Supervised and unsupervised mapping of binary variables: a proximity perspective

Rooij M.J. de; Woestenburg, D.H.A.; Busing, F.M.T.A.

Behaviormetrika
https://doi.org/10.1007/s41237-024-00248-z

INVITED PAPER

Supervised and unsupervised mapping of binary variables: a proximity perspective

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

Received: 4 March 2024 / Accepted: 25 November 2024

Abstract

We propose a new mapping tool for supervised and unsupervised analysis of multivariate binary data with multiple items, questions, or response variables. The mapping assumes an underlying proximity response function, where participants can have multiple reasons to disagree or say "no" to a question. The probability to endorse, or to agree with an item depends on an item specific parameter and the distance in a joint space between a point representing the item and a point representing the participant. The item specific parameter defines a circle in the joint space around the location of the item such that for participants positioned within the circle the endorsement probability is larger than 0.5. For map estimation, we develop and test an MM-algorithm in which the negative log-likelihood function is majorized with a weighted least squares function. The weighted least squares function can be minimized with standard algorithms for multidimensional unfolding. To illustrate the new mapping, two empirical data sets are analyzed. The mappings are interpreted in detail and the unsupervised map is compared to a visualization based on correspondence analysis. In a Monte Carlo study, we test the performance of the algorithm in terms of recovery of population parameters and conclude that this recovery is adequate. A second Monte Carlo study investigates the predictive performance of the new mapping compared to a similar mapping with a monotone response function.

Keywords Bernoulli Variables · Euclidean Distance · Proximity items · Single-peaked · Visualization

Communicated by Kohei Adachi.

✉ Mark de Rooij
rooijm@fsw.leidenuniv.nl

Dion Woestenburg
d.h.a.woestenburg@fsw.leidenuniv.nl

Frank Busing
busing@fsw.leidenuniv.nl

[1] Methodology and Statistics department, Institute of Psychology, Leiden University, PO Box 9555, 2300 RB Leiden, The Netherlands

Published online: 21 January 2025

1 Introduction

Multivariate binary data are often collected in different fields of research. In such investigations, for a set of participants, dichotomous responses are collected on a set of response variables or items. In different sciences, mapping might be considered an important tool. For example, researchers in political science are interested in vote intentions for an upcoming election. In the Dutch parliamentary election studies (Irwin et al. 2003), for example, participants are asked to indicate which political parties were still under consideration for a vote a few months prior to the election. In psychiatry, researchers are interested in mental disorders, such as depression and anxiety and comorbidity of such 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 (NESDA, Penninx et al. 2008) data were collected on a large number of participants on different mental disorders (Spinhoven et al. 2009). In health sciences, researchers are interested in drug consumption profiles, that is, which people use the same set of drugs. Fehrman et al. (2017) are interested in subjects' drug consumption profiles and collected data about the consumption of 18 different drugs. As a final example, Sugiyama (1975) collected data from 4243 Japanese persons, where each participant had to pick any of six different religious practices, leading to binary variables on these six items.

In some studies, besides these binary response variables also characteristics of the persons are available. In the NESDA study and in Fehrman et al. (2017), personality characteristics of the participants are available and the researchers are interested in how these personality characteristics influence mental disorders or drug use. In the Dutch parliamentary election studies, opinions on several topics are available for the participants and researchers are interested in the link between the opinions and the vote intentions. Without such predictor variables the analysis is *unsupervised* while the analysis is *supervised* with such predictors.

For the analysis of binary data, it is important to distinguish between two types of response processes (Thurstone and Chave 1929; Coombs 1964; Polak 2011). In a unipolar scale or map, item responses are monotonically related to the position of the person on the map. The items are so-called *dominance items*. Mathematical problems are a typical example of dominance items where subjects with a higher mathematical ability have a higher probability of getting the problem right. For dominance items the subjects are partitioned into two homogeneous group, that is, both the group of subjects who answer the item correct (1) and the group who answer the item wrong (0) constitute homogeneous groups.

In a bipolar scale or map the item responses are characterized by the proximity between the item and the respondent: The item responses are functions of the distance between the position of an item and the position of a person. The items are so-called *proximity items*. For proximity items, only the respondents who answer yes (or 1) form a homogeneous group. The respondents who answer no, might do so because of a diverse set of reasons.

In classical multivariate analysis, principal component analysis (PCA, Pearson 1901; Hotelling 1936; Jolliffe 2002) is the standard tool for the analysis of dominance items whereas multidimensional unfolding (MDU, Coombs 1964; Heiser 1981; Busing 2010) and Correspondence Analysis (CA, Heiser 1981; Ter Braak 1985; Polak et al. 2009) are the standard tools for the analysis of proximity processes. PCA, MDU, and CA provide low-dimensional geometric mappings of the data.

For binary data a logistic framework is most natural, for both supervised and unsupervised analysis. For the analysis of a single binary response variable and a set of predictors logistic regression is preferred over linear regression. Let us define the probability that person i answers yes (1) for response variable r by $\pi_{ir} = P(Y_{ir} = 1)$, where Y_{ir} is the response variable. In logistic models, these probabilities are defined by the function

$$\pi_{ir} = \frac{\exp(\theta_{ir})}{1 + \exp(\theta_{ir})} = \frac{1}{1 + \exp(-\theta_{ir})},$$

where θ_{ir} is the canonical log-odds form or, in generalized linear model terms, the linear predictor.

For unsupervised analysis of a binary data matrix within the logistic framework, De Leeuw (2006) developed logistic PCA, where θ_{ir} is defined in geometric terms as $\theta_{ir} = m_r + \langle \mathbf{u}_i, \mathbf{v}_r \rangle$ with $\langle \cdot, \cdot \rangle$ the inner product of the *person scores* ($\mathbf{u}_i$) and *factor loadings* ($\mathbf{v}_r$) and m_r an offset for item r. This geometric representation gives a dominance perspective on the data. De Leeuw (2006), generalizing earlier work of Groenen et al. (2003), developed a majorization-minimization (MM) algorithm that transforms the likelihood problem into an iterative least squares problem, such that in every iteration a PCA of *working responses* has to be solved, which can be computed using the singular value decomposition.

Instead of the dominance perspective (i.e., the inner product representation) we can also use a proximity perspective with a distance representation. In that case, we represent the observations and response variables with points in an S-dimensional Euclidean space. The coordinates of the person points are given by the vectors $\mathbf{u}_i$ and those of the items by $\mathbf{v}_r$. The Euclidean distance between these two points ($d(\mathbf{u}_i, \mathbf{v}_r)$) represents, in an inverse manner, the probability for a "yes" or 1 on response variable r for observation i: the smaller the distance the larger the probability. This two-mode distance function is central in Multidimensional Unfolding, a generalization of Multidimensional Scaling to rectangular proximity data and often used in the analysis of preference data.

When predictors ($\mathbf{x}_i$) are available we can perform a supervised analysis and the vector $\mathbf{u}_i$ can be constrained to be a function of these predictors, that is, $\mathbf{u}_i = \mathbf{B}'\mathbf{x}_i$ with $\mathbf{B}$ a matrix of regression coefficients. In the context of dominance items this leads to a logistic reduced-rank regression model (Yee and Hastie 2003; De Rooij 2023). De Rooij (2023) generalized the MM-algorithm of De Leeuw for this supervised case, where again in each iteration of the MM-algorithm a least squares problem needs to be solved, that is, updates are given by a generalized singular value decomposition of *working responses*. For proximity items, we can apply a similar

constraint. The person points are constrained to be linear functions of the predictor variables. Instead of multidimensional unfolding, we need restricted multidimensional unfolding to incorporate these constraints (Busing et al. 2010).

In this paper, we will develop and investigate a logistic mapping based on this proximity perspective using two-mode distances, with and without restrictions. Specifically, we define $\theta_{ir} = m_r - d(\mathbf{u}_i, \mathbf{v}_r)$, where either the vectors $\mathbf{u}_i$ are estimated freely (unsupervised) or constrained to be functions of the predictors $\mathbf{x}_i$ (supervised). We will show that such a mapping allows for a larger number of response profiles compared to a mapping based on the dominance perspective. We develop an MM algorithm for estimation of the map.

This paper is organized as follows. In the next Section, we show our mapping in detail and develop an MM-algorithm for computing the map. We also discuss model selection and assessment issues. In Sect. 3, we analyze Sugiyama's religious practices data and the Dutch election data. For the religious data, we compare our unsupervised solution with a correspondence analysis solution as shown in Heiser (1981). For the Dutch election data we first show an unsupervised analysis followed by a supervised analysis. In Sect. 4, we discuss Monte Carlo studies investigating the parameter recovery of the unsupervised and supervised algorithms and the predictive performance of the supervised mapping. We conclude the paper with a discussion.

2 The mapping

2.1 Multivariate binary data

We have a data set $\{\mathbf{y}_i\}_{i=1}^n$ where $\mathbf{y}_i \in \{0, 1\}^R$, with R the number of response variables. Based on these observations, we define $\mathbf{q}_i = 2\mathbf{y}_i - 1$, such that q_{ir} is either 1 or -1. When characteristics of the participants are available, we collect them in $\{\mathbf{x}_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^P$ where P denotes the number of predictor variables.

2.2 Probabilities

We impose a geometric structure on the probabilities $\pi_{ir} = Pr(Y_{ir} = 1)$ for $r = 1, \ldots, R$ as

$$\pi_{ir} = \frac{\exp(m_r - d(\mathbf{u}_i, \mathbf{v}_r))}{1 + \exp(m_r - d(\mathbf{u}_i, \mathbf{v}_r))} = \frac{1}{1 + \exp(d(\mathbf{u}_i, \mathbf{v}_r) - m_r)},$$

where $d(\mathbf{u}_i, \mathbf{v}_r)$ is the two-mode Euclidean distance

$$d(\mathbf{u}_i, \mathbf{v}_r) = \sqrt{\sum_{s=1}^{S} (u_{is} - v_{rs})^2}$$

in pre-chosen dimensionality S. The S-vectors $\mathbf{u}_i$ and $\mathbf{v}_r$ denote the coordinates of points representing the persons and items in the Euclidean space. The parameter m_r is related to the response variables and determines the maximum height of the probabilities, that is, when $m_r = 0$, for example, the probabilities cannot exceed 0.5.

It is instructive to see how the m_r parameter influences the probabilities. Therefore, Fig. 1 shows the probabilities for a joint unidimensional space with a single item located at the origin ($v = 0$) and persons on the real line ($u \in [-3, 3]$) for different values of m_r, that is 0 (dotted black), 1 (dashed blue), 2 (solid green), and 3 (dashed-dotted red). We see that the probabilities peak at the position of the item, that is, when the distance between item and subject equals zero. Larger m_r lead to higher probabilities, but also to wider *regions of endorsement*, that are, regions in the joint space where the probabilities for participants are larger than 0.5. More precisely, when $d(\mathbf{u}_i, \mathbf{v}_r) = m_r$ the probability equals 0.5. Consequently, when $m_r < 0$ all probabilities are smaller than 0.5. When $m_r > 0$ there are two points on the joint scale where the probability equals 0.5, one at a distance of m_r from the item to the left and one at the same distance to the right. From Fig. 1 it also becomes clear that the group of participants that does not endorse the item is heterogeneous, that is, persons at both extremes have low probabilities.

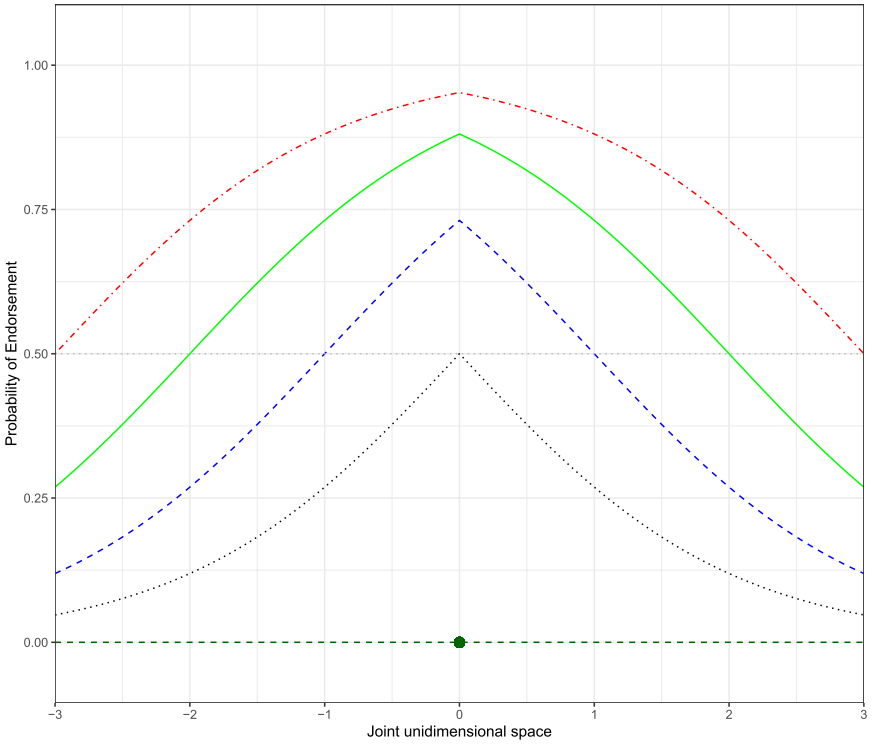


Fig. 1 Probability of endorsement for subjects on the real line and an item located at 0, for different values of m_r. The dotted black curve is for $m_r = 0$, dashed blue for $m_r = 1$, solid green for $m_r = 2$ and dashed-dotted red for $m_r = 3$

In two-dimensional solutions, the regions of endorsement become circles centered at the item point and with radius equal to m_r. When a person point falls within such a circle the probability of endorsement is larger than 0.5 and we classify the subject for that response variable in the class responding 1, whereas if the person point falls outside the circle we classify it in the class responding with 0. Persons may fall outside the circle in any direction, showing the heterogeneity of these participants. An example configuration is shown in Fig. 2, where we show four response variables or items ($R = 4$). Each item is represented by a point and a circle. The circle indicates the region of endorsement, that is, a region in the two-dimensional joint plane such that for persons who are positioned inside the circle the probability is larger than a half, whereas for persons positioned outside the circle the probability is smaller than a half. The probability of endorsement becomes smaller the further away the person lies from the circle (or item point).

In Fig. 2, the four circles partition the two-dimensional space into 14 regions, where each region corresponds to a predicted response profile, that is, a vector of R zeros and ones indicating which items are endorsed. Thirteen of these regions fall within one or more circles, the fourteenth regions falls outside all circles and corresponds to the profile 0000. Note that the point for this profile can be anywhere outside the circles. With four items the maximum number of response profiles equals 16, therefore 2 possible response profiles are not represented, which in this case are the profiles where A and D are endorsed without endorsing B and C (1001) and vice versa (0110). More generally, given our proximity model, the maximum number of regions for R items in an S-dimensional space is

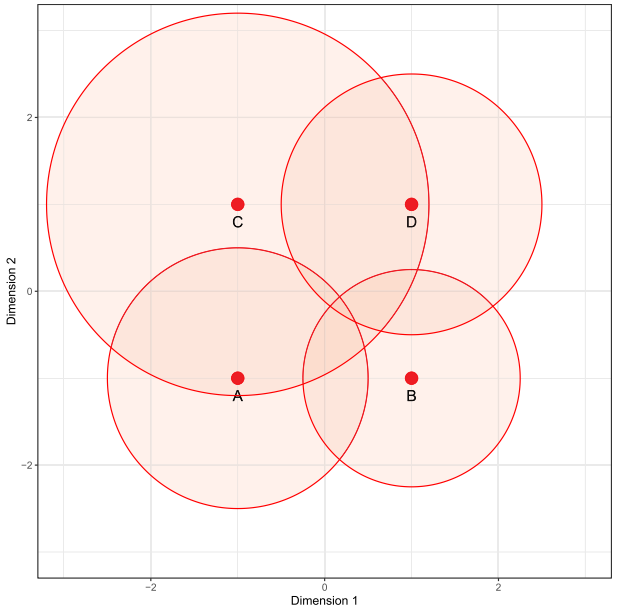


Fig. 2 A two-dimensional mapping with four items or response variables. The four points indicate the positions of the items, while the circles indicate the regions of endorsement with $P(Y_{ir}) \geq 0.5$, that are regions where participants have a probability larger than 0.5 to endorse the specific item

$$\binom{R-1}{S}+\sum_{s=0}^{S}\binom{R}{s}$$

(Sloane 2003; Yaglom and Yaglom 1987). A perfect representation can be found in dimensionality $S = R - 1$. In lower dimensional spaces, not all possible response profiles are represented. Note that in logistic principal component analysis the maximum number of regions for R items in an S-dimensional space is $\sum_{s=0}^{S}\binom{R}{s}$ (Coombs and Kao 1955; De Leeuw 2006), therefore our representation will generally give a better fit.

When predictor variables are available, we constrain the persons points to be functions of these predictors. In this paper, we only consider linear additive functions, that is, $\mathbf{u}_i = \mathbf{B}'\mathbf{x}_i$, but more general nonlinear functions could be considered. In the mapping, these predictor variables will be represented by variable axes with markers. Geometrically, person points can be obtained from these variable axes by the vector sum method also called *interpolation* (Gower and Hand 1996).

2.3 Related approaches

DeSarbo and Hoffman (1986, 1987) defined the log-odds form as $\theta_{ir} = m_i - d^2(\mathbf{u}_i, \mathbf{v}_r)$ with $d^2(\cdot,\cdot)$ the squared two-mode Euclidean distance. The use of squared distances changes the geometry and interpretation of the model. Whereas in our mapping the offsets (m_r) are related to response variables or items, DeSarbo and Hoffman consider the offsets to be a person characteristic (m_i), that is, some persons (observations) have a large *area of acceptance* whereas others have a small area of acceptance. In geometric terms, every person has a circle (with radius $\sqrt{m_i}$) around their position. When an item falls within this circle of a person the probability of a 1 response is higher than 0.5, when it falls outside the circle the probability is smaller than 0.5. Usually, the number of observations n is much larger than the number of response variables or items. Therefore, the number of parameters of the mapping of DeSarbo and Hoffman (1986, 1987) is usually much larger compared to our mapping for the unsupervised case.

DeSarbo and Hoffman (1986, 1987) also allow for external variables. Like in our approach they constrain $\mathbf{u}_i = \mathbf{B}'\mathbf{x}_i$ in their supervised approach. DeSarbo and Hoffman (1986, 1987) developed and described a conjugate gradient algorithm for estimation of the model parameters. To the best of our knowledge, no software is available anymore to estimate this map.

Takane (1998) proposed another closely related model, called MAXSC, where the probabilities are defined as $\pi_{ir} = \exp(\theta_{ir})/(a + \exp(\theta_{ir}))$ with $\theta_{ir} = -d^2(\mathbf{u}_i, \mathbf{v}_r)$, also using squared distances. Defining $m^* = \log(a)$ we may write Takane's proposal as

$$\pi_{ir} = \frac{\exp(-d^2(\mathbf{u}_i, \mathbf{v}_r))}{\exp(m^*) + \exp(-d^2(\mathbf{u}_i, \mathbf{v}_r))}.$$

Now dividing numerator and denominator by $\exp(m^*)$ gives

$$\pi_{ir} = \frac{\exp(-d^2(\mathbf{u}_i, \mathbf{v}_r))/\exp(m^*)}{1 + \exp(-d^2(\mathbf{u}_i, \mathbf{v}_r))/\exp(m^*)}$$

and therefore

$$\pi_{ir} = \frac{\exp(m - d^2(\mathbf{u}_i, \mathbf{v}_r))}{1 + \exp(m - d^2(\mathbf{u}_i, \mathbf{v}_r))},$$

where $m = -m^* = -\log(a)$. Except for the distance or squared distance form, this gives a special case of our mapping with $m_r = m$ for all r, but also a special case of the mapping of DeSarbo and Hoffman with $m_i = m$ for all i. The threshold parameter a in Takane's model can be considered either a person or item characteristic. In a geometric representation, the threshold can be either included as a one-sized circle around the person points or around the item points. For interpretation, larger values of a lead to smaller probabilities, whereas lower values of a lead to higher probabilities.

Takane (1998) did not consider external information, so only investigated the unsupervised case. Whereas in our approach as well as that of DeSarbo and Hoffman the $\mathbf{u}_i$ parameters are fixed effects, Takane assumes them to be random effects with a multivariate normal distribution with mean equal to zero and a diagonal covariance matrix. Takane (1998) developed an EM-algorithm to maximize the marginal likelihood, with in the M-step a Fisher's scoring algorithm. To the best of our knowledge, no software is available anymore to estimate this map.

A third technique often used for the analysis of proximity items is correspondence analysis (Heiser 1981; Greenacre 1984; Ter Braak 1985; Polak 2011; Beh and Lombardo 2021). Correspondence Analysis (CA) is an exploratory data analysis tool for tables with non-negative entries. CA decomposes the observed non-negative data as

$$y_{ir} = y_{++}p_{i+}p_{+r}\left[1 + \sum_{s=1}^{S^*} f_{is}\lambda_s g_{rs}\right]$$

subject to the constraints $\sum_i p_{i+}f_{is} = \sum_r p_{+r}g_{rs} = 0$ and $\sum_i p_{i+}f_{is}^2 = \sum_r p_{+r}g_{rs}^2 = 1$ and where S^* is the maximum dimensionality. The next step is to determine an optimal dimensionality S, often by an analysis of explained inertia. The fitted values follow the equation above, but the sum over dimensions is from 1 till the optimum S. The part belonging to higher dimensions become the residuals. Based on this optimal dimensionality a graphical representation is made of the data. There are several options, we focus on the *row principal normalization*. In this normalization, the row categories are plotted as points with coordinates $u_{is} = \lambda_s f_{is}$, and the column categories as vectors with coordinates g_{rs}. In row principal normalization, the Euclidean distances between the row points in the representation approximate differences in centered row profiles, that is, chi-square distances, in the data. The column vectors have a direction and a length. The association with the row categories is reconstructed by projection, and the length indicates how well a column fits the chosen

dimensionality. Other normalizations include the column principal normalization, the symmetric normalization where the singular values are evenly distributed over the row and column points, and the principal normalization (see Greenacre 1984, for details).

The reason we include CA in our discussion is that Ter Braak (1985) showed relationships between correspondence analysis and the following unimodal logistic model

$$\log\left(\frac{\pi_{ir}}{1-\pi_{ir}}\right) = m_r - \frac{1}{2}\sum_s (u_{is} - v_{rs})^2/t_r^2,$$

a model closely related to the squared distance models pointed out above. The t_r parameters are so-called tolerances. Ter Braak concludes that under four conditions, that are, equal tolerances (t_r), equal or independent maxima (m_r), equally spaced or uniformly distributed participants scores (u_{is}) and item scores (v_{rs}), correspondence analysis provides an approximation to the maximum likelihood solution of this model. In his analysis, the focus of approximation is in terms of the participant ($\mathbf{u}_i$) and item points ($\mathbf{v}_r$). How to exactly translate other estimates of correspondence analysis to, for example, the parameters m_r remains unclear. In Sect. 3, we will show a comparison of correspondence analysis with our model using empirical data. Ter Braak (1986) extended CA for supervised analysis, by restricting a set of points to be a linear combination of external variables.

2.4 Data compression for unsupervised analysis

Before we delve into the algorithm, we will reorganize the data to improve the efficiency of our algorithm. For the unsupervised case with R response variables only 2^R response profiles are possible. For Sugiyama's data about religious practices, for example, $R = 6$ so there are 64 possible response profiles. This provides the opportunity to organize the data into a 64 by 6 matrix instead of a 4243 by 6 matrix of individual responses. The 64 profiles should be weighted by their frequency of occurrence. We denote this frequency by n_i, the number of times the specific profile occurs. There is one profile that is uninformative, the one with only zeros. In an unsupervised analysis, that profile needs to be removed from the data before the analysis. We will denote by I the number of response profiles, each weighted by n_i for $i = 1, \ldots, I$. The frequencies will be collected in the vector $\mathbf{n}$. Note that the original data can also be cast in this formulation with $n_i = 1$ for all i and $I = n = \sum_i n_i$.

In the following, we will develop an MM algorithm. In the outer loop of the algorithm, the negative log-likelihood is majorized by a weighted least squares function. In the inner loop, updates of our parameters are computed.

2.5 An MM-algorithm for estimating the mapping

The beauty of MM algorithms is that the idea is quite simple, powerful, and has guaranteed descent. The idea of MM (De Leeuw and Heiser 1977; Groenen 1993; Heiser 1995; Hunter and Lange 2004) 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 support 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 support point. The support point is defined by the current estimates of $\boldsymbol{\theta}$. The two properties define an iterative sequence for a convergent algorithm because by construction

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

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

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

the updated vector or parameters.

2.5.1 Outer loop

Logistic models are often fitted by maximizing the likelihood, or equivalently minimizing the negative log likelihood,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{I} n_i \sum_{r=1}^{R} -\log \frac{1}{1+\exp(-q_{ir}\theta_{ir})} = \sum_{i=1}^{I}\sum_{r=1}^{R} -n_i \log \frac{1}{1+\exp(-q_{ir}\theta_{ir})} \\ &= \sum_{i=1}^{I}\sum_{r=1}^{R} \mathcal{L}_{ir}(\theta_{ir}),\end{aligned}$$

with $q_{ir} = 2y_{ir} - 1$ and $\theta_{ir} = m_r - d(\mathbf{u}_i, \mathbf{v}_r)$, as before. Because majorization is closed under summation, we can focus on a single element, $\mathcal{L}_{ir}(\theta_{ir})$.

The quadratic majorization theorem states that

$$\mathcal{L}_{ir}(\theta_{ir}) \leq \mathcal{L}_{ir}(\vartheta_{ir}) + \mathcal{L}'_{ir}(\vartheta_{ir})(\theta_{ir} - \vartheta_{ir}) + \frac{1}{2}(\theta_{ir} - \vartheta_{ir})H(\theta_{ir} - \vartheta_{ir}),$$

for a support point ϑ_{ir} and for any H that is larger or equal to the second derivative. An upper bound to the second derivative is $H = \frac{n_i}{4}$.

The first derivative of $\mathcal{L}_{ir}(\theta_{ir})$ with respect to θ_{ir} is

$$\begin{aligned}\mathcal{L}'_{ir}(\theta_{ir}) = \frac{\partial \mathcal{L}_{ir}(\theta_{ir})}{\partial \theta_{ir}} &= -n_i(1+\exp(-q_{ir}\theta_{ir}))\frac{-q_{ir}\exp(-q_{ir}\theta_{ir})}{-(1+\exp(-q_{ir}\theta_{ir}))^2} \\ &= -n_i q_{ir}\frac{\exp(-q_{ir}\theta_{ir})}{1+\exp(-q_{ir}\theta_{ir})},\end{aligned}$$

such that

$$\mathcal{L}_{ir}(\theta_{ir}) \leq \mathcal{L}_{ir}(\vartheta_{ir}) - n_i q_{ir} \frac{\exp(-q_{ir}\vartheta_{ir})}{1+\exp(-q_{ir}\vartheta_{ir})}(\theta_{ir}-\vartheta_{ir}) + \frac{n_i}{8}(\theta_{ir}-\vartheta_{ir})(\theta_{ir}-\vartheta_{ir}).$$

Define $\xi_{ir} = q_{ir}\frac{\exp(-q_{ir}\vartheta_{ir})}{1+\exp(-q_{ir}\vartheta_{ir})}$, and work out the majorization function

$$\begin{aligned}\mathcal{L}_{ir}(\theta_{ir}) &\leq \mathcal{L}_{ir}(\vartheta_{ir}) - n_i\xi_{ir}(\theta_{ir}-\vartheta_{ir}) + \frac{n_i}{8}(\theta_{ir}-\vartheta_{ir})(\theta_{ir}-\vartheta_{ir}) \\ &\leq \mathcal{L}_{ir}(\vartheta_{ir}) - n_i\xi_{ir}\theta_{ir} + n_i\xi_{ir}\vartheta_{ir} + \frac{n_i}{8}(\theta_{ir}^2+\vartheta_{ir}^2-2\theta_{ir}\vartheta_{ir}) \\ &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \frac{n_i}{8}\theta_{ir}^2 - n_i\xi_{ir}\theta_{ir} - 2\frac{n_i}{8}\theta_{ir}\vartheta_{ir} + n_i\xi_{ir}\vartheta_{ir} + \frac{n_i}{8}\vartheta_{ir}^2.\end{aligned}$$

Let $\lambda_{ir} = \vartheta_{ir} + 4\xi_{ir}$ to obtain

$$\begin{aligned}\mathcal{L}_{ir}(\theta_{ir}) &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \frac{n_i}{8}\theta_{ir}^2 - 2\frac{n_i}{8}\theta_{ir}\lambda_{ir} + \frac{n_i}{8}\lambda_{ir}^2 - \frac{n_i}{8}\lambda_{ir}^2 + n_i\xi_{ir}\vartheta_{ir} + \frac{n_i}{8}\vartheta_{ir}^2 \\ &\leq \mathcal{L}_{ir}(\vartheta_{ir}) + \frac{n_i}{8}(\theta_{ir}-\lambda_{ir})^2 - \frac{n_i}{8}\lambda_{ir}^2 + n_i\xi_{ir}\vartheta_{ir} + \frac{n_i}{8}\vartheta_{ir}^2.\end{aligned}$$

Let us define $c = \mathcal{L}_{ir}(\vartheta_{ir}) - \frac{n_i}{8}\lambda_{ir}^2 + n_i\xi_{ir}\vartheta_{ir} + \frac{n_i}{8}\vartheta_{ir}^2$, a constant with respect to θ_{ir}, so that we can write

$$\begin{aligned}\mathcal{L}_{ir}(\theta_{ir}) \leq& \frac{1}{8}w_{ir}(\theta_{ir}-\lambda_{ir})^2 + c \\ \leq& \mathcal{M}_{ir}(\theta_{ir}|\vartheta_{ir}) + c.\end{aligned}$$

with $w_{ir} = n_i$ for all r.

As

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{I}\sum_{r=1}^{R}\mathcal{L}_{ir}(\theta_{ir}),$$

we have that

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &\leq \sum_{i=1}^{I}\sum_{r=1}^{R}\frac{1}{8}w_{ir}(\theta_{ir}-\lambda_{ir})^2 + c \\ &\leq \mathcal{M}(\boldsymbol{\theta}|\boldsymbol{\vartheta}) + c,\end{aligned}$$

so that the majorization function is a weighted least squares function with weights equal to $w_{ir} = n_i$ for all r.

2.5.2 Inner loop: minimizing the weighted least squares function

The geometric structure for θ_{ir} equals

$$\theta_{ir} = m_r - d(\mathbf{u}_i, \mathbf{v}_r),$$

such that the parameters of our optimization function are the offsets $\mathbf{m}$, the coordinates $\mathbf{u}_i$ collected in the matrix $\mathbf{U}$ in case of an unsupervised analysis or the regression weights $\mathbf{B}$ for the supervised analyses, and the coordinates $\mathbf{v}_r$ collected in the matrix $\mathbf{V}$. Given current values of these parameters (ϑ), the majorization function becomes

$$\mathcal{M}(\mathbf{m}, \mathbf{U}, \mathbf{V}|\vartheta) = \sum_i \sum_r w_{ir}(\lambda_{ir} - m_r + d(\mathbf{u}_i, \mathbf{v}_r))^2.$$

We will alternate between updating the offsets and the coordinates to find the minimum of our loss function.

Update of m_r

For updating m_r, we consider $\mathbf{U}$ (or $\mathbf{B}$) and $\mathbf{V}$ as fixed. Then we need to minimize

$$\begin{aligned}\mathcal{M}(\mathbf{m}|\mathbf{U}, \mathbf{V}, \vartheta) &= \sum_i \sum_r w_{ir}(\lambda_{ir} - m_r + d(\mathbf{u}_i, \mathbf{v}_r))^2 \\ &= \sum_i \sum_r w_{ir}(t_{ir} - m_r)^2,\end{aligned}$$

where $t_{ir} = \lambda_{ir} + d(\mathbf{u}_i, \mathbf{v}_r)$. The update for m_r is given by the weighted mean of t_{ir} for every r, that is

$$m_r^+ = \frac{\sum_i w_{ir} t_{ir}}{\sum_i w_{ir}}.$$

If we would like to constrain $m_1 = m_2 = \cdots, m_R = m$, the update of m becomes

$$m^+ = \frac{\sum_i \sum_r w_{ir} t_{ir}}{\sum_i \sum_r w_{ir}}.$$

As a side note, we highlight that person specific offsets (m_i), as proposed by DeSarbo and Hoffman (1986, 1987), could be incorporated in our mapping. Estimates can be obtained as

$$m_i^+ = \frac{\sum_r w_{ir} t_{ir}}{\sum_r w_{ir}}.$$

As noted before, this would increase the number of parameters substantially. It is even possible to go one step further and estimate both item and person specific offsets with even more parameters to estimate. These options are not incorporated in our software.

Update geometric parameters: Unsupervised analysis

For updating the coordinate matrices we treat m_r as fixed. Let us rewrite our majorization function as

$$\begin{aligned}\mathcal{M}(\mathbf{U},\mathbf{V}|\mathbf{m},\boldsymbol{\vartheta}) &= \sum_i \sum_r w_{ir}(\lambda_{ir} - m_r + d(\mathbf{u}_i,\mathbf{v}_r))^2 \\ &= \sum_i \sum_r w_{ir}(\delta_{ir} - d(\mathbf{u}_i,\mathbf{v}_r))^2,\end{aligned} \tag{1}$$

where $\delta_{ir} = -(\lambda_{ir} - m_r)$.

This minimization function in Eq. (1) is the usual raw STRESS function often used in multidimensional scaling and unfolding. De Leeuw (1977) and De Leeuw and Heiser (1977) proposed the SMACOF algorithm for minimization of this STRESS function for multidimensional scaling. The SMACOF algorithm is itself an MM algorithm. Convergence properties of this algorithm are described by De Leeuw (1988). Heiser (1981, 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. Advances in the algorithm 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, the cross product term of the dissimilarities δ_{ir} with the distances $d(\mathbf{u}_i,\mathbf{v}_r)$ is majorized using the Cauchy-Schwarz inequality by a linear function. Heiser (1987) defined preliminary updates based on the current $\mathbf{U}$ and $\mathbf{V}$ as

$$\begin{aligned}\mathbb{U} &= \mathbf{P}\mathbf{U} - \mathbf{A}\mathbf{V} \\ \mathbb{V} &= \mathbf{Q}\mathbf{V} - \mathbf{A}'\mathbf{U}\end{aligned}$$

where the matrix $\mathbf{A}$ has elements

$$a_{ir} = \begin{cases} w_{ir}\delta_{ir}/d(\mathbf{u}_i,\mathbf{v}_r), & \text{if } d(\mathbf{u}_i,\mathbf{v}_r) > 0, \\ 0, & \text{if } d(\mathbf{u}_i,\mathbf{v}_r) = 0 \end{cases} \tag{2}$$

and $\mathbf{P} = \text{diag}(\mathbf{A1})$, and $\mathbf{Q} = \text{diag}(\mathbf{1}'\mathbf{A})$. Collecting the weights w_{ir} in the matrix $\mathbf{W}$ and defining the diagonal matrices $\mathbf{R} = \text{diag}(\mathbf{W1})$ and $\mathbf{C} = \text{diag}(\mathbf{1}'\mathbf{W})$, the majorization function for the raw STRESS function (Eq. 1) is given by (Heiser 1987; Busing 2010)

$$\begin{aligned}\mathcal{M}(\mathbf{U},\mathbf{V}|\mathbf{m},\boldsymbol{\vartheta}) \leq \sigma^2(\mathbf{U},\mathbf{V}) = c + \text{tr}\big(\mathbf{U}'\mathbf{R}\mathbf{U}\big) + \text{tr}\big(\mathbf{V}'\mathbf{C}\mathbf{V}\big) \\ - 2\text{tr}\big(\mathbf{U}'\mathbf{W}\mathbf{V}\big) - 2\text{tr}\big(\mathbf{U}'\mathbb{U}\big) - 2\text{tr}\big(\mathbf{V}'\mathbb{V}\big).\end{aligned} \tag{3}$$

Alternating between updates for $\mathbf{U}$ and $\mathbf{V}$ provides the minimum for (3): keeping $\mathbf{V}$ fixed, the update for $\mathbf{U}$ is given as

$$\mathbf{U} = \mathbf{R}^{-1}(\mathbb{U} + \mathbf{W}\mathbf{V})$$

and keeping $\mathbf{U}$ fixed, the update for $\mathbf{V}$ is given as

$$\mathbf{V} = \mathbf{C}^{-1}\big(\mathbb{V} + \mathbf{W}'\mathbf{U}\big).$$

This standard SMACOF algorithm, as just described, was derived under the assumption that the (working) dissimilarities are non-negative. However, in our case we

cannot guarantee that this assumption is true in every cycle of the algorithm. Heiser (1991) showed a way to deal with negative dissimilarities in multidimensional scaling. We will generalize that approach to the two-mode distance case.

The line of thought of Heiser's contribution is that two majorizing functions are defined: one for the case that the dissimilarity is non-negative and one for the case that the dissimilarity is negative. When the dissimilarities are non-negative, we can majorize, as described above, by a linear function. When the dissimilarities are negative, the STRESS function can be majorized by a quadratic function. Heiser (1991) showed that the updating formulae are still valid but that some elements of the matrices $\mathbf{W}$ and $\mathbf{A}$ need to be defined differently, depending on the sign of the dissimilarity. Matrix $\mathbf{W} = \{w_{ir}\}$ is redefined as

$$w_{ir} = \begin{cases} w_{ir} & \text{if } \delta_{ir} \geq 0, \\ \left[w_{ir}(d(\mathbf{u}_i, \mathbf{v}_r) + |\delta_{ir}|)\right]/d(\mathbf{u}_i, \mathbf{v}_r) & \text{if } \delta_{ir} < 0 \text{ and } d(\mathbf{u}_i, \mathbf{v}_r) > 0, \\ \left[w_{ir}(\epsilon + \delta_{ir}^2)\right]/\epsilon & \text{if } \delta_{ir} < 0 \text{ and } d(\mathbf{u}_i, \mathbf{v}_r) = 0, \end{cases}$$

where ϵ is a small positive constant. The elements of $\mathbf{A}$ depend on the sign of the working dissimilarities, that is, for negative δ_{ir} the corresponding a_{ir} are set to zero, while for non-negative δ_{ir} the a_{ir} are defined as in Eq. (2).

Within the inner loop we could iterate till convergence, but a few iterations of updating $\mathbf{U}$ and $\mathbf{V}$ are enough. In other words, instead of finding the minimum of the majorization function in every iteration we only need to take a few steps in the right direction. This iterative sequence still guarantees convergence to a (local) minimum.

Note that, the unsupervised mapping has rotational and translational freedom. To obtain identified solutions, we require $\sum_i n_i u_{is} = 0$ for every dimension s to curb the translational freedom. To deal with the rotational freedom, we rotate the solution such that $\mathbf{U}$ is in principal coordinates. Therefore, we define the diagonal matrix $\mathbf{D}_n$ with elements $d_{ii} = n_i$ and perform an eigenvalue decomposition $\mathbf{U}'\mathbf{D}_n\mathbf{U} = \mathbf{E}\mathbf{\Phi}\mathbf{E}'$ and rotate (i.e., post multiply) both $\mathbf{U}$ and $\mathbf{V}$ by $\mathbf{E}$.

Update geometric parameters: Supervised analysis

De Leeuw and Heiser (1980) considered multidimensional scaling with restrictions on the configuration, like we use in the supervised mapping. Heiser (1987) described how to use linear constraints in multidimensional unfolding, which was further developed in Busing et al. (2010).

In the supervised analysis, we have that $\mathbf{U} = \mathbf{X}\mathbf{B}$ and we need to estimate $\mathbf{B}$ instead of $\mathbf{U}$. Given the other parameters, the update of $\mathbf{B}$ is given as

$$\mathbf{B} = \left(\mathbf{X}'\mathbf{R}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbb{U} + \mathbf{X}'\mathbf{W}\mathbf{V}\right)$$

and before updating $\mathbf{V}$ we compute $\mathbf{U} = \mathbf{X}\mathbf{B}$.

For the supervised mapping there is only rotational freedom because the origin corresponds to $\mathbf{x} = \mathbf{0}$. Like in the unsupervised analysis we rotate the solution such that $\mathbf{U}$ is in principal coordinates. We find the rotation matrix $\mathbf{E}$ as in the unsupervised case and rotate both $\mathbf{B}$ and $\mathbf{V}$ by $\mathbf{E}$.

2.6 Local optima and initialisation

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 good (or rational) initial parameter values or by using many random starts. Good rational starting values can be obtained by performing, for example, a (canonical) correspondence analysis. However, even these good starts do not guarantee to find the global minimum. Therefore, it is recommended to also perform a number of random starts. In our implementation, we draw elements of the parameter matrices $\mathbf{U}$ or $\mathbf{B}$ and $\mathbf{V}$ from independent standard normal distributions as initial values. For the offset parameter, we use the average of the q_{ir} as initial values for m_r.

2.7 Algorithm scheme and implementation

An overview of the algorithms for the unsupervised and supervised mappings can be found in Algorithm 1 and Algorithm 2, respectively. Note that the matrix $\mathbf{\Pi}$ has elements π_{ir}. We implemented the algorithm in the R-package `lmap` (De Rooij et al. 2024). For the estimation of the map, the function `lmdu` can be used. The package also has a function for plotting the map.

Algorithm 1 Unsupervised Algorithm

Input: $\mathbf{Y}, \mathbf{n}, \mathbf{m}, \mathbf{U}, \mathbf{V}, S$
predefine: maxouter, maxinner, ϵ_1, ϵ_2
assess $\mathcal{L}^0(\mathbf{m}, \mathbf{U}, \mathbf{V})$
for $t_1 \leftarrow 1$, maxouter **do**
  compute $\mathbf{\Pi}$
  compute $\mathbf{\Lambda} \leftarrow \mathbf{1}\mathbf{m}' - d(\mathbf{U}, \mathbf{V}) + 4(\mathbf{Y} - \mathbf{\Pi})$
  compute $\mathbf{T} \leftarrow \mathbf{\Lambda} + d(\mathbf{U}, \mathbf{V})$
  compute $\mathbf{m} \leftarrow (\mathbf{W} \odot \mathbf{T})'\mathbf{1}/n$
  assess $\mathcal{M}^0(\mathbf{m}, \mathbf{U}, \mathbf{V}|\boldsymbol{\vartheta})$
  for $t_2 \leftarrow 1$, maxinner **do**
    compute $\mathbf{A}$ and $\mathbf{P}, \mathbf{Q}$
    compute $\mathbf{W}$ and $\mathbf{R}, \mathbf{C}$
    compute $\mathbf{U} \leftarrow \mathbf{R}^{-1}(\mathbf{P}\mathbf{U} - \mathbf{A}\mathbf{V} + \mathbf{W}\mathbf{V})$
    compute $\mathbf{V} \leftarrow \mathbf{C}^{-1}(\mathbf{Q}\mathbf{V} - \mathbf{A}'\mathbf{U} + \mathbf{W}'\mathbf{U})$
    assess $\mathcal{M}^{t_2}(\mathbf{m}, \mathbf{U}, \mathbf{V}|\boldsymbol{\vartheta})$
    if $\mathcal{M}^{t_2}(\mathbf{m}, \mathbf{U}, \mathbf{V}|\boldsymbol{\vartheta}) - \mathcal{M}^{t_2-1}(\mathbf{m}, \mathbf{U}, \mathbf{V}|\boldsymbol{\vartheta}) < \epsilon_1$: break
  end for
  assess $\mathcal{L}^{t_1}(\mathbf{m}, \mathbf{U}, \mathbf{V})$
  if $\mathcal{L}^{t_1}(\mathbf{m}, \mathbf{U}, \mathbf{V}) - \mathcal{L}^{t_1-1}(\mathbf{m}, \mathbf{U}, \mathbf{V}) < \epsilon_2$: break
end for
eigenvalue decomposition $\mathbf{U}'\mathbf{D}_n\mathbf{U}$: $\mathbf{E}\mathbf{\Phi}\mathbf{E}'$
rotate $\mathbf{U} \leftarrow \mathbf{U}\mathbf{E}$ and $\mathbf{V} \leftarrow \mathbf{V}\mathbf{E}$
return($\mathbf{m}, \mathbf{U}, \mathbf{V}$)

Algorithm 2 Supervised Algorithm

Input: $\mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}, \mathbf{B}, \mathbf{V}, S$
predefine: maxouter, maxinner, ϵ_1, ϵ_2
assess $\mathcal{L}^0(\mathbf{m}, \mathbf{B}, \mathbf{V})$
for $t_1 \leftarrow 1$, maxouter **do**
  compute $\mathbf{\Pi}$
  compute $\mathbf{\Lambda} \leftarrow \mathbf{1m}' - d(\mathbf{XB}, \mathbf{V}) + 4(\mathbf{Y} - \mathbf{\Pi})$
  compute $\mathbf{T} \leftarrow \mathbf{\Lambda} + d(\mathbf{XB}, \mathbf{V})$
  compute $\mathbf{m} \leftarrow (\mathbf{W} \odot \mathbf{T})'\mathbf{1}/n$
  assess $\mathcal{M}^0(\mathbf{m}, \mathbf{XB}, \mathbf{V}|\boldsymbol{\vartheta})$
  for $t_2 \leftarrow 1$, maxinner **do**
    compute $\mathbf{A}$ and $\mathbf{P}, \mathbf{Q}$
    compute $\mathbf{W}$ and $\mathbf{R}, \mathbf{C}$
    compute $\mathbf{B} \leftarrow (\mathbf{X}'\mathbf{RX})^{-1}\left[\mathbf{X}'\left(\mathbf{PXB} - \mathbf{AV}\right) + \mathbf{X}'\mathbf{WV}\right]$
    compute $\mathbf{V} \leftarrow \mathbf{C}^{-1}\left(\mathbf{QV} - \mathbf{A}'\mathbf{XB} + \mathbf{W}'\mathbf{XB}\right)$
    assess $\mathcal{M}^{t_2}(\mathbf{m}, \mathbf{XB}, \mathbf{V}|\boldsymbol{\vartheta})$
    if $\mathcal{M}^{t_2}(\mathbf{m}, \mathbf{XB}, \mathbf{V}|\boldsymbol{\vartheta}) - \mathcal{M}^{t_2-1}(\mathbf{m}, \mathbf{XB}, \mathbf{V}|\boldsymbol{\vartheta}) < \epsilon_1$: break
  end for
  assess $\mathcal{L}^{t_1}(\mathbf{m}, \mathbf{B}, \mathbf{V})$
  if $\mathcal{L}^{t_1}(\mathbf{m}, \mathbf{B}, \mathbf{V}) - \mathcal{L}^{t_1-1}(\mathbf{m}, \mathbf{B}, \mathbf{V}) < \epsilon_2$: break
end for
eigenvalue decomposition $\mathbf{B}'\mathbf{X}'\mathbf{D}_n\mathbf{XB}$: $\mathbf{E\Phi E}'$
rotate $\mathbf{B} \leftarrow \mathbf{BE}$ and $\mathbf{V} \leftarrow \mathbf{VE}$
return$(\mathbf{m}, \mathbf{B}, \mathbf{V})$

2.8 Model selection and assessment

For the application of our mapping (both the unsupervised as well as the supervised) to empirical data, the user has to define or choose a dimensionality S. Although we have a likelihood method, likelihood ratio statistics cannot be used for dimensionality selection (Takane et al. 2003), because a regularity condition for this statistic to be chi-square distributed is not satisfied. To find an optimal dimensionality, we will use the AIC. AIC is based on the entropic or information-theoretic interpretation of the maximum likelihood method as well as the minimization of the Kullback-Leibler information quantity (Akaike 1974; Burnham and Anderson 2004; Anderson 2007). The AIC for any model can be defined as

$$\text{AIC} = 2\hat{\mathcal{L}}(\boldsymbol{\theta}) + 2\text{npar},$$

where npar denotes the number of parameters of the model. The first term $2\hat{\mathcal{L}}(\boldsymbol{\theta})$ in AIC is twice the negative log likelihood (usually called the deviance) and it acts as a measure of lack of fit to the data, for which smaller values will be preferred. The second term, 2npar, acts as a penalty term which penalizes complex models for having many parameters. The aim is to reach a balance between the lack of fit and the model complexity: models with smaller AIC values will indicate a better balance. The optimal model choice minimizes the AIC. For the AIC, the number of parameters is needed. For the unsupervised case the number of

parameters is npar $= (I-1)S + RS + R - S(S-1)/2$, while for supervised analysis it is npar $= PS + RS + R - S(S-1)/2$.

For supervised models, there is also the question which predictor variables have an effect on the response variables. Again there is a problem with the likelihood ratio statistic. It compares the fit of two nested models by dividing the likelihood value obtained under a null hypothesis by the likelihood obtained under an alternative hypothesis. If the model under the null hypothesis is true and certain regularity conditions are satisfied, minus two times the log of this ratio is known to be asymptotically distributed as a chi-square variable with degrees of freedom equal to the difference in the number of parameters under the two hypotheses. The assumption that the model under the null hypothesis should be true is problematic in our case. Suppose that we would like to test whether the first predictor has an effect on the responses and therefore estimate the model with and without this first predictor. Even if the first predictor has no effect on the outcomes, the model can be miss-specified in many other aspects. One example is that some of the other predictors have a nonlinear effect on the log odds of some of the response variables. Therefore, we also propose to use the AIC for selection of predictor variables.

For the assessment of model fit, we can evaluate several statistics. First, we can look at different types of residuals. The simplest type of residuals are the *raw residuals*, $e_{ir} = y_{ir} - \hat{\pi}_{ir}$. These residuals are positive for participants with $y_{ir} = 1$ and negative for the others. The range of these raw residuals is from -1 to 1.

Another type of residuals are the *deviance residuals*. These deviance residuals partition the total deviance in small pieces, such that the sum of squared deviance residuals equal the total deviance. These deviance residuals are defined as

$$\mathcal{D}_{ir} = \text{sign}(e_{ir})\sqrt{2\mathcal{L}_{ir}}.$$

These residuals are a by-product of the estimation procedure, as the $\mathcal{L}_{ir}$ are computed in every iteration. From the deviance residual, we can evaluate the contribution of each participants or each item to the overall deviance, by squaring and summing over items or participants, respectively.

In the supervised mapping, we make an assumption about the shape of the relationship between the predictor variables and the response variable. Whereas in linear models the implied functional form is linear, in our mapping the functional form is defined by the distance. To verify whether this functional form is correct, we will generalize component plus residual plots. Such plots have been proposed for linear and logistic regression, see Fox (2015). For our mapping, such a plot is created for each combination of a predictor and a response variable. On the horizontal axis the predictor variable is depicted. On the vertical axis we show the partial fit plus four times the raw residual (e_{ir}). The partial fit is defined as $m_r - d(\boldsymbol{x}_p\boldsymbol{b}'_p, \boldsymbol{v}_r)$. We add four times the raw residual, as this is also used in our algorithm (as shown in Sect. 2.5). We add two lines to this scatterplot: the first line shows the assumed functional form, the second is a smooth loess curve (Cleveland 1979). When the two lines strongly deviate from each other this indicates model misspecification.

One aspect of model assessment is to check for influential cases. To detect the influence of observations on the fit, usually the model is fitted n times, every time leaving one participant out. Indicators of influence are given by the (standardized) change in estimated parameters or fit measures. For linear regression models it has been shown that such statistics can be derived without actually having to refit the model n times. For logistic regression, Pregibon (1981) derived approximations to these indicators. As computational power has increased dramatically over the last decades, we will not develop approximations, but use the computational power to find indicators of influence. The model will be estimated by leaving out participant i. The estimated values when leaving out an observation will be denoted by $\hat{\mathbf{B}}_{-i}$, and similarly for other parameters. We propose to use the following indicators. First the change in deviance

$$\delta_{\mathcal{D}}(i) = 2\big(\mathcal{L}(\boldsymbol{m}, \boldsymbol{B}, \boldsymbol{V}) - \mathcal{L}(\boldsymbol{m}_{-i}, \boldsymbol{B}_{-i}, \boldsymbol{V}_{-i})\big)$$

which is similar to Cooks distance (Williams 1987). Note that the likelihood is evaluated on the complete sample using the estimates obtained when leaving participant i out. Second, the overall change in regression weights is defined by

$$\delta_B(i) = \|\hat{\mathbf{B}} - \hat{\mathbf{B}}_{-i}\|^2,$$

where $\|\boldsymbol{Q}\|^2$ takes the sum of squares of all elements of the matrix $\boldsymbol{Q}$. Finally, the overall change in item locations is defined by the following measure

$$\delta_V(i) = \|\hat{\mathbf{V}} - \hat{\mathbf{V}}_{-i}\|^2.$$

For all three indicators, higher values represent larger influence of observation i.

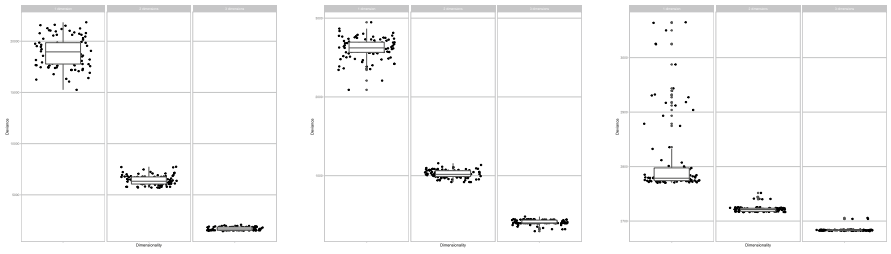
3 Applications

In this section, we show applications on two empirical data sets. The first data set is Sugiyama's data on religious practices, the second data set concerns vote intentions for parliamentary elections in The Netherlands.

We apply our unsupervised mapping to the first data set and compare the results against correspondence analysis. For the vote intention data, we first apply an unsupervised mapping and thereafter include the predictor variables in the supervised mapping. Before we delve into the applications, we discuss for all three analyses the occurrence of local optima. Afterwards, we interpret the solutions for both data sets.

3.1 Severity of local optima

For each of the three analyses, we performed 100 different starts in dimensionalities 1, 2, and 3. The deviances (twice the negative loglikelihood) of these analyses are shown in Fig. 3, for the unsupervised analysis of Sugiyama's data in Fig. 3a, for the unsupervised analysis of the Dutch election data in Fig. 3b. and for the supervised analysis of the Dutch election data in Fig. 3c. Generally, we see that local optima



(a) (b) (c)

Fig. 3 Deviances of 100 random starts in 1, 2, and 3 dimensions for **a** Sugiyama's data, **b** the unsupervised analysis of the Dutch election data, and **c** the supervised analysis of the Dutch election data. A little bit of horizontal jitter is used to better visualize the deviance of the 100 solutions

occur in all dimensionalities, but that especially the unidimensional analyses are prone to local optima. The local optima problem for the supervised analysis seems to be less severe as the median is very close to the minimum.

3.2 Sugiyama's data on religious practices

For Sugiyama's data on religious practices, the respondents had to answer *yes* or *no* to the following 6 questions:

A Do you make it a rule to practice religious conduct, such as attending religious services, religious worship, and missionary works and do you occasionally offer prayers or chant sutras?
B Do you visit a grave once or twice a year?
C Do you occasionally read religious books, such as the Bible or the Buddhist Scriptures?
D Do you visit shrines and temples to pray for business prosperity, success in an entrance examination, and so forth?
E Do you keep a talisman, such as an amulet, charm, or mascot near you?
F Did you draw a fortune, consult a diviner, or had you your fortune told within the last year?

The 64 answer patterns with response frequencies can be found in Heiser (1981) and Takane (1998). Three response patterns do not occur in the data. In total, there were 4243 participants. Because all zeros are uninformative, the pattern with only zeros is left out of the analysis. This reduces the sample size with 718 to 3525.

The deviance of the intercepts only model (i.e., $\theta_{ir} = m_r$) equals 24,011.13

The smallest deviance in one dimension is 15,254.3, in two dimensions 5690.7, and in three dimensions 1451.8, so that 36.5, 76.3 and 93.9% of the deviance is explained by these three mappings, respectively. The AICs are 22328.3, 19824.7, and 22643.8 for the one, two, and three dimensional solution, respectively. The maximum number of represented profiles are 12, 32, and 52 for the uni-, two-, and

three-dimensional solution, respectively. We will further look at the two-dimensional solution.

The two dimensional solution is displayed in the left panel of Fig. 4, where each of the six items is represented by a point with coordinate $\mathbf{v}_r$ and a circle with radius m_r. Persons with locations inside the circle have a probability larger than 0.5 for the corresponding item. We see that item C has the smallest circle, that is having a relatively small area of endorsement, while item B has the largest circle. We shaded the regions of endorsement in such a way that when two (or more) circles overlap the shading becomes darker.

Considering the positions of the items, we see that items B, E, and D lie close together, as well as A and C, while item F is a bit isolated. We see that the circles for items C and F only slightly overlap, indicating that these two items are seldom picked together. Circles for items E and D, however, do overlap considerably indicating that these two items are often picked together. The person points are labeled by the response pattern. For example, the point labelled as 110010 represent participants that picked items A, B, and E while not picking items B, C, and F. We see that the response patterns with a single 1 fall on the outside of the joint space, within a single circle. For example, at the top there are the response patterns for only A (100000) and only C (001000), the points fall within the regions of endorsement of the corresponding items. In the middle of these two points is the response pattern for both A and C (101000), falling in the region of endorsement of both items. In this representation 23 profiles can be seen clearly, where many other profiles clutter together. These 23 profiles represent 2811 participants, that is 79.7% of the respondents. These 23 profiles all fall in the correct regions.

There is a large bulk of subject points near the intersection of all circles, representing mainly answer profiles with 2, 3, and 4 picks. In the right hand panel of Fig. 4, we zoom into this region with many response patterns. Note that many regions, contain multiple profiles points, such that relatively many of these profiles

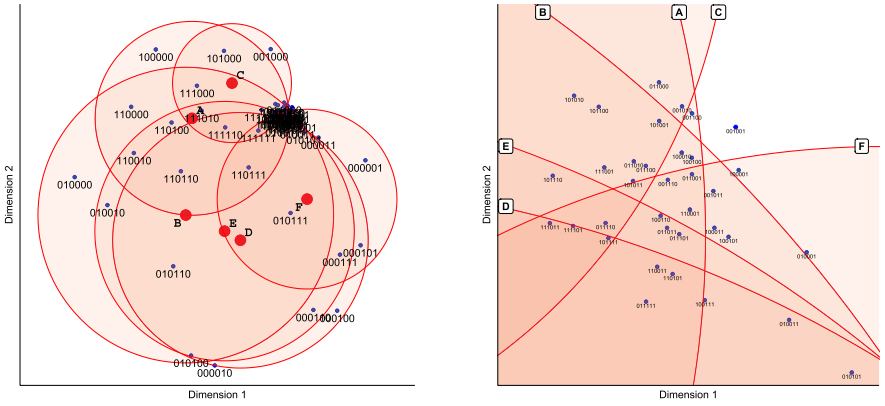


Fig. 4 Two-dimensional solution for Sugiyama's data. The six religious practice items are indicated by the letters A till F. Subject points are labelled with their response patterns. Left panel: the estimated configuration; right panel: a zoom into the clutter of response profiles

are missclassified. These missclassified points, however, only represent a small part of the data (714 participants) and response profiles with small weights. The profile with most participants in this zoomed-in visualization is 010001 (on the right hand side of the figure) and has 137 participants.

Heiser (1981, pp. 142–143) analysed these data using correspondence analysis. In CA, a researcher has to choose between the row-principal, column-principal, or symmetric normalization. Heiser opted for the row-principal normalization, such that the answer patterns are in the centroid of the item points. We reproduced the analysis in the Fig. 5. We focus on the unimodal model approximation (i.e., distance rule) as suggested by Ter Braak (1985), but sometimes also discuss the standard inner product interpretation (i.e., projection rule). The first two dimensions explain 26.4 and 21.6 percent of the inertia, respectively. The column masses are 0.18, 0.35, 0.06, 0.13, 0.17, and 0.11, for items A till F respectively.

The solution is in some aspects similar to ours. The order of the items in the CA solution along the V-shape from upper left, to bottom center, to upper right is FDE-BAC. This same ordering can also be witnessed from our solution (Fig. 4) along a nonlinear curve, although the spacing in our solution is different. Furthermore, in both solutions the items B, E, and D are close together, as well as A and C.

Inspecting the correspondence analysis solution in Fig. 5 in more detail, we can find some interesting things. In the row principal normalization as depicted here, the points for positions of participants are in the center of the item positions chosen by

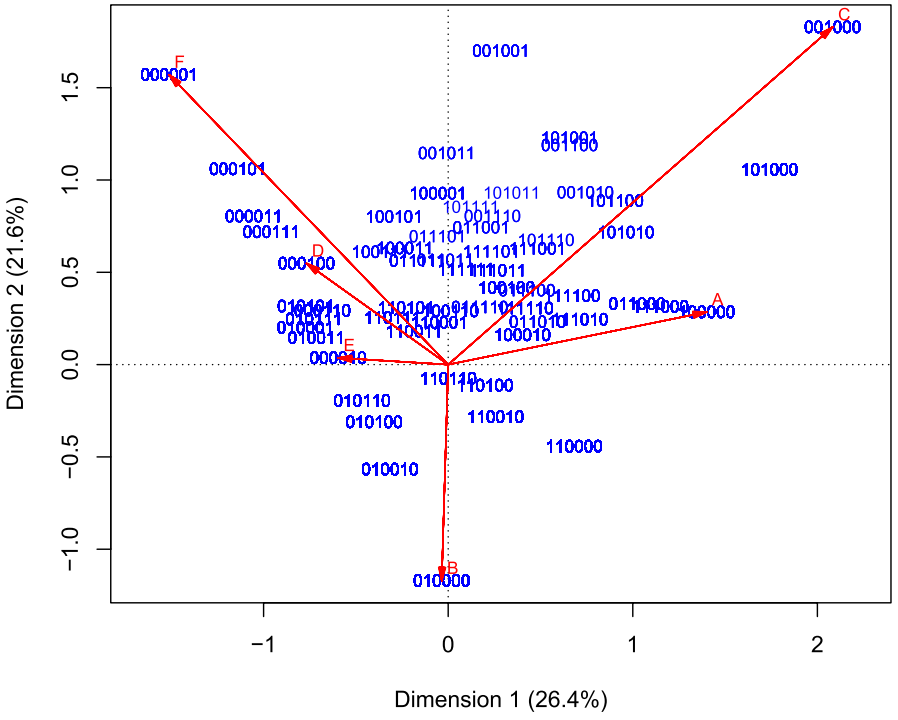


Fig. 5 Two dimensional correspondence analysis solution for religious practices data

the participant. Therefore, participants with response pattern 000001 are exactly on top of the item point for item F (the last one), and participants with response pattern 000101 are exactly in the middle of items D and F. The usual correspondence analysis interpretation uses the projection rule. This projection rule states that we project the participant point onto the vector and we multiply the length of the vector times the distance from the origin towards the projection point. Following this rule, we can conclude that these participants (i.e., those with response pattern 000101) have a higher association with item F than with item D, because the length of vector F is longer. In contrast, following the distance model interpretation the two items are equally distant, which would result in this interpretation of equal association. The expected values for items D and F for those participants, are 0.64 and 0.95. The higher expected value cannot be explained from the distance perspective, as the distances are equal and the mass for item D is higher than the mass for item F.

Inspecting the fitted values of this two dimensional correspondence analysis solution, we find values in the range −0.70 till 1.52, with values below zero and above 1. Therefore, we can not interpret these values as probabilities and the approximation of correspondence analysis to a unimodal logistic model, as discussed in Sect. 2.3, fails in that sense. This result is, of course, similar to the comparison of fitting a usual linear regression model and a logistic regression model to a binary response variable.

What is unclear from the correspondence analysis solution is how to make the classification when we use the distance rule. Whereas, in our distance model it is clear whether the probability of a participant for an item is smaller or larger than 0.5 (i.e., whether a participant point falls within or outside a circle), in the correspondence analysis solution no such regions are available.

We further like to remark that choosing a different normalization in the correspondence analysis visualization, does not alter the fitted values nor the association values using the projection rule, but it does change the distances. So, the interpretation in terms of a distance model alters depending on the normalization chosen. Whereas in CA the user has to specify a normalization, for our mapping, we do not have to make a choice between different normalizations as the between sets distances are optimized. Furthermore, our analysis includes the offsets (m_r) as circles in the display, whereas such effects are usually not displayed in CA. As a consequence, from our display we can immediately see whether the probability of endorsement is larger than 0.5 or not for a given subject point. Such information cannot be retrieved from the CA solution.

3.3 Dutch election data

This data set consists of 352 Dutch inhabitants and their vote intentions for the parliamentary election in 2002 (Irwin et al. 2003). The responses in this data set correspond to vote intentions for 8 different political parties: PvdA (110), CDA (123), VVD (119), D66 (89), GL (114), LN (27), LPF (77), and SP (57). Ninety-three respondents indicated only one party, 169 respondents reported a vote

intention for two parties, 79 for three parties, 8 for four parties, 2 respondents for five parties, and 1 respondent still had 6 parties under consideration.

Furthermore, respondents were asked their opinion on five issues. These opinion data can be used as predictors. On a seven point scale, they had to indicate whether they think that *Euthanasia* (E) should always be forbidden (1) or that a doctor should always be allowed to end a life upon a patientâ€™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 Netherlands 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 towards *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 an 11-point *Left-Right* (LR) scale, where 0 indicates left and 10 right wing. We centered these predictor variables around 4 for the seven points scales and around 5 for the left-right variable.

3.3.1 Unsupervised analysis

In this unsupervised analysis, we only consider the vote intention data as responses. The deviance of the intercepts only model for these data equals 3056.7. In this first analysis of the Dutch election data, we do not take the predictors into account, so we perform an unsupervised analysis. The optimal deviance for the unidimensional representation is 2089.7 (31.6% explained deviance), for the two-dimensional representation 918.52 (70.0% explained), and for the three-dimensional representation 295.1 (90.0% explained). The AICs are 2823.7, 2368.5, and 2459.1 for the one, two, and three dimensional solution, respectively. The two-dimensional map is displayed in Fig. 6.

Both PvdA and VVD have large areas of endorsement, while D66 has the smallest. We see that SP and GL are close together with large overlap in the region of endorsement, both are left-wing oriented parties. More right-wing oriented parties are VVD, LPF and LN; they group together on the other end. D66 and CDA are traditionally more centered parties, where CDA is a Christian party, while D66 is a progressive party. The PvdA is the Dutch labour party.

From the solution, we computed some classification statistics for each political party: the proportion correctly classified, the sensitivity and specificity, the positive predictive value and negative predictive value, the F1-score and the Area under the ROC curve (AUC). See Lever et al. (2016) for a definition and discussion about the use of these metrics. These statistics depend on the observed values y_{ir} and the fitted values $\hat{\pi}_{ir}$. These statistics are shown in the top of Table 1 and show that, generally, the representation is good. Note that, we evaluated the statistics *in sample*, that is, the statistics are computed using the same sample that was used to fit the model. Although usually, these statistics would be evaluated *out of sample*, the main

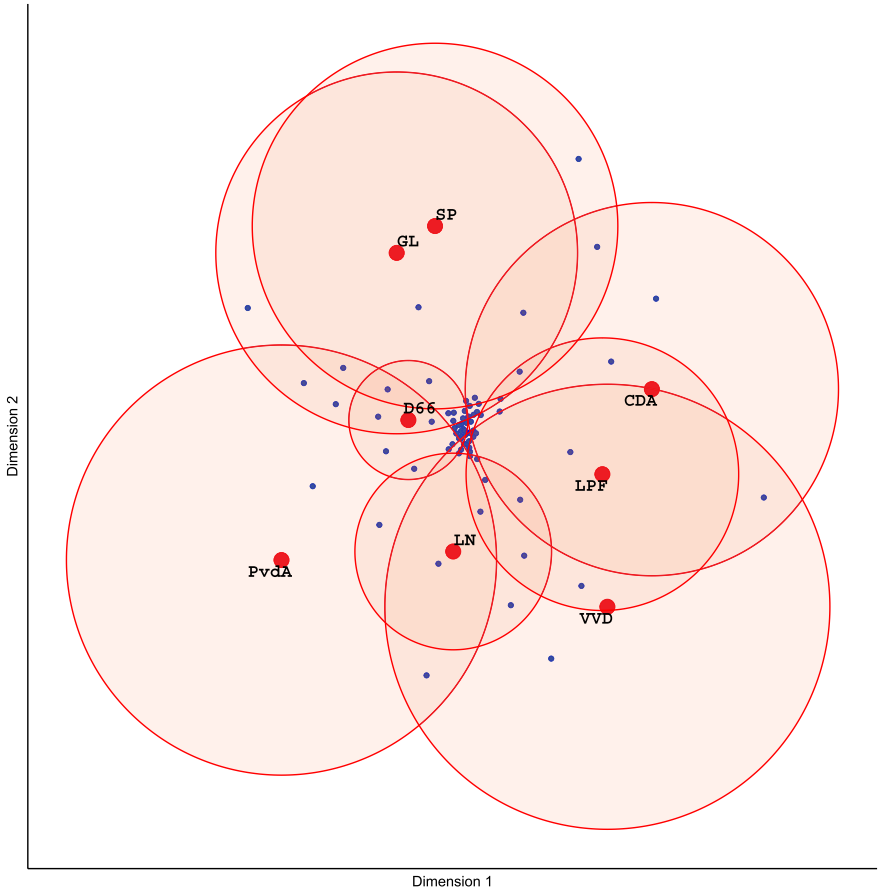


Fig. 6 Two dimensional unsupervised solution for Dutch Election data

reason for in sample evaluation is to compare these statistics for the unsupervised and supervised analysis in the next section.

3.3.2 Supervised analysis

In this subsection, we add the predictors to the analysis. The vote intention are the response variables, the opinions the predictor variables. For the supervised analysis, there is no need to remove the zero profiles from the analysis, as now the predictors have information for those participants.

The optimal deviances are 2770.2, 2716.3, and 2681.4 for the one, two, and three-dimensional solution respectively. The two dimensional solution explains 11.1% of the total deviance (i.e., compared to the intercept only model) and explains 15.9%

Table 1 Classification statistics for each political party based on the unsupervised solution and supervised solution

	PvdA	CDA	VVD	D66	GL	LN	LPF	SP
Unsupervised analysis								
Proportion correct	0.770	0.784	0.784	0.878	0.838	0.905	0.851	0.851
Sensitivity (TP/P)	0.824	0.727	0.792	0.800	0.792	0.909	0.833	1.000
Specificity (TN/N)	0.754	0.808	0.780	0.918	0.860	0.905	0.857	0.828
Pos Pred Value	0.500	0.615	0.633	0.833	0.731	0.625	0.652	0.476
Neg Pred Value	0.935	0.875	0.886	0.900	0.896	0.983	0.941	1.000
F1	0.622	0.667	0.704	0.816	0.760	0.741	0.732	0.645
AUC	0.880	0.904	0.906	0.953	0.917	0.955	0.932	0.892
Supervised analysis								
Proportion correct	0.679	0.662	0.699	0.750	0.764	0.923	0.807	0.841
Sensitivity (TP/P)	0.475	0.533	0.587	1.000	0.707	–	0.696	0.538
Specificity (TN/N)	0.722	0.688	0.729	0.749	0.780	0.923	0.815	0.853
Pos Pred Value	0.264	0.260	0.370	0.011	0.465	–	0.208	0.123
Neg Pred Value	0.868	0.878	0.867	1.000	0.908	–	0.975	0.980
F1	0.339	0.350	0.454	0.022	0.561	–	0.320	0.200
AUC	0.679	0.697	0.743	0.618	0.786	0.768	0.749	0.748

of the deviance of the unsupervised solution. The AICs are 2810.2, 2778.3, and 2763.4 for the one, two, and three dimensional solution, respectively. Although the AIC favors the three dimensional solution, we further discuss the two-dimensional solution in Fig. 7, such that we can compare the supervised solution with the unsupervised one.

Before, we delve into the interpretation, however, we further assess the quality of the mapping. Inspecting the deviance residuals (results not shown), we notice that for 5 participants these residuals are larger than twice the average. Having a closer look at these five observations shows no worrying signs, however. The contributions of the 8 response variables to the overall deviance are 15% for PvdA, 15% for CDA, 14% for both VVD and D66), 13% for GL, 6% for LN, 12% for LPF, and 10% for SP. There is some variation in fit, but there are no political parties that have an extreme contribution to the overall deviance. We also assessed the influence of individual observations. Results are shown in Fig. A1 in Appendix A, where it can be seen that removal of observation 162 leads to a relatively large change in weights and class points, but not in the deviance. Removing this participant from the data did, however, not results in a large change of interpretation. Finally, we also checked the model specification. The component plus residual plots for our mapping are shown in Fig. A2 (Appendix). Overall, the mapping seems to be well specified. For the predictor variable crime there seems to be some minor misspecification in the lower values, especially for response variables CDA, VVD, and LPF. We could try to improve the model specification, by including the quadratic term of crime. Note, however, that including such a quadratic term also influences the relationship of this predictor with the responses that seem to be well specified.

Fig. 7 Two dimensional supervised solution for Dutch Election data. The political parties (red points) represent the responses (i.e., intention to vote for these parties), the circles around these point represent regions of endorsement where the probability is larger than 0.5, the smaller dots represent the participants and the predictor variables are represented by variable axes with labels indicating the values of the predictor variable

Compared to the unsupervised analysis, in this supervised analysis variable axes are included for the predictor variables. The solid part of the lines correspond to values within the observed range (minimum to maximum) of the predictor variable, while the dotted lines extend the variable axis to the border of the display. The length of the solid part of the variable axes can be considered a type of effect size as it displays how much difference a specific predictor variable makes in the positions of the persons in the joint map. The higher the contribution of the variable to the scatter of the person positions, the larger the contribution. Variable names are printed at the positive side of the variable (i.e., high scores). The left-right variable for example, for which positive scores indicate that the subject considers him or herself right-wing, runs from upper right to lower left. Participants that consider themselves right wing are therefore located in the lower left quadrant of the space. Subject positions can be obtained from the variable axes by *interpolation* or a process called completing parallelograms (see, Gower and Hand 1996; Gower et al. 2011).

Considering the positions, we see that PvdA is now close to the other two left-wing parties, SP and GL. The three right-wing parties (VVD, LPF, and LN) are

also closer together. D66 is positioned more in the center, while the CDA stands out a bit. When participants consider themselves left-wing the probability for voting left-wing parties PvdA, GL, and SP is higher. Similarly, when participants consider themselves right-wing, the probability of voting one of the right wing parties (LPF, VVD, LN, or CDA) is higher. When participants indicate that the Netherlands should send back as many asylum seekers as possible and act tougher on crime, they have a higher probability of voting for LPF, LN, or VVD, while those participants that indicate the opposite have a high probability for voting either SP or GL. Participants that indicate euthanasia should always be forbidden, have a higher probability for voting either CDA or SP.

For this supervised analysis we computed the same classification statistics as for the unsupervised analysis (again, in sample). The results are shown in lower half of Table 1. Overall, we see that the classification evaluation metrics become less good compared to the unsupervised analysis. That is expected as we restrict the coordinates of the participants to be linear combinations of the issue opinions, that are the predictor variables. Because the estimated offset parameter for LN ($\hat{m}_{LN}$) is negative, the probability of choosing LN never exceeds 0.5 and therefore some statistics cannot be computed for this party. The positive predictive value and the F1-score are overall quite low compared to the unsupervised solution.

4 Monte Carlo experiments

In this section, we report on two Monte Carlo experiments. The first considers parameter recovery, the second considers predictive performance.

4.1 Parameter recovery

In this section, we will discuss numerical experiments investigating the ability of the algorithm to recover a population distribution. The population distributions are based on the empirical examples of the previous section.

4.1.1 Data generation

In our experiments, we take the estimated parameters from Sect. 3 and some characteristics of the data sets as population parameters.

For the *unsupervised analysis*, we take the mean and covariance matrix from the estimated subject positions and draw a population **U** of 100,000 subjects. The estimated response locations (from Fig. 4 or 6) are taken as the population parameters **V**.

For the *supervised analysis*, we draw 100,000 values from a multivariate normal distribution with mean zero and covariance matrix equal to the observed covariance matrix of the predictors in the Dutch Election data. With these generated predictor variables and the population regression weights (**B**), we compute subject locations (**U**).

For both unsupervised and supervised analysis, we calculate probabilities using the distances between the person locations and the item locations and using the estimated offsets (m_r) as population values. In every replication, we draw a subsample from the persons and use the probabilities (π_{ir}) to draw responses variables (y_{ir}) from the binomial distribution. The subsamples have different sample sizes, that is $n = 100$, 200, 500, and 1000. We fit the mapping on the generated data. One-hundred replications are used.

4.1.2 Evaluation

Let a population configuration be defined by the matrix $\mathbf{Z}$ and its estimate as $\hat{\mathbf{Z}}$. The *congruence coefficient* is a measure of recovery of the population configuration and is defined as (Borg and Groenen 2005, p. 350)

$$\phi = \frac{\sum_{i<j} d_{ij}(\mathbf{Z}) d_{ij}(\hat{\mathbf{Z}})}{\sqrt{\sum_{i<j} d_{ij}^2(\mathbf{Z})}\sqrt{\sum_{i<j} d_{ij}^2(\hat{\mathbf{Z}})}}. \tag{4}$$

We define two such measures, the first ϕ_{uv} for the complete configuration including subject and item points, such that $\mathbf{Z} = \left[\mathbf{U}', \mathbf{V}'\right]'$, whereas the second ϕ_v only uses the item points, that is $\mathbf{Z} = \mathbf{V}$.

Another measure of configuration similarity is obtained by computing the *product-moment correlation coefficient* over the coordinate matrices $\mathbf{Z}$ and $\hat{\mathbf{Z}}$, defined as (Borg and Mair 2022)

$$r = \frac{\text{tr}(\mathbf{Z}'\hat{\mathbf{Z}}\mathbf{T})}{\sqrt{\text{tr}(\mathbf{Z}\mathbf{Z}') \cdot \text{tr}(\hat{\mathbf{Z}}\hat{\mathbf{Z}}')}}, \tag{5}$$

where $\mathbf{T}$ is the estimated Procrustean rotation matrix, and $\text{tr}(\mathbf{A}) = \sum_i a_{ii}$. For evaluation of recovery of the population configuration we focus again on the complete configuration with subject and item points (r_{uv}) and only the item point configuration (r_v)

4.1.3 Results for unsupervised mapping

The simulation results for the Monte Carlo study based on the Sugiyama data and on the election data are given in Table 2. The recovery of the population configuration is overall very good for both studies. The recovery becomes better with larger sample sizes, that is the mean values increase while the standard deviations decrease. Both observations are especially visible for ϕ_v and r_v.

Although the recovery is very good according to these statistics, we need to be somewhat cautious. The congruence coefficient and the correlation coefficient assume that the geometric structure may be translated, dilated, and rotated. Our mapping method, however, only has translational and rotational freedom. Therefore, it is informative to geometrically show the estimated configurations after

Table 2 Results from the simulation study for unsupervised analyses

Data set	Measure		Sample size			
			100	200	500	1000
Religious	ϕ_{uv}	Mean	0.948	0.948	0.952	0.955
		std	0.007	0.008	0.006	0.003
	ϕ_{v}	Mean	0.976	0.979	0.987	0.993
		std	0.015	0.011	0.008	0.003
	r_{uv}	Mean	0.889	0.889	0.896	0.900
		std	0.018	0.015	0.011	0.005
	r_{v}	Mean	0.952	0.957	0.974	0.986
		std	0.028	0.021	0.015	0.006
Election	ϕ_{uv}	Mean	0.949	0.949	0.948	0.947
		std	0.007	0.005	0.003	0.002
	ϕ_{v}	Mean	0.982	0.991	0.996	0.997
		std	0.009	0.005	0.002	0.001
	r_{uv}	Mean	0.893	0.896	0.899	0.898
		std	0.014	0.009	0.005	0.004
	r_{v}	Mean	0.961	0.979	0.990	0.992
		std	0.018	0.010	0.003	0.002

(1) an orthogonal Procrustes analysis towards the population distribution, and (2) a Procrustes similarity analysis, which also takes the dilation into account. In Fig. 8, we show the results of the estimated configurations after orthogonal Procrustes (left) and after a similarity transformation (right). The diamonds show the population points, the dots show the estimated values. The bags include 90% of

(a) (b)

Fig. 8 Graphical representation of the results of the simulation study based on the politics data with sample size equal to 500. Panel (**a**) presents the results after an orthogonal Procrustes analysis without dilation; Panel (**b**) after a similarity transformation with dilation. The diamonds show the population points, the dots show the estimated values. The bags include 90% of the estimated locations

the estimated locations. Whereas in the similarity Procrustes the population values are generally within the 90% bags, for the orthogonal Procrustes analysis we see that estimated positions are more extreme (further away from the origin) than the population values and the population points often fall outside the bags. This feature represents a kind of overfitting, that is, with larger distances and simultaneously larger estimated m_r parameters, the algorithm produces estimated probabilities closer to zero and one. For minimizing the negative log-likelihood this is advantageous.

4.1.4 Results for supervised mapping

The recovery results for the supervised algorithm are shown in Table 3, where it can be seen that the recovery is very good for all sample sizes. The mean values of the coefficients increase with sample size and the standard deviation of the congruence and correlation coefficients decrease with larger sample size. In contrast to the results of the unsupervised case, in this case we found no signals of overfitting.

4.2 Predictive performance

In this second Monte Carlo study, we compare the predictive performance of our logistic multidimensional unfolding with logistic reduced rank regression. Both methods are comparable in the sense that they produce a low dimensional mapping of the data with approximately the same number of parameters. In Sect. 2.2 we showed that multidimensional unfolding can represent a larger number of response profiles in the same dimensional space. The number of represented profiles increases as the offset parameters become larger because then the circles, representing the regions of endorsement, overlap. We expect that this leads to better predictive performance.

Table 3 Simulation results for supervised algorithm

Measure		Sample size			
		100	200	500	1000
ϕ_{uv}	Mean	0.956	0.980	0.994	0.997
	std	0.020	0.009	0.003	0.002
ϕ_{v}	Mean	0.954	0.980	0.994	0.998
	std	0.029	0.015	0.004	0.001
r_{uv}	Mean	0.908	0.958	0.986	0.994
	std	0.047	0.019	0.006	0.003
r_{v}	Mean	0.896	0.952	0.986	0.995
	std	0.062	0.030	0.009	0.002

4.2.1 Data generation

We generate data with $P = 3$ predictor variables from a multivariate standard normal distribution with uncorrelated predictors. The population parameters b_{ps} are (1,0) for the first predictor, (0,1) for the second, and ($\sqrt{2}, \sqrt{2}$ for the third. We use $R = 13$ response variables, with the following coordinates in two-dimensional space

$$\mathbf{V}' = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -1 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 1 & \frac{1}{2} & 0 & -\frac{1}{2} & -1 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 \end{bmatrix}$$

Values for the offsets m_r are drawn in every replication from a uniform distribution. We consider three ranges, the first from -1 till 0, the second range is from -0.5 till 0.5, and the third range from 0 to 1. With these settings, we generated training data sets with 200, 500, and 1000 observations from either the distance model or the inner product model. We also generated for each training set a test set with 1000 observations. We used 100 replications.

4.2.2 Evaluation

Each generated training data set is analysed by both a logistic restricted multidimensional unfolding and a logistic reduced rank regression. When the data are generated by the distance model, we start logistic multidimensional unfolding with the population parameters as initial values and trust this will give good results. When the data are generated by the inner product model, we start logistic multidimensional unfolding with the population parameters as initial values but also perform 25 random starts. The solution with the lowest negative log-likelihood is saved. Logistic reduced rank regression is not hampered by the local optima problem, so for this procedure starting values do not matter.

With the estimated parameters of these two procedures and the predictor variables in the test set, we compute predicted values $\hat{\pi}_{ir}$ for the observations in the test set. We compare the predictive performance of the two methods using the Brier score

$$\sum_{i=1}^{1000}\sum_{r=1}^{13}(y_{ir} - \hat{\pi}_{ir})^2/13000,$$

and compare them using boxplots.

4.2.3 Results

The results of this simulation study are shown in Fig. 9, where the upper row shows the results for data sets generated with the distance model, whereas the lower row shows the results for data sets generated with the inner product model. The upper

and lower row cannot simply be compared because the characteristics of the data, such as the proportion of ones for each response variable differ. The focus therefore is on the predictive performance of the distance model as compared to the inner product model. The three columns in Fig. 9 correspond to the different ranges from which the offsets are drawn.

In case the data are generated with the distance model (upper row), we see that the distance model predicts the test data better than the inner product model. The differences between the two become larger with higher values of the offsets. The reason is that with higher offset values the regions of endorsement become overlapping, leading to a larger number of possible response patterns. The distance model can accommodate a larger number of response patterns than the inner product model.

When the data are generated from a inner product model (lower row), the inner product model predicts the test data slightly better than the distance model. No matter what the values for the offset parameters are (i.e., comparing the left, middle, and right columns of plots), the *difference* in predictive performance of the two models

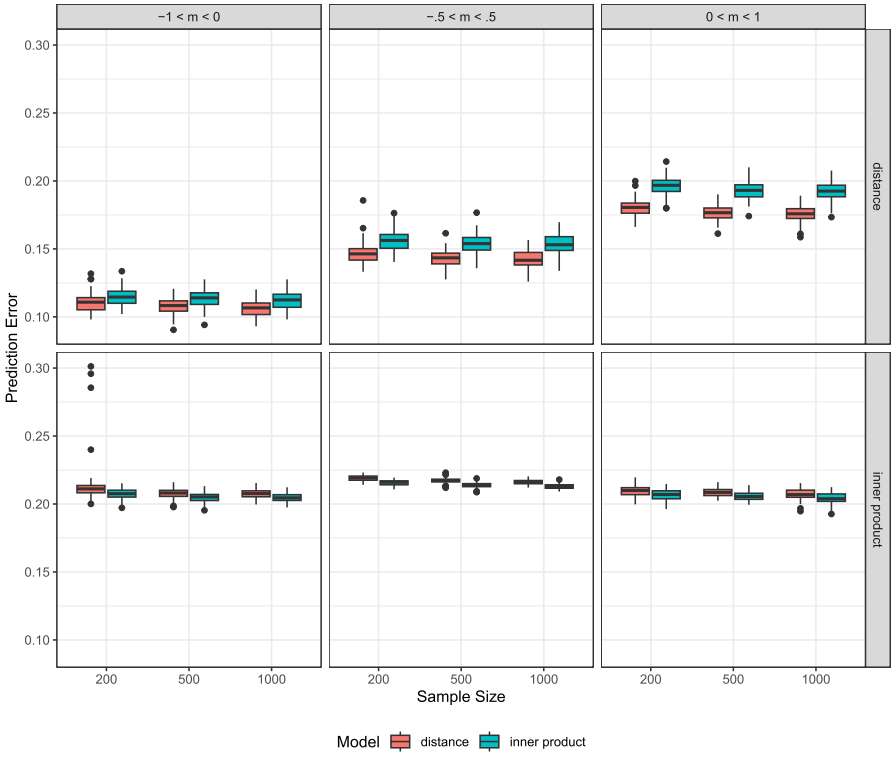


Fig. 9 Prediction error for restricted logistic multidimensional unfolding (distance, left boxplot) compared to logistic reduced rank regression (inner product, right boxplot) for data generated with both population models. In the upper row, data are generated following the distance model whereas in the lower row data are generated with the inner product model. The three different columns represent different ranges of population offset parameters

is approximately the same. As the response curves of the inner product model can be considered a special case of those of the distance model (i.e., take only part of the response curve till (or from) the peak of the curve), we did not expect the differences to be large. The slightly better predictive performance of the inner product model is probably due to an overfitting effect of the distance model, that is, the positions of the response variables are on the boundary of the solution with large offsets, so that the decision boundaries resemble the straight lines of the reduced rank model. Furthermore, the distance model uses a few more parameters, generally leading to an increase in variance and therefore possibly worse predictions. The difference between the two sets of predictions are, however, really small.

In the left lower plot, we see a few outliers with high prediction errors for the distance model with small sample size. This is probably due to a local optimum and could be resolved by more random starts.

5 Discussion

We proposed a new mapping methodology for binary variables based on the two-mode distance function. The distances in the map are connected through a logistic function with the probabilities or expected values. The distinctive feature of the mapping is that it assumes a proximity answer process, where the distance between a person point and an item point determines the probability to endorse or agree with an item: A small distance corresponds to a large probability, whereas a large distance corresponds to a small probability. Every item or response variable is represented by a point and a circle. Distances towards the point are directly related to probability of endorsement. The circles define a region of endorsement, that is, a region where the probability of endorsement is larger than 0.5. When an estimated offset parameter is negative, the corresponding item does not have a region of endorsement because the probability of endorsing never exceeds 0.5. In that case, for each participant we would predict that this item is not endorsed. However, for 100 participants each having a probability of endorsement of 0.3 we expect 30 of these participants to endorse the item.

The proposed mapping can be used for supervised and unsupervised analyses. In a supervised analysis a set of explanatory or predictor variables is available for the participants that restrict the estimates of the points of the participants to be functions of these predictor variables. By including predictors in the map, we specify a functional form of the relationship between the predictors and the response variables. To check for misspecification, we generalized component plus residual plots that are sometimes used in generalized linear models. The plots are applied in the supervised analysis shown in Sect. 3.3.2. In the current paper, we only used additive linear functions but nonlinear and non-additive functions can simply be incorporated by, for example, using spline bases and/or interaction terms. Such terms, would alter the relationship between the variable axes and the participant points and depending on the precise specification might become difficult to interpret. The relationship

between the positions of the participants and the item locations, however, remains the same.

We theoretically compared our mapping to proximity models developed earlier by DeSarbo and Hoffman (1986, 1987) and Takane (1998). Unfortunately, no software is available anymore for these methods. The methods are similar, but these methods use squared Euclidean distances, where we use the distances itself. Usually, researchers interpret distances when looking at a map, not squared distances. DeSarbo and Hoffman (1986, 1987) use offsets for persons, whereas we use offsets for items. Using offsets for persons greatly increases the number of parameters to estimate. In the algorithm section, we showed how offsets for persons could be estimated. We did not incorporate that in our software (yet). In theory it is possible to use both person and item offsets, however this would make the interpretation more difficult and further interpretational guidelines for this case need to be developed. For example, what would it mean if the circles of a person and an item overlap?

An MM-algorithm is proposed for estimation of the map. The main majorization step transforms the negative log-likelihood to a least squares function as shown by Groenen et al. (2003) and De Leeuw (2006). We extended this majorization step with weights for the response profiles. In the inner loop a weighted least squares multidimensional unfolding has to be performed, which can be done by the SMACOF algorithm. The dissimilarities in this step are, however, not guaranteed to be positive. Therefore, we adopted an idea of Heiser (1991) to the unfolding situation. This step is again an MM algorithm, so that the algorithm is a *double MM* algorithm. MM algorithms generally have a linear rate of convergence (Hunter and Lange 2004), that is, they often need many iterations to converge. On the other hand, the computations within the iterations are usually quite simple. As we have a double MM algorithm, where we majorize the majorization function of the negative log-likelihood, the algorithm is slow. Heiser (1995) discussed ways to increase the speed of MM algorithms.

We applied the mapping to two empirical data sets. The first data set is about religious practices. We applied our mapping and compared the results against results obtained by Heiser (1981) using correspondence analysis, a least squares mapping technique for single peaked data. Correspondence analysis can be interpreted using a proximity perspective, as shown by Ter Braak (1985). We showed that such an interpretation is problematic. Our solution has a clearer link between the geometric structure and the probabilities of endorsement. In correspondence analysis, participants are in the center of the items they endorse, but there is not a direct function that translates distances to expected values or probabilities.

The second data set is about vote intentions, where not only the intentions (yes or no) are available but we also have opinions of the participants on a set of opinions, that can serve as explanatory variables. We showed both the unsupervised and supervised analysis and compared the results in terms of mapping but also in terms of classification diagnostics. Including explanatory variables constraints a set of points and therefore classification statistics become worse. A supervised analysis, however, shows relationships between opinions and vote intentions, or more

generally between explanatory variables and responses like in regression models. In that sense, our mapping is similar to multiple logistic regression models where we assume a single peaked relationship between the predictors and the responses. The dependencies between the different response variables are "modelled" by using a low-dimensional map. If we can verify the assumption that given the low dimensional relationship the responses are independent we could extend the mapping with likelihood-based statistics for model selection.

We performed two Monte Carlo experiments. In the first experiment, we evaluated the algorithm and saw that the unsupervised analysis leads to *overfitting*. Consider, for example, the factitious configuration in Fig. 2 and assume that the two response profiles that are not represented in the visualization are also not observed. In that case, the profiles would all fall in the correct region, that is, for each $y_{ir} = 1$ the corresponding probability ($\hat{\pi}_{ir}$) is larger than 0.5, and for each $y_{ir} = 0$ the corresponding probability is smaller than 0.5. Making the complete visualization twice as large, that is multiplying $\mathbf{U}$, $\mathbf{V}$, and each m_r by two, changes the probabilities such that probabilities larger than 0.5 become even larger and probabilities smaller than 0.5 become even smaller. This makes the negative log-likelihood smaller. Therefore, the algorithm will make the map larger and larger, without really changing the overall appearance of the configuration. In empirical data analysis, not all response profiles are correctly represented and these incorrect profiles act as a counterforce against blowing up the map. Similar overfitting issues are also found for the logistic PCA (Song et al. 2019). Possible solutions for this overfitting can be found by including a penalty in the optimization function, for example on the sum of squares of $\mathbf{U}$ or $\mathbf{V}$. In the supervised analysis, we did not find any signs of overfitting, so including predictor variables is also a solution.

In the first Monte Carlo study we used the congruence coefficient and the product-moment correlation as outcome measures. Both measures are generally large even for random data (Borg and Mair 2022) and unfortunately no clear guidelines are available for interpretation. Another measure is the *alienation* coefficient which is simply $\kappa = \sqrt{1 - \phi^2}$. The alienation yields values that vary over a greater range and might be easier to distinguish. All methods have been used mainly for multidimensional scaling with a relatively small set of objects. How these coefficients behave for multidimensional unfolding and our mapping needs further investigation.

In the second Monte Carlo experiment, we evaluated the predictive performance of our mapping against that of a logistic reduced rank model. Data were generated from two population models, one based on inner products the other based on distances. We showed that when the data are generated with a distance model, the distance model has better predictive performance than the reduced rank model. The difference in predictive performance increases as the values of the offsets increase. With larger offsets the number of regions of endorsement increases, each corresponding to a response profile. As the distance model can represent a larger number of response profiles than the inner product model this

was an expected result. When the data are generated using an inner product population model, the reduced rank model predicts slightly better than the distance model. The difference does increase of decrease with changing offsets. The difference in performance might be explained by overfitting. The distance model is capable of fitting monotone response patterns by moving the position of the item on the boundary of the configuration and creating a large offset. This can be inferred from Fig. 1, where the curves are monotonically increasing from the left till the position of the item or decreasing from the point of the time to the right. The distance model uses slightly more parameters than the reduced rank approach because of different identification issues. In our unsupervised analysis we have translational and rotational freedom, which amounts to $S(S+1)/2$ and $S(S-1)/2$ indeterminacies, respectively. In the corresponding inner product model, the number of indeterminacies is S^2. More parameters usually leads to higher variance and reduced predictive accuracy.

In our mapping we use the two-mode distance function. Not all two-mode distance functions give rise to a proximity answer process. In spatial voting models for roll call data (Poole and Rosenthal 1985; Clinton et al. 2004; Poole et al. 2011) and in the MELODIC family (De Rooij and Groenen 2023) also a two-mode distance function is used. The crucial difference between our mapping and these approaches is that in the latter the distance is defined between a subject and a category of an item, whereas in the current paper it is the distance between a subject and an item. When the two-mode distance function is defined towards categories of a binary item or response variable, the structure implies a dominance answer process, such as in (logistic) principal component analysis.

Our unsupervised mapping in the unidimensional case ($S = 1$) is similar to so called single-peaked item response models (Andrich 1988; Hoijtink 1990; Andrich and Luo 1993; Roberts et al. 2000). In these item response models often the person parameters (our $\mathbf{u}_i$) are random effects and the focus is on creating a good measurement scale. Our mapping is not focussed on measurement per se, although in the unidimensional case it could be used for that. Our supervised mapping in the unidimensional case is similar to an explanatory multidimensional single-peaked item response model. Explanatory item response models include predictor variables for either the participants or items. While for the dominance answer process there have been quite some developments (De Boeck and Wilson 2004), for single-peaked response processes explanatory variants are largely lacking although there has been some recent work on explanatory variants of the generalized graded unfolding model (Joo et al. 2022; Usami 2011). All these single-peaked item response models assume a unidimensional latent trait, whereas our mapping is multidimensional. For (single-peaked) item response models often many items are needed while our mapping can deal with a small number of items.

To conclude this paper, the supervised mappings we developed in this manuscript are solutions to what computer scientists call the *multi-label classification* problem

(Gibaja and Ventura 2014; Herrera et al. 2016). Each of the observations is characterised by a set of labels, that is, the set of responses for which $y_{ir} = 1$. Gibaja and Ventura (2014) discuss three different tasks for multi-label learning: the label ranking task, the multi-label classification task, and the multi-label ranking task. The latter task generalizes the first two and provides at the same time a ranking of the labels as well as a bipartition for each observation. The approach we develop in this paper, as well as logistic reduced rank regression, is also a multi-label ranking task. Gibaja and Ventura (2014) and Herrera et al. (2016) discuss three ways of learning from multi-label data set: the data transformation approach, the method adaptation approach, and the ensemble of classifiers. Furthermore, Gibaja and Ventura (2014) describe an overview of transformation and adaptation methods. The approach that we develop in this paper is a method adaptation approach. Siblini et al. (2019) give a review on dimension reduction in multi-label classification. They discussed that both the feature space (i.e., the predictors) as well as the labels space (i.e. the outcomes) can be reduced. Often the dimension reduction is performed independent of the classification. We present a dimension reduction approach that takes into account the multi-label classification task. That is, our loss function is targeted towards classification performance but the mapping finds a *joint space* in which we embed points for the response variables, such that classification performance is optimized. This joint space lies within the column space of the predictor space. An issue in multi-label classification is label dependence, the correlation among labels. In the approach we developed, the dependency between labels is explicitly taken into account through the dimension reduction.

Appendix: Model assessment of supervised map for Dutch election data

In Fig. A1, we show the change in deviance, change in regression weights, and change in item locations statistics when deleting an observation for our analysis in Sect. 3.3.2. It seems that participant 162 has large influence as the regression weights and item locations change substantially when deleting this observations. This change, however, does not affect the deviance much as in that plot participant 162 does not stand out. We verified the solution (not shown) and see no reason to delete this participant from the data.

In Fig. A2, we show the component plus residual plots for our mapping. Although there seems to be some misfit for certain predictor response relationships, e.g., income differences for PvdA or crime for VVD, overall there seems to be no reason for large concern. Note that if we would change the functional form of income differences by for example also including a squared effect term, this would not only affect the relationship to PvdA but also the relationship towards all other response variables.

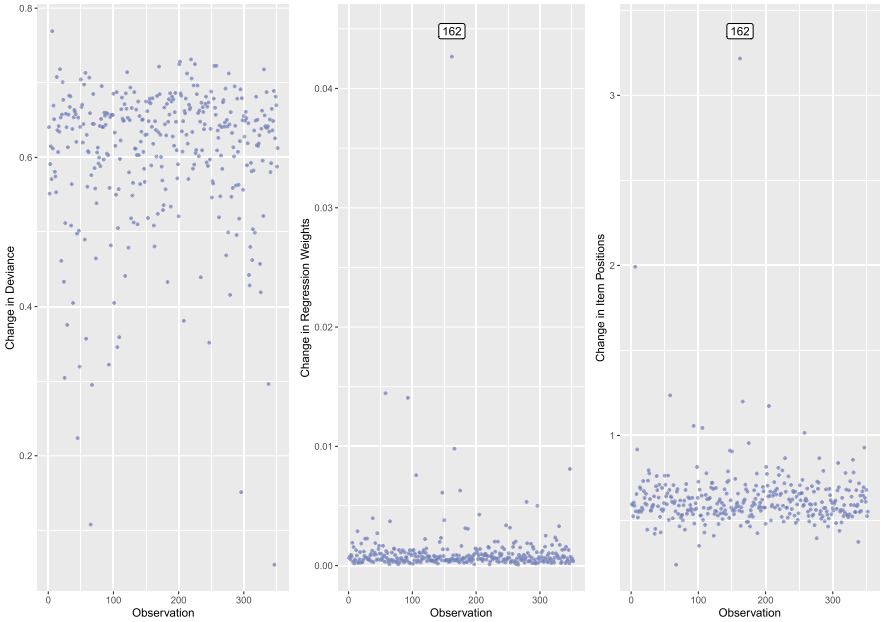


Fig. A1 Influence plots. Observation number (horizontal axis) against Change in deviance $\delta_{\mathcal{D}}$ (left), change in regression weights δ_B (middle), and change in item positions δ_V (right) on the vertical axis. Observation 162 has large values for change in regression weights (middle plot, point at the top) and change in item position (right plot, point at the top)

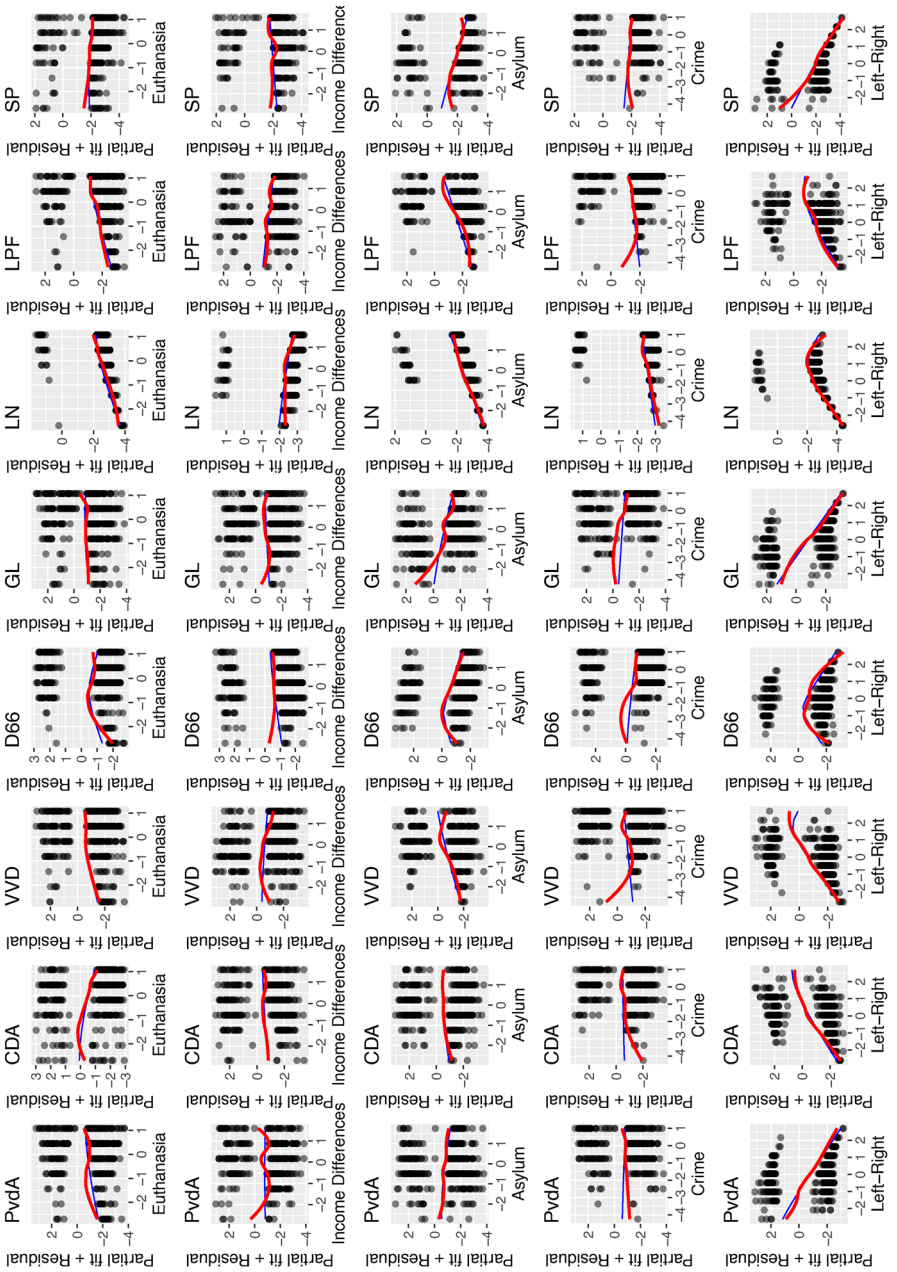


Fig. A2 Component plus residual plots

Acknowledgements 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 published here. We would like to thank the reviewers and guest editors for their constructive remarks on an earlier version of this paper.

Author contributions MDR developed the supervised and unsupervised maps and derived properties of the maps. MDR with FB derived the algorithm and implemented it in R-code. MDR wrote initial draft of the paper and finalized the manuscript. MDR supervised DW in the data analyses and simulation studies; DW performed the data analyses and simulation studies, commented on the first draft of the paper; FB translated parts of the R-code to C++ code. FB commented on the first draft.

Funding No funding was received for conducting this study.

Availability of data and material The Dutch Election data are publicly available after registration from (https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:31979). The Sugiyama data are reported in the paper by Takane (1998) and can be obtained from the corresponding author.

Code availability The code for estimation of the supervised and unsupervised maps is implemented in the R-package `lmap`.

Declarations

Conflict of interest None.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Anderson DR (2007) Model based inference in the life sciences: a primer on evidence. Springer, New York

Andrich D (1988) The application of an unfolding model of the pirt type to the measurement of attitude. Appl Psychol Meas 12(1):33–51

Andrich D, Luo G (1993) A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. Appl Psychol Meas 17(3):253–276

Beh EJ, Lombardo R (2021) An introduction to correspondence analysis. Wiley, Chichester

Borg I, Groenen PJ (2005) Modern multidimensional scaling: theory and applications. Springer, New York

Borg I, Mair P (2022) A note on procrustean fittings of noisy configurations. Austrian J Stat 51(4):1–9
Burnham KP, Anderson DR (2004) Multimodel inference: understanding aic and bic in model selection. Sociol Methods Res 33(2):261–304
Busing FMTA (2010) Advances in multidimensional unfolding. Doctoral thesis, Leiden University
Busing FMTA, Heiser WJ, Cleaver G (2010) Restricted unfolding: preference analysis with optimal transformations of preferences and attributes. Food Qual Prefer 21(1):82–92
Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74(368):829–836
Clinton J, Jackman S, Rivers D (2004) The statistical analysis of roll call data. Am Polit Sci Rev 98(2):355–370
Coombs CH, Kao R (1955) Nonmetric factor analysis. University of Michigan. Department of Engineering Research. Bulletin
Coombs CH (1964) A theory of data. Wiley, Chichester
De Boeck P, Wilson M (2004) Explanatory item response models: a generalized linear and nonlinear approach. Springer, New York
De Leeuw J (1977) Applications of convex analysis to multidimensional scaling. In: Barra J, Brodeau F, Romier G, Van Cutsem B (eds) Recent developments in statistics. North Holland Publishing Company, Amsterdam, pp 133–146
De Leeuw J (1988) Convergence of the majorization method for multidimensional scaling. J Classif 5:163–180
De Leeuw J (2006) Principal component analysis of binary data by iterated singular value decomposition. Comput Stat Data Anal 50(1):21–39
De Leeuw J, Heiser WJ (1977) Convergence of correction matrix algorithms for multidimensional scaling. In: Lingoes J, Roskam E, Borg I (eds) Geometric representations of relational data. Mathesis Press, Ann Arbor, pp 735–752
De Leeuw J, Heiser WJ (1980) Multidimensional scaling with restrictions on the configuration. In: Krishnaiah P (ed) Multivariate analysis, vol V. North Holland, Amsterdam, pp 501–522
De Rooij M (2023) A new algorithm and a discussion about visualization for logistic reduced rank regression. Behaviormetrika 51:389–410
De Rooij M, Groenen PJF (2023) The melodic family for simultaneous binary logistic regression in a reduced space. In: Okada A, Shigemasu K, Yoshino R, Yokoyama S (eds) Facets of behaviormetrics: the 50th anniversary of the behaviormetric society. Springer, Singapore, pp 67–98
De Rooij M, Busing F, Claramunt Gonzalez J (2024) lmap: Logistic Mapping. R package version 0.1.2
DeSarbo WS, Hoffman DL (1986) Simple and weighted unfolding threshold models for the spatial representation of binary choice data. Appl Psychol Meas 10(3):247–264
DeSarbo WS, Hoffman DL (1987) Constructing mds joint spaces from binary choice data: a multidimensional unfolding threshold model for marketing research. J Mark Res 24(1):40–54
Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN (2017) The five factor model of personality and evaluation of drug consumption risk. In: Palumbo F, Montanari A, Vichi M (eds) Data science: innovative developments in data analysis and clustering. Springer, Cham, pp 231–242
Fox J (2015) Applied regression analysis and generalized linear models. Sage, Thousand Oaks
Gibaja E, Ventura S (2014) Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdiscip Rev 4(6):411–444
Gower J, Hand D (1996) Biplots. Taylor & Francis, London
Gower J, Lubbe S, Le Roux N (2011) Understanding biplots. Wiley, Chichester
Greenacre MJ (1984) Theory and applications of correspondence analysis. Chapman and Hall, Boca Raton
Groenen PJF, Giaquinto P, Kiers HAL (2003) Weighted majorization algorithms for weighted least squares decomposition models. Econometric Institute Research Papers EI 2003-09, Erasmus University Rotterdam
Groenen PJF (1993) The majorization approach to multidimensional scaling. DSWO Press, Leiden
Heiser WJ (1981) Unfolding analysis of proximity data. Doctoral dissertation, Leiden University
Heiser WJ (1987) Joint ordination of species and sites: the unfolding technique. In: Legendre P, Legendre L (eds) Developments in numerical ecology. Springer, Berlin, pp 189–221
Heiser WJ (1991) A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative. Psychometrika 56(1):7–27

Heiser WJ (1995) Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. In: Krzanowski WJ (ed) Recent advances in descriptive multivariate analysis. Clarendon Press, New York, pp 157–189
Herrera F, Charte F, Rivera AJ, Del Jesus MJ (2016) Multilabel classification: problem analysis, metrics and techniques. Springer, Cham
Hoijtink H (1990) A latent trait model for dichotomous choice data. Psychometrika 55(4):641–656
Hotelling H (1936) Simplified calculation of principal components. Psychometrika 1(1):27–35
Hunter DR, Lange K (2004) A tutorial on MM algorithms. Am Stat 58(1):30–37
Irwin G, van Holsteyn J, den Ridder J (2003) Nationaal Kiezersonderzoek, NKO 2002 2003. DANS
Jolliffe IT (2002) Principal component analysis. Springer, New York
Joo S-H, Lee P, Stark S (2022) The explanatory generalized graded unfolding model: incorporating collateral information to improve the latent trait estimation accuracy. Appl Psychol Meas 46(1):3–18
Lever J, Krzywinski M, Altman N (2016) Classification evaluation: it is important to understand both what a classification metric expresses and what it hides. Nat Methods 13(8):603–605
Pearson K (1901) On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572
Penninx BW, Beekman AT, Smit JH, Zitman FG, Nolen WA, Spinhoven P, Cuijpers P, De Jong PJ, Van Marwijk HW, Assendelft WJ et al (2008) The Netherlands study of depression and anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatric Res 17(3):121–140
Polak MG (2011) Item analysis of single-peaked response data: the psychometric evaluation of bipolar measurement scales. Doctoral thesis, Leiden University
Polak M, Heiser WJ, De Rooij M (2009) Two types of single-peaked data: correspondence analysis as an alternative to principal component analysis. Comput Stat Data Anal 53(8):3117–3128
Poole KT, Rosenthal H (1985) A spatial model for legislative roll call analysis. Am J Polit Sci 29(2):357–384
Poole K, Lewis JB, Lo J, Carroll R (2011) Scaling roll call votes with wnominate in r. J Stat Softw 42:1–21
Pregibon D (1981) Logistic regression diagnostics. Ann Stat 9(4):705–724
Roberts JS, Donoghue JR, Laughlin JE (2000) A general item response theory model for unfolding unidimensional polytomous responses. Appl Psychol Meas 24(1):3–32
Siblini W, Kuntz P, Meyer F (2019) A review on dimensionality reduction for multi-label classification. IEEE Trans Knowl Data Eng 33(3):839–857
Sloane NJ et al (2003) The on-line encyclopedia of integer sequences. https://oeis.org/A046127
Song Y, Westerhuis JA, Aben N, Michaut M, Wessels LF, Smilde AK (2019) Principal component analysis of binary genomics data. Brief Bioinformatics 20(1):317–329
Spinhoven P, De Rooij M, Heiser W, Smit JH, Penninx BW (2009) The role of personality in comorbidity among anxiety and depressive disorders in primary care and specialty care: a cross-sectional analysis. General Hosp Psychiatry 31(5):470–477
Sugiyama M (1975) Religious behavior of the japanese: Execution of a partial order scalogram analysis based on quantification theory. In US-Japan Seminar of Multidimensional Scaling and Related Techniques, La Jolla
Takane Y (1998) Choice model analysis of the "pick any/n'' type of binary data. Japan Psychol Res 40(1):31–39
Takane Y, Van der Heijden PG, Browne MW (2003) On likelihood ratio tests for dimensionality selection. In: Higuchi T, Iba Y, Ishiguro M (eds) Proceedings of science of modeling: The 30th anniversary meeting of the information criterion (AIC). The Institute of Statistical Mathematics, Tokyo, pp. 348–349
Ter Braak CJ (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41:859–873
Ter Braak CJ (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67(5):1167–1179
Thurstone LL, Chave EJ (1929) The measurement of attitude. University of Chicago Press, Chicago
Usami S (2011) Generalized graded unfolding model with structural equation for subject parameters. Japan Psychol Res 53(3):221–232
Williams D (1987) Generalized linear model diagnostics using the deviance and single case deletions. Journal of the Royal Statistical Society: Series C (Applied Statistics) 36(2):181–191

Yaglom AM, Yaglom IM (1987) Challenging mathematical problems with elementary solutions. Dover Publications, Mineola
Yee TW, Hastie TJ (2003) Reduced-rank vector generalized linear models. Stat Modell 3(1):15–41