated regression coefficients became really large, so that predicted probabilities became zero and one. This happened 39 times, two times for a sample size of 200 and 37 times for a sample size of 100. Of the 39 occasions, 9 times the data were generated following the distance model (mru), 10 times the data were generated with the squared distance model (ipc), and the other 20 times the data were generated by multinomial logistic regression (mlr). In Fig. 5 we see that for small sample sizes (i.e., 100) multinomial restricted unfolding gives the best predictions no matter what the population model is. For large sample sizes the predictions are about equally good.

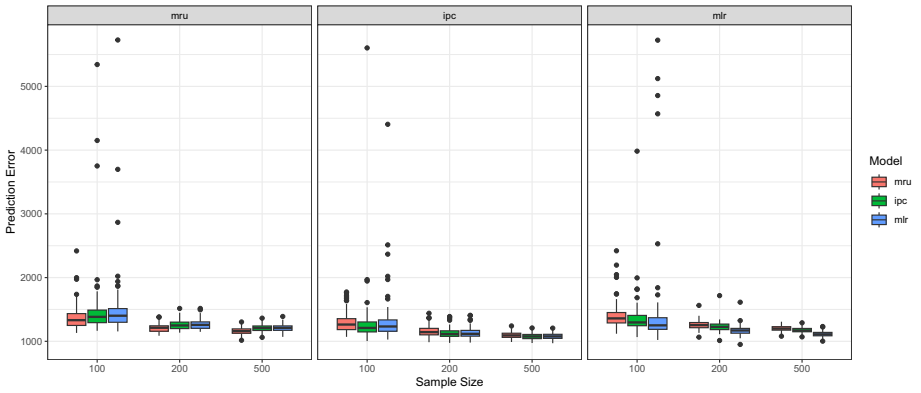

Fig. 4 Prediction error for multinomial restricted unfolding (mru; distances), ideal point classification (ipc; squared distances) and multinomial logistic regression (mlr) for data generated from three different population models based on the Liver data

5 Applications

In this section, we will apply multinomial restricted unfolding to two empirical data sets. With the first data set, we compare in detail the results and interpretation of our model against the model with squared distances. Using the second data set, we showcase model selection and interpretation.

5.1 Liver Data

This data set (Plomteux, 1980; Lesaffre & Albert, 1989) consists of 218 patients classified in 4 different disease classes: Acute Viral Hepatitis (AVH, 57 patients), Persistent Chronic Hepatitis (PCH, 44 patients), Aggressive Chronic Hepatitis (ACH, 40 patients) and Post-Necrotic Cirrhosis (PNC, 77 patients). As predictor variables, three liver function tests are available: ASpartate aminotransferase (AS), ALanine aminotransferase (AL), and Glutamate

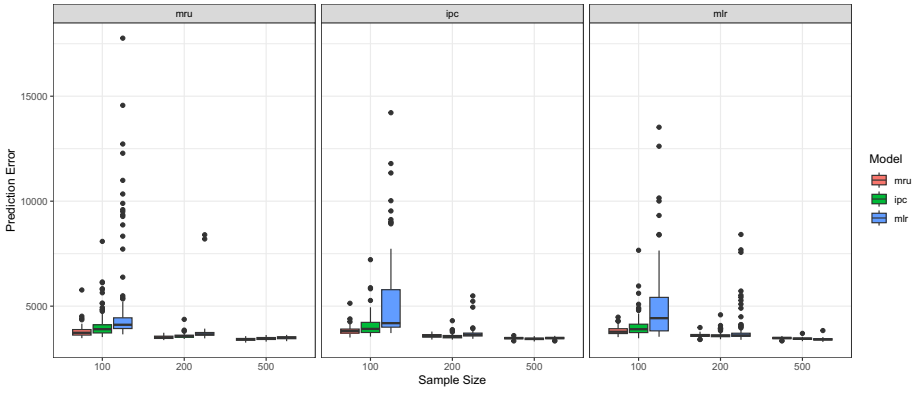

Fig. 5 Prediction error for multinomial restricted unfolding (mru; distances), ideal point classification (ipc; squared distances) and multinomial logistic regression (mlr) for data generated from three different population models based on the Dutch Election data

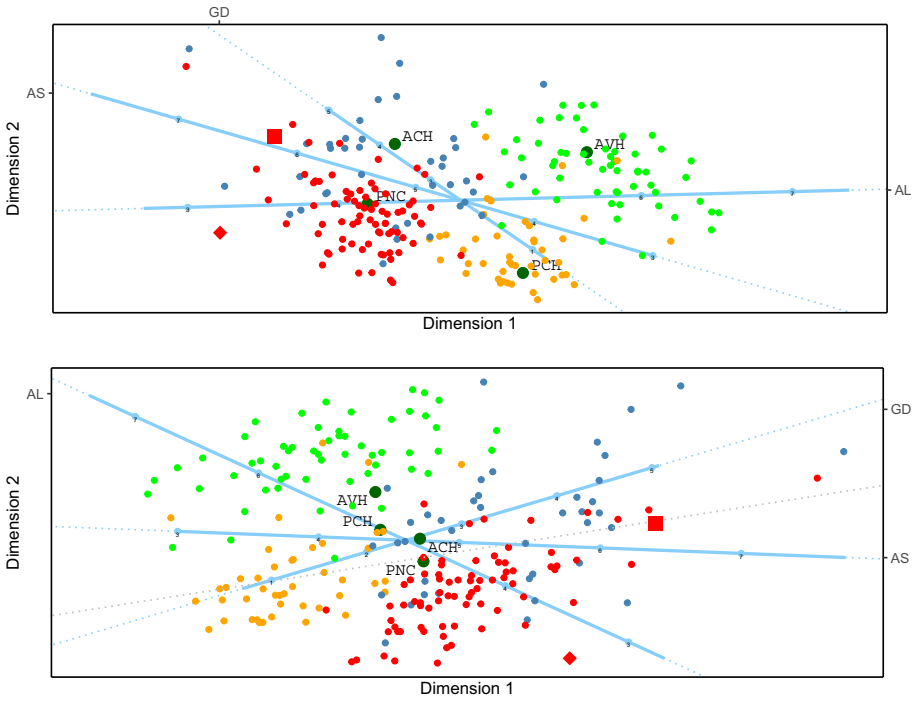

Fig. 6 Biplots for distance model (top) and squared distance model (bottom) for the liver data. Points are colored by observed class

Dehydrogenase (GD). All three predictor variables are first log-transformed to deal with skewness and then centered.

5.1.1 Analysis

The deviances of the one-, two-, and three-dimensional models, are 282.88, 196.46, and 178.39, with 7, 13, and 18 parameters, respectively. Therefore, the AIC is smallest for the three-dimensional model, while the BIC is smallest for the two-dimensional model. So, either the two-, or three-dimensional model seems to be optimal.

The two-dimensional model with squared distances has an estimated deviance value of 203.00, with 11 parameters[1]. Compared to the two-dimensional distance model, this results in a somewhat higher AIC and a somewhat lower BIC. The in-sample classification performance shows that the distance model classifies 181 participants correctly, while the squared distance model classifies 178 correctly. The statistics do not point out which model is preferred for these data.

The two-dimensional solutions for both the distance and squared distance model are given in Fig. 6. We see large congruences between the two biplots, that is, the variable axis for the three predictors are very much alike and also the configurations of class points are very

[1] The distance model only has rotational freedom and identification can be obtained by rotating to principal orientation. Identification of the model based on squared distances is more involved and depends on both the number of classes and number of dimensions. Sometimes besides rotational freedom there also is scaling freedom. See De Rooij (2009) for details.

similar (up to a rotation of the solution). A noticeable difference is that in the distance models the class points are much more in the center of the patients belonging to the classes, whereas in the squared distance model the class points are more in the center and the patient positions more in the periphery.

The distances between the class points give an upper limit for the log-odds in the distance model. The smallest distance is 2.53 between ACH and PNC, therefore, the log-odds in favor of ACH instead of PNC (or the other way around) can never exceed this 2.53. The largest distance is between PNC and AVH, so these two classes are better distinguished by the model. Such upper bounds on the log-odds are not implied by the distances between the classes in the squared distance model (see also Fig. 2), so in the squared distance models these distances are less informative.

Furthermore, in the squared distance solution there is one participant (red point) located almost on top of the class point of PNC. From a distance perspective, one would expect that this person has the highest probability for PNC. This is, however, not the case in the squared distance model, as the participant only has a probability of 0.51 for the class PNC. The participant with the highest probability for PNC (0.98) is located at the bottom of the figure, indicated by a diamond shape symbol. In the biplot of the distance model we also show the same person with a diamond although it does not have the largest probability for this solution. For the distance model the probability peaks at the class point, so that, when a person is located on top of a class points they have the highest probability for that specific class.

To point out another difference between the distance and squared distance model, we added a gray dotted line to the squared distance plot. This line is perpendicular on the line joining the class points PNC and ACH. Therefore, participants located on this line all have the same estimated log-odds in the squared distance model (compare Fig. 2). One participant on this line lies almost on top of the class point for PNC (same person as in previous paragraph). However, there are also other participants on this line further to the right and quite distant from both these two classes. Notice the participant represented by the red square located on this line (the same participant is also indicated by a red square in the distance model). In the squared distance model, this participant has the same estimated log-odds of PNC against ACH as the participant located on the PNC class point. Its distances, however, to the two class points are almost equal: 6.16 and 6.19. Yet, although the distances are almost equal, the estimated log-odds for this participant is the same as for the participant which clearly has a smaller distance towards PNC than to ACH. For the distance model, the distances towards the class points matter (instead of distances to the decision line) and such aberrant interpretation does not occur.

5.2 Dutch Election Data

This data set consists of 275 inhabitants of the Netherlands for whom their vote in the Dutch parliamentary election of 2002 was recorded (Irwin et al., 2003). The response variable is the political party the participant voted for with 8 categories abbreviated as PvdA (43, Labour Party), CDA (72, Christian Democrats), VVD (49, Conservative Liberals), D66 (27, Progressive Liberals), GL (28, Green Left), CU (7, Christian Union), LPF (28, right-wing populist and nationalist), and SP (21, Socialist Party), where the number between parentheses indicate the number of occurrences.

Additionally, participants were asked their opinion on five issues. On a 7-point scale they indicated whether they think that Euthanasia (E) should always be forbidden (1) or

Table 2 Model selection for Dutch Election data

(a) Dimensionality selection				(b) Variable selection					
Dimensionality	Deviance	npar	AIC	Left out X	Deviance	npar	AIC	LRT	p
1	968.56	13.00	994.56	–	924.55	25.00	974.55	–	–
2	924.55	25.00	974.55	E	940.78	23.00	986.78	16.24	0.00
3	905.15	36.00	977.15	ID	929.93	23.00	975.93	5.39	0.07
4	895.99	46.00	987.99	AS	934.64	23.00	980.64	10.10	0.01
5	895.62	55.00	1005.62	C	931.26	23.00	977.26	6.71	0.04
				LR	975.11	23.00	1021.11	50.56	0.00

npar, number of estimated parameters; AIC, Akaike's Information Criterion; LRT, Likelihood Ratio Test; p, p-value

that a doctor should always be allowed to end a life upon a patient's request (7). Similarly, whether Income Differences (ID) should be increased (1) or decreased (7). The third issue concerns Asylum Seekers (AS), and whether the participants have the opinion that the Dutch Government should allow more asylum seekers to enter (1) or should send back as many asylum seekers as possible (7). The next issue is about the acting of the government about Crime (C), that is whether the government is acting too tough on crime (1) or should act tougher on crime (7). Finally, the participants had to indicate their location on a Left-Right (LR) continuum, where 0 indicates left and 10 right. We centered and scaled these predictor variables.

5.2.1 Analyses

The deviances, number of parameters, and corresponding AIC statistics for the 1 till 5 dimensional solutions are given in Table 2a. The two-dimensional solution seems to be optimal. Within this two-dimensional solution, we left out, one by one, each of the predictor variables to check their contribution. Results of this variable selection procedure are shown in Table 2b, where it can be seen that all variables have a significant contribution except for "Income Differences". For this variable the likelihood ratio statistic indicates a lack of evidence of an effect, whereas the AIC statistic indicates the predictor should not be removed from the model. As there is no clear evidence, we keep the predictor in the model.

The final representation of the data is shown in Fig. 7. The results show a cluster of left-wing parties (SP, PvdA, D66, and GL) at the top and a cluster of more right-wing parties (CDA, VVD, LPF and CU) at the bottom. Predictor variables with a large effect are left-right self scaling (LR), crime (C), and euthanasia (E), while asylum seekers (AS) and income

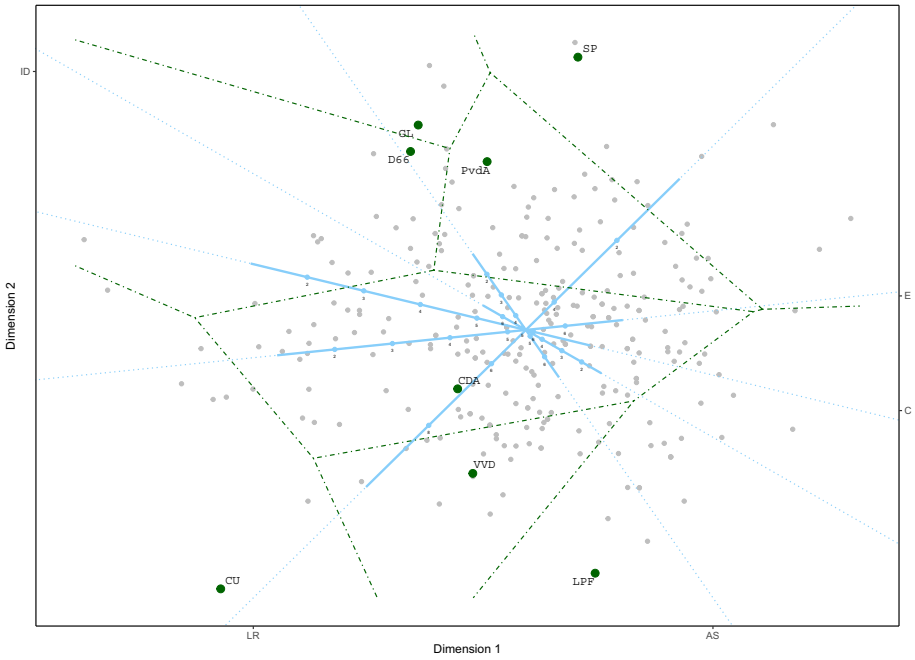

Fig. 7 Biplot for the Dutch Election data

differences (ID) have smaller effects as can be seen from the length of the solid part of the variable axes.

The labels of the predictor variables are printed on the positive side of the variable, so that with increasing left-right scaling (i.e., from left (negative) to right (positive)) participants tend to vote more for CDA, VVD, LPF and CU and less for SP, PvdA, GL and D66. Participants that have the opinion that more Aylum Seekers should be sent back have a higher probability to vote for LPF. Participants scoring low on Euthanasia, that is, Euthanasia should always be forbidden, have a higher probability for voting CU. Participants who indicate the government should act tougher on Crime have a higher probability to vote for the LPF.

Briefly, a numerical comparison of the distance model as compared to the squared distance model. The deviance of the two-dimensional model with squared distances (including the same predictors) has a deviance of 931.33 as compared to 924.55 for the distance model, with the same number of parameters. The AIC for the squared distance model is 981.33, somewhat higher than the AIC for the distance model (974.55). The biplot (not shown) is quite similar to the one in Fig. 7. The classification performance of the distance model is slightly better than that of the squared distance model, that is, the within-sample percentage correct is 34.18 for the distance model and 32.72 for the squared distance model.

6 Discussion

We proposed and investigated multinomial restricted unfolding, a new multinomial logistic distance model for supervised classification of multiclass data, based on Euclidean distances instead of squared Euclidean distances as proposed earlier (Takane et al., 1987; Takane, 1987; De Rooij, 2009).

The distance model has three major advantages over the squared distance model: (1) The probability for a class peaks at its location in the low dimensional configuration; (2) The interpretation is in terms of distances towards the class points, whereas in the squared distance model the interpretation is in terms of the distance towards the decision boundary; and (3) The distance between two class points represents an upper bound for the estimated log-odds of choosing one of these classes over the other.

A main consequence of the first advantage is that the points for the participants are located around a class point (see Fig. 6), whereas in the squared distance model locations of participants are often pushed to the boundaries as higher probabilities are attained there. When interpreting a distance plot, researchers look at the distances between points. However, when the distance from a participant towards two class points is almost equal the researcher will reason that the probabilities are almost equal as well. In Section 5, we pointed out that this leads to incorrect conclusions for the squared distance model, whereas such a conclusion is correct for the distance model. Finally, although in both types of displays the distance between two class points can be interpreted as discriminatory power, only in the distance model this distance gives an upper bound to the log-odds.

In a simulation study (Section 4.2), we compared the predictive performance of the distance model (mru) against the predictive performance of the squared distance model (ipc) and multinomial logistic regression (mlr). A main finding of this simulation study is that the distance model is less vulnerable to overfitting in the case of small sample sizes. The prediction errors for both the squared distance model and multinomial logistic regression have much larger outliers (see Figs. 4 and 5). Furthermore, for the populations based on the Dutch

Election data, the distance model consistently outperforms the two other model in terms of prediction error for small sample sizes.

In both the distance and the squared distance model, there is an additive and a linear relationship between the predictor variables and the coordinates of the participants in the low dimensional configuration. For the squared distance model, this linear additive relationship also transfers to the log-odds. For the distance model, however, this is not the case (e.g., Fig. 3). In multinominal restricted unfolding, we have a nonlinear, non-additive relationship between the predictors and the log-odds. As a result, whereas the squared distance model becomes equivalent to multinomial regression in maximum dimensionality, multinomial restricted unfolding might fit better in terms of the likelihood than multinomial regression in maximum dimension.

A main contribution of this paper is the MM algorithm for estimating the model parameters of multinomial restricted unfolding. By combining earlier results of De Leeuw (2006) and Heiser (1991), we developed a monotonically convergent algorithm. In the main MM step the negative log-likelihood function is majorized with a least squares function. For multidimensional unfolding, algorithms for least squares estimation are available (Heiser, 1981; Busing, 2010) also for restricted multidimensional unfolding. In our algorithm, however, some working dissimilarities in a given iteration might be negative. Therefore, we generalized the algorithm of Heiser (1991) to the restricted unfolding case. The latter algorithm is also an MM algorithm, therefore we have a *double MM* algorithm. It is well known that MM algorithms can be slow (Heiser, 1995). As a result our double MM algorithm is also slow. We therefore translated our original R code to C code to speed up the estimation process. The algorithm is implemented in the `lmap` R-package (De Rooij & Busing, 2022). In a small simulation study, we checked the performance of the algorithm and concluded that it worked well (see Section 4).

Appendix. Population parameters

Simulation Study 1: Parameter Recovery

For the parameter recovery study the population values of the parameters are given in the next section considering the simulation study about predictive performance, where the values are given after the bullets for the distance model (mru).

Simulation Study 2: Predictive Performance

Populations Based on the Liver Data

We generated three predictor variables from a multivariate normal distribution with mean equal to zero and covariance matrix

$$\begin{bmatrix} 0.73 & 0.75 & 0.35 \\ 0.75 & 1.29 & 0.31 \\ 0.35 & 0.31 & 0.48 \end{bmatrix}.$$

This same covariance matrix was used for three different population models. The three population models are:

- For data generation with the distance model, multinomial restricted unfolding (mru), we used

$$\mathbf{B}' = \begin{bmatrix} -3.88 & 6.59 & -1.36 \\ 1.12 & 0.17 & 0.93 \end{bmatrix}$$

and

$$\mathbf{V}' = \begin{bmatrix} 4.85 & 2.39 & -2.52 & -3.56 \\ 1.84 & -2.79 & 2.16 & -0.14 \end{bmatrix}.$$

- For data generation with the squared distance model, ideal point classification (ipc), we used

$$\mathbf{B}' = \begin{bmatrix} 0.53 & -2.15 & -0.28 \\ 3.11 & -3.43 & 1.79 \end{bmatrix}$$

and

$$\mathbf{V}' = \begin{bmatrix} -1.36 & -0.36 & 0.01 & 0.61 \\ -0.63 & -0.63 & 0.37 & 0.38 \end{bmatrix}.$$

- For data generation with multinomial logistic regression (mlr) we used the following regression weights

$$\mathbf{B} = \begin{bmatrix} 0.00 & 1.70 & 2.24 & 1.86 \\ 0.00 & -0.18 & 7.51 & 9.50 \\ 0.00 & -3.54 & -10.77 & -13.49 \\ 0.00 & 0.66 & 5.61 & 4.47 \end{bmatrix}.$$

Populations Based on the Dutch Election Data

We generated five predictor variables from a multivariate normal distribution with mean equal to zero and covariance matrix

$$\begin{bmatrix} 1.00 & 0.04 & 0.06 & 0.08 & 0.01 \\ 0.04 & 1.00 & -0.11 & -0.01 & -0.29 \\ 0.06 & -0.11 & 1.00 & 0.42 & 0.35 \\ 0.08 & -0.01 & 0.42 & 1.00 & 0.33 \\ 0.01 & -0.29 & 0.35 & 0.33 & 1.00 \end{bmatrix}.$$

This same covariance matrix was used for three different population models. The three population models are:

- For data generation with the distance model, multinomial restricted unfolding (mru), we used

$$\mathbf{B}' = \begin{bmatrix} 1.03 & -0.32 & 0.22 & 0.74 & -0.65 \\ 0.11 & 0.18 & -0.31 & -0.18 & -0.64 \end{bmatrix}$$

and

$$\mathbf{V}' = \begin{bmatrix} -0.43 & -0.76 & -0.59 & -1.28 & -1.20 & -3.40 & 0.77 & 0.58 \\ 1.88 & -0.66 & -1.60 & 1.99 & 2.28 & -2.89 & -2.71 & 3.04 \end{bmatrix}.$$

- For data generation with the squared distance model, ideal point classification (ipc), we used

$$\mathbf{B}' = \begin{bmatrix} -0.07 & -0.06 & 0.09 & 0.04 & 0.28 \\ -0.52 & 0.10 & -0.19 & -0.22 & 0.17 \end{bmatrix}$$

and

$$\mathbf{V}' = \begin{bmatrix} -1.00 & 0.58 & 0.93 & -1.08 & -1.10 & 1.53 & 1.32 & -1.44 \\ 0.29 & 0.29 & 0.21 & 0.64 & 0.66 & 1.33 & -0.49 & -0.15 \end{bmatrix}.$$

- For data generation with multinomial logistic regression (mlr) we used the following regression weights

$$\mathbf{B} = \begin{bmatrix} 0.00 & 0.69 & 0.17 & -0.45 & -0.54 & -3.37 & -0.89 & -1.03 \\ 0.00 & -0.12 & -0.04 & -0.07 & -0.19 & -1.68 & 0.72 & 0.32 \\ 0.00 & -0.13 & -0.41 & 0.06 & -0.01 & -0.12 & -0.40 & -0.06 \\ 0.00 & 0.47 & 0.18 & -0.12 & -0.03 & 0.52 & 1.00 & 0.34 \\ 0.00 & -0.17 & 0.16 & -0.53 & -0.56 & -0.17 & -0.09 & 0.12 \\ 0.00 & 1.01 & 1.20 & 0.57 & 0.20 & 1.50 & 1.21 & -0.50 \end{bmatrix}.$$

Acknowledgements The authors would like to thank Willem Heiser for his comments on an earlier version of this manuscript. The Dutch Election data utilized in this manuscript were originally collected for the Dutch Parliamentary Election Studies 2002 and 2003 by Galen A. Irwin, Joop J.M. van Holsteyn and Josje M. den Ridder on behalf of the Foundation for Electoral Research in the Netherlands (Stichting Kiezersonderzoek Nederland, SKON). These studies have been made possible by grants from Dutch Organization for Scientific Research (NWO), the Ministry of the Interior and Kingdom Relations (BZK), the Remote E-Voting Project (Kiezen op Afstand, KOA) of the Ministry of the Interior and Kingdom Relations (BZK), the Ministry of Health, Welfare and Sports (VWS), the Social and Cultural Planning Office (SCP), and the Department of Political Science, Leiden University. The original collectors of the data do not bear any responsibility for the analyses or interpretations in this manuscript.

Data Availability The two data sets and R code for the analyses are available on the `github`-page of the first author.

References

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics, 44*(1), 197–200.

Böhning, D., & Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics, 40*(4), 641–663.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

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

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

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. Barra, F. Brodeau, G. Romier, & B. Van Cutsem (Eds.), *Recent Developments in Statistics* (pp. 133–146). North Holland Publishing Company.

De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification, 5*, 163–180.

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., & Heiser, W. J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. In J. Lingoes, E. Roskam, & I. Borg (Eds.), *Geometric representations of relational data* (pp. 735–752). Mathesis Press.

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

De Rooij, M. and Busing, F. M. T. A. (2022). *lmap: Logistic Mapping*. R package version 0.1.1.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*, 54–75.
Evans, G. W. (2014). *Logistic Gifi: A Logistic Distance Association Model for Exploratory Analysis of Categorical Data*. PhD thesis, UCLA.
Gower, J., & Hand, D. (1996). *Biplots*. Taylor & Francis.
Gower, J., Lubbe, S., & Le Roux, N. (2011). *Understanding biplots*. Wiley.
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.
Groenen, P. J. F. and Josse, J. (2016). Multinomial multiple correspondence analysis. arXiv:1603.03174.
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. (1987). Joint ordination of species and sites: the unfolding technique. In P. Legendre & L. Legendre (Eds.), *Developments in Numerical Ecology* (pp. 189–221). Springer.
Heiser, W. J. (1991). A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative. *Psychometrika, 56*(1), 7–27.
Heiser, W. J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In W. J. Krzanowski (Ed.), *Recent Advances in Descriptive Multivariate Analysis* (pp. 157–189). Clarendon Press.
Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician, 58*(1), 30–37.
Irwin, G., van Holsteyn, J., and den Ridder, J. (2003). *Nationaal Kiezersonderzoek, NKO 2002 2003*. DANS.
Lesaffre, E., & Albert, A. (1989). Multiple-group logistic regression diagnostics. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 38*(3), 425–440.
Plomteux, G. (1980). Multivariate analysis of an enzymic profile for the differential diagnosis of viral hepatitis. *Clinical Chemistry, 26*(13), 1897–1899.
Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika, 52*(4), 493–513.
Takane, Y. (1998). Visualization in ideal point discriminant analysis. In J. Blasius & M. Greenacre (Eds.), *Visualization of Categorical Data* (pp. 441–459). Academic Press.
Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika, 52*(3), 371–392.
Takane, Y., van der Heijden, P. G., and Browne, M. W. (2003). On likelihood ratio tests for dimensionality selection. In Higuchi, T., Iba, Y., and Ishiguro, M., editors, *Proceedings of science of modeling: The 30th anniversary meeting of the information criterion (AIC)*, pages 348–349. The Institute of Statistical Mathematics Tokyo.