the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2022)

published by John Wiley & Sons Ltd
on behalf of British Psychological Society.

www.wileyonlinelibrary.com

# Distance-based logistic model for cross-classified categorical data

José Fernando Vera
Department of Statistics and O.R. Faculty of Sciences, University of Granada, Spain

Logistic regression models are a powerful research tool for the analysis of cross-classified data in which a categorical response variable is involved. In a logistic model, the effect of a covariate refers to odds, and the simple relationship between the coefficients and the odds ratio often makes these the parameters of interest due to their easy interpretation. In this article we present a distance-based logistic model that allows a simple graphical interpretation of the association coefficients using the odds ratio in a contingency table. Two configurations are estimated, one for the rows and one for the columns, as the categories of a polytomous predictor and a nominal response variable respectively, such that the local odds ratio and the distances between the predictor and response categories are inversely related. The associations in terms of the odds ratios, or the ratios of the odds to their geometric means, are interpreted through distances for the most common coding schemes of the predictor variable, and the relationship between the distances related to different codings is investigated in its full dimension. The performance of the estimation procedure is analysed with a Monte Carlo experiment. The interpretation of the model and its performance, as well as its comparison with a two-step procedure involving first a logistic regression and then unfolding, is illustrated using real data sets.

## 1. Introduction

Logistic regression is the most important model for categorical response data. In logistic regression, the odds ratio is usually the parameter of interest, due among other factors to its ease of interpretation. The simple relationship between the coefficients and the odds ratio in logistic regression is one of the main reasons for the widespread use of this procedure.

The odds ratio and the joint probabilities usually receive most attention in the analysis of contingency tables. A wide range of numerical measures, basic inferential procedures, and graphical representations can be used to help visualize and summarize the relationships between two categorical variables. For instance, for nominal response variables, the *IJ* association factor (Good, 1956) focuses on comparing the number of subjects in a cell with the expected number if the variables are independent, while the uncertainty coefficient (Theil, 1970) measures the proportional reduction in entropy. For ordinal variables, concordant and discordant pairs can be used to describe the degree to



Correspondence should be addressed to J. Fernando Vera, Department of Statistics and O.R. Faculty of Sciences, University of Granada, 18071 Granada, Spain (email: jfvera@ugr.es).

DOI:10.1111/bmsp.12264

which a relationship is monotonic, as in the gamma measure (Goodman & Kruskal, 1954), or correlation-based measures, as in the tetrachoric (Pearson, 1904) or the polychoric (Tallis, 1962) correlations. In addition, well known inferential procedures such as the Wald confidence interval for the log odds ratio (Woolf, 1955), score-test-based confidence intervals (Cornfield, 1956), tests for proportions such as those of Agresti and Coull (1998) or Agresti and Caffo (2000), or independence tests such as the well known chi-square test, together with basic graphical procedures such as bar diagrams or mosaic plots (Friendly, 1994), are other widely employed procedures that are also used to analyse associations, usually complementary to modelling.

However, when the analyst wishes to model the relationship between these variables and to determine association patterns, log-linear and logistic models are usually employed (see, e.g., Agresti, 2013). In this article we consider the particular situation of modelling cross-classified data when two categorical variables – a response variable and a related explanatory variable – are involved. In this situation, the odds ratios are of particular importance because they represent parameters in models that pattern the relationships within these cross-classification data, particularly in multinomial logistic regression.

The idea of parameterizing models in terms of graphical representations to facilitate interpretation has provoked great interest as a means of analysing associations within categorical data. For non-sparse data, graphical models for association such as the RC($M$) association model (Goodman, 1985) and the distance association (DA) model (de Rooij & Heiser, 2005) have been proposed. Although these two models are equivalent (de Rooij, 2007, 2008), the latter is based on Euclidean distances and is easier to interpret. DA models are mainly used to estimate one configuration for the row categories and another for the column categories of the table, such that the Euclidean distances between points inversely describe the association between the categories of the two sets, in a log-linear analysis framework.

Many studies have considered the case of a categorical response variable related to several categorical explanatory variables, taking a person-oriented approach to analyse the personal profiles of the variables (Bergman & Magnusson, 1997). In this situation, sparse tables may arise, and the presence of a large number of zero entries may lead the DA model estimation to fail. This problem is aggravated when multiple categorical explanatory and response variables must be considered. When a large number of objects are present, a useful procedure in multidimensional scaling is to combine latent class analysis and graphical representation (Vera, Macías, & Angulo, 2009; Vera, Macías, & Heiser, 2009a). This is also the case with unfolding (Vera, Macías, & Heiser, 2009b). These models are also useful for non-sparse tables involving profiles, for which the DA model can still be estimated. However, in this case the association plot may be difficult to interpret due to the presence of a large number of points (profiles). In this framework, a practical alternative is to combine latent class models (Vera, de Rooij, & Heiser, 2014) and latent block models (Vera & de Rooij, 2020), in conjunction with DA models. This approach makes it possible to represent associations in relation to the estimated clusters.

For cross-classified data, the expression of the odds ratio in terms of Euclidean distances is equivalent both for the DA model in a log-linear analysis framework and for the ideal point discriminant analysis (IPDA) model (Takane, Bozdogan, & Shibayama, 1987). In the latter case, in maximum dimension it can be viewed as a multinomial logistic regression model (de Rooij, 2009). Originally developed for discriminant analysis, the IPDA model is well suited for the situation in which the rows are samples of a multinomial

categorical variable (columns) with fixed row margins (Takane, 1987), but less so for dealing with cross-classified data (de Rooij & Heiser, 2005). In the IPDA model (which is originally constrained, thus reducing the number of parameters to be estimated), the coordinates for the row categories are generally related to mixed predictor variables that will be projected into a low-dimensional space. Moreover, the coordinates of the column categories are not usually estimated (as they are assumed to represent known clusters to which the row points belong), but are represented by weighted averages of the coordinates of the row points in the reduced space. This model provides a graphical interpretation of the conditional probability of a column category given a row category in a multinomial sampling scheme. Some visualization singularities of the IPDA model, together with general properties of the DA model in this framework, are described by de Rooij (2009).

The above models are mainly used to obtain a graphical visualization of the direct associations between row and column categories (RC(*M*) and DA), and of conditional probabilities (IPDA). In both the IPDA and the DA models, the estimated distances can be used to calculate the corresponding odds ratio for each model. Although this process is simple, the graphical display of an odds ratio is not straightforward with these models, as this would require the combined addition and subtraction of four distances representing associations or conditional probabilities (see, e.g., de Rooij & Heiser, 2005). Therefore, when common coding schemes are imposed to interpret parameters in a logistic framework, the configuration derived from these models is not feasible to explain associations in terms of the odds ratio.

In this paper, our aim is to facilitate the interpretation of the association coefficients in a logistic regression problem for cross-classified data when two nominal variables are involved. We propose a distance-based logistic (DBL) model that provides a simple representation of the associations in terms of local odds ratios in a multinomial baseline-category logit model framework. The model is related to the analysis of cross-classified data for a polytomous explanatory variable and a multinomial response variable, regarding the usual set of local odds ratios in which the last two categories of the corresponding contingency table are established as a reference. The DBL model enables us to estimate one configuration for the row categories and another for the column categories, such that the distances between the points inversely represent the associations in terms of the corresponding set of local odds ratios, while the response probabilities are estimated in terms of Euclidean distances.

Although different coding schemes for categorical variables lead to the same substantive results concerning the effects in traditional logistic regression, differences in the estimated parameter values are evident for different codings. In the two coding schemes most commonly used, the local odds ratio and the parameter interpretation are analysed in terms of distances. In general, the Euclidean distances inversely represent the local odds ratio, while for deviations from the mean coding the Euclidean distances also inversely represent the ratio between the corresponding odds and the geometric mean of the odds, which facilitates the interpretation of the parameters in this model.

In the next section, we formulate the DBL model. Section 3 then describes the relationship of the odds ratio with the estimated distances for the two most common coding schemes for a polytomous predictor variable, and the connection between the estimated distances in both codings is investigated in full dimension. Section 4 gives an overview of how the odds ratio can be interpreted in terms of distances for equivalent log-linear models and for the related IPDA model. Section 5 focuses on the estimation

of the parameters, the resulting indeterminacies and the problem of model selection. In Section 6, we analyse the behaviour of the model, based on a Monte Carlo experiment; its performance for empirical data is then illustrated and compared with that of a two-step estimation procedure based on traditional logistic regression and unfolding. Finally, we discuss the results obtained and present the main conclusions drawn.

## 2. Distance-based response probabilities for odds ratio representation

Let us denote by $A$ a polytomous predictor variable with $I$ categories and by $B$ a multinomial response variable with $J$ unordered categories, and consider the $I \times J$ contingency table $\mathbf{F} = (f_{ij})$ collecting the observed counts in a sample of size $N$. Under the traditional multinomial model, let $\pi_{j/i} = P[B = B_j | A = A_i]$ be the probability of the response $B_j$ at the fixed setting $A_i$, for $i = 1\ldots, I, j = 1, \ldots, J$, and denote by $m_{ij} = f_{i.}\pi_{j/i}$ the expected value of the $j$th response in the $i$th row category, where $f_{i.} = \sum_{j=1}^{J} f_{ij}$ is the fixed marginal count of this row. Hence, $\sum_{j=1}^{J}\pi_{j/i} = 1$, $\forall i = 1, \ldots, I$, and the usual multinomial baseline-category logit model (setting the $J$th category as a reference) is given by the equations

$$L_{j/i} = \ln\left[\frac{\pi_{j/i}}{\pi_{J/i}}\right] = \ln\left[\frac{m_{ij}}{m_{ij}}\right] = \alpha_j + \tau_{ij} \quad j = 1, \ldots, J-1; i = 1, \ldots, I, \tag{1}$$

where $\alpha_j$ is the intercept coefficient, and $\tau_{ij} = \boldsymbol{\tau}_j^t \mathbf{a}_i$ is the association coefficient, with $\boldsymbol{\tau}_j$ and $\mathbf{a}_i$ vectors of dimension $I-1$, corresponding respectively to the regression coefficients and the values of the design variables $\tilde{A}_1, \ldots, \tilde{A}_{I-1}$, related to the $i$ th category of $A$. In this model, the response probabilities are given by

$$\pi_{j/i} = \frac{\exp(\alpha_j + \tau_{ij})}{1 + \sum_{j=1}^{J-1}\exp(\alpha_j + \tau_{ij})} \quad \forall i = 1\ldots, I, \quad j = 1, \ldots, J-1, \tag{2}$$

and $\sum_{j=1}^{J}\pi_{j/i} = 1, \quad i = 1, \ldots, I$. As usual for categorical predictors, redundancy in coding is avoided by imposing constraints on the design variables. If the $I$th category is set, in the most usual coding schemes $a_{ii} = 1$ for $i = 1, \ldots, I-1$, and $a_{is} = 0$, for $s \neq i = 1, \ldots, I-1$. For reference cell coding, $\mathbf{a}_I = 0$, which makes $\tau_{Ij} = 0$ and therefore $\alpha_j = L_{j/I}$ and $\tau_{ij} = L_{j/i} - L_{j/I}, i = 1\ldots, I-1, j = 1, \ldots, J-1$.Reference cell coding is a widely used coding scheme, since estimating the risk of a group relative to a control group is usually of interest. If there is no particular group of interest, deviation from the mean coding is the most commonly used method, for which $a_{Is} = -1, s = 1, \ldots, I-1$, and then it is assumed that $\tau_{Ij} = -\sum_{i=1}^{I-1}\tau_{ij}, \forall j = 1, \ldots, J-1$. Hence, it follows that $\alpha_j = (1/I)\sum_{i=1}^{I} L_{j/i}$, and $\tau_{ij} = L_{j/i} - \alpha_j$.Parameter $\tau_{ij}$ represents the effect as the deviation of the logit for the $i$th category from that for the $I$th category in reference cell coding, or from the average logit over all categories of the predictor variable in deviation from mean coding. It is easy to show that for fixed $B_j$ if $\tau_{ij} > \tau_{i'j}$ the logit for the $i$th row is larger than the logit for the $i'$th row, and hence the odds $\Omega_{ij} = \pi_{j/i}/\pi_{J/i}$ will be larger than the odds $\Omega_{i'j} = \pi_{j/i'}/\pi_{J/i'}$.For reference cell coding, exponentiation of the coefficient produces the corresponding odds ratio $e^{\tau_{ij}} = \Omega_{ij}/\Omega_{Ij}$, while for deviation from the mean coding it yields the ratio of the odds for a given category to the geometric mean of the odds,

$$e^{\tau_{ij}} = \frac{\Omega_{ij}}{e^{\alpha_j}} = \frac{\Omega_{ij}}{\prod_{i=1}^{I}(\Omega_{ij})^{1/I}}. \tag{3}$$

In general, the interpretability of this coefficient depends on whether the average odds value is meaningful. However, a simple graphical interpretation of the estimated coefficients can be obtained in terms of Euclidean distances in a model related to unfolding.

### 2.1. Distance-based logistic model (DBL)

In logistic regression, the estimated value of a given parameter only makes sense when compared with that for another category. However, a suitable parameterization may provide an interesting graphical role for the association parameter values. In a baseline logistic regression framework for cross-classified data with a nominal predictor and a multinomial response variable, a new formulation is introduced to obtain a better interpretation of the association parameters in terms of the odds ratio (or the ratio of the odds to their geometric means), by Euclidean distances.

Let us define the $(I-1)\times M$ matrix $\mathbf{X}$ and the $(J-1)\times M$ matrix $\mathbf{Y}$, whose row vectors, $\mathbf{x}_i$ and $\mathbf{y}_j$, are the coordinates of the points representing the row and column categories of the table, respectively, in dimension $M$. Note that the representation of categories $I$ and $J$ is not considered in this model, and therefore the dimension for the unfolding representation is $M \leq \min(I,\ J) - 2$. In the DBL model, we introduce the expression

$$\tau_{ij} = \log\left(\frac{1}{d_{ij}^2}\right), \tag{4}$$

where $d_{ij}^2 = d^2\left(\mathbf{x}_i,\ \mathbf{y}_j\right)$ represents the non-zero squared Euclidean distance between points representing categories $A_i$ and $B_j$ respectively in a space of low dimension $M$, for $i = 1\ldots,\ I-1,\ j = 1,\ \ldots,\ J-1$, and $d_{Ij}^2 = \exp\left(-\tau_{Ij}\right),\ j = 1\ldots,\ J-1$, is defined according to each coding scheme to avoid redundancy in the parameter estimation. Under this formulation, the DBL model is given by the equations

$$L_{j/i} = \ln\left[\frac{\pi_{j/i}}{\pi_{J/i}}\right] = \alpha_j - \log\left(d_{ij}^2\right) \quad j = 1,\ \ldots,\ J-1; i = 1,\ \ldots,\ I, \tag{5}$$

and the expression for the response probabilities is given in terms of the distances by

$$\pi_{j/i} = \frac{\exp(\alpha_j)/d_{ij}^2}{1 + \sum_{j=1}^{J-1}\exp(\alpha_j)/d_{ij}^2}, \qquad \forall i = 1,\ \ldots,\ I, \quad \forall j = 1,\ \ldots,\ J-1, \tag{6}$$

where $\log\left(d_{Ij}^2\right) = 0$ for reference cell coding or $\log\left(d_{Ij}^2\right) = -\sum_{i=1}^{I-1}\log\left(d_{ij}^2\right)$, in deviation from the mean coding, $j = 1,\ \ldots,\ J-1$. From these constraints, the values $d_{Ij}^2$, $j = 1,\ \ldots,\ J-1$, are obtained without the need to assume that they are Euclidean distances, so the estimation of $\mathbf{x}_I$ is not required. In the DBL model $d_{ij}^2 = \exp\left(\alpha_j - L_{j/i}\right)$, and therefore for a fixed category $B_j$ it follows that $d_{ij}^2 \leq d_{i'j}^2$ if and only if $L_{j/i} \geq L_{j/i'}$. Hence, the distance from response category $B_j$ to explanatory category $A_i$ is less than that

to explanatory category $A_{i'}$ when logit $L_{j/i}$ is greater than logit $L_{j/i'}$. This relationship is more readily interpreted in terms of the odds ratio.

## 3. Local odds ratio, odds, and distances

Although both coding schemes in logistic regression produce similar results regarding the odds ratio, the interpretation of the parameters is usually more complicated with deviation from the mean coding (see, e.g., Hosmer, Lemeshow, & Sturdivant, 2013). In the DBL model, different coding schemes lead to different configurations, which facilitates the interpretation of the model.

In the DBL model, $e^{\tau_{ij}} = 1/d_{ij}^2$, which allows the model parameters to be interpreted directly in terms of distances. As mentioned above, exponentiation of the estimated coefficients $\tau_{ij}$ using deviation from the mean coding (3) expresses the odds relative to the corresponding geometric mean of the odds, but this cannot be considered a true odds ratio because the values in the numerator and denominator do not represent the odds for two different categories. However, this expression is of interest when we wish to investigate the main effect for each category without referring to a fixed one. In the DBL model, this parameter can more readily be interpreted in terms of distances: the shorter the distance between an observed category $A_i$ and a response category $B_j$, the greater the deviation of the corresponding odds from the overall (geometric) mean of the odds for this response. For reference cell coding, the distance relationship is obtained directly in terms of the local odds ratio.

The usual basic set of local odds ratio when categories $A_I$ and $B_J$ are set as a reference can be expressed as

$$\Theta_{ij} = \frac{\pi_{j/i}\pi_{J/I}}{\pi_{J/i}\pi_{j/I}} = \exp\left(\tau_{ij} - \tau_{Ij}\right) = \frac{d_{Ij}^2}{d_{ij}^2}, \qquad \forall i = 1, \ \ldots, \ I-1; j = 1, \ \ldots, \ J-1, \tag{7}$$

and their interpretation is easier in terms of distances. Thus, $\Theta_{ij} > 1$ means $d_{ij} < d_{Ij}$, $\Theta_{ij} < 1$ means $d_{ij} > d_{Ij}$, and independence would be equivalent to $d_{ij} = d_{Ij}$, $\forall i = 1, \ \ldots, \ I-1$, $j = 1, \ \ldots, \ J-1$. The latter would correspond to a representation in which all the points for the row (column) categories are condensed into one point while the column (row) points are located equidistant around it. This also applies to a configuration of only two point-clusters, that is, a configuration in which all the row points are condensed into one point and all the column points into another. Any odds ratio can be written in terms of the distances as follows:

$$\begin{aligned} \Theta\left(A_i, \ A_{i'}, \ B_j, \ B_{j'}\right) &= \frac{m_{ij}m_{i'j'}}{m_{i'j}m_{ij'}} = \frac{\pi_{j/i}\pi_{j'/i'}}{\pi_{j/i'}\pi_{j'/i}} = \frac{\Theta_{ij}\Theta_{i'j'}}{\Theta_{i'j}\Theta_{ij'}} \\ &= \frac{\exp\left(\tau_{ij}\right)\exp\left(\tau_{i'j'}\right)}{\exp\left(\tau_{i'j}\right)\exp\left(\tau_{ij'}\right)} = \frac{d_{ij'}^2 d_{i'j}^2}{d_{ij}^2 d_{i'j'}^2}. \end{aligned} \tag{8}$$

As usual, when category $B_J$ is considered the baseline category in the response variable, the odds ratio are more readily interpretable in terms of distances, since for any $B_j$, $j = 1, \ \ldots, \ J-1$,

$$\Theta\left(A_i,\ A_{i'},\ B_j\right) = \frac{\Theta_{ij}}{\Theta_{i'j}} = \frac{d^2_{i'j}}{d^2_{ij}}. \tag{9}$$

Then, as for the logits, for a fixed effect $B_j$, $d^2_{ij} \leq d^2_{i'j}$ if and only if $\Theta\left(A_i,\ A_{i'},\ B_j\right) \geq 1$. Hence, when the distance from response category $B_j$ to explanatory category $A_i$ is less than that from explanatory category $A_{i'}$, the odds of response category $B_j$ versus $B_J$ are higher for predictor category $A_i$ than for $A_{i'}$ and vice versa.

For deviation from the mean coding, $d^2_{Ij} = \left(\prod_{i=1}^{I}\Theta_{ij}\right)^{1/I}$ is the geometric mean of the set of local odds ratios related to $B_j$. This follows from the fact that $d^2_{Ij} = \Theta_{ij}d^2_{ij}$, and that in this coding scheme $\alpha_j = (1/I)\sum_{i=1}^{I} L_{j/i}$, and therefore we can write

$$d^2_{ij} = \frac{\exp(\alpha_j)}{\frac{\pi_{j/i}}{\pi_{J/i}}} = \left(\prod_{m=1}^{I}\frac{\frac{\pi_{j/m}}{\pi_{J/m}}}{\frac{\pi_{j/i}}{\pi_{J/i}}}\right)^{1/I} = \frac{\prod_{m=1}^{I}\Theta_{mj}^{1/I}}{\Theta_{ij}}. \tag{10}$$

For reference cell coding, the estimated configuration allows for much easier interpretation. Since $d^2_{Ij} = 1$, $j = 1,\ \ldots,\ J-1$, it follows that $d_{ij} \leq d_{i'j'}$ if and only if $\Theta_{ij} \geq \Theta_{i'j'}$.

Different coding schemes produce different estimated configurations. In full dimension, the DBL model can be viewed as a multinomial logistic regression model (see Section 5). In this situation, let us denote by $\tilde{\mathbf{D}}$ the matrix of Euclidean distances $\tilde{d}_{ij} = d\left(\tilde{\mathbf{x}}_i,\ \tilde{\mathbf{y}}_j\right)$ for the configurations $\tilde{\mathbf{X}},\ \tilde{\mathbf{Y}}$ obtained under deviation from the mean coding, and by $\dot{\mathbf{D}}$ the matrix of distances $\dot{d}_{ij} = d(\dot{\mathbf{x}}_i,\ \dot{\mathbf{y}}_i)$ for the configurations $\dot{\mathbf{X}}, \dot{\mathbf{X}}$ given under reference cell coding. Then, from (7) if follows that $\tilde{d}^2_{ij} = \tilde{w}_j\dot{d}^2_{ij}$, where $\tilde{w}_j = \prod_i \Theta_{ij}^{1/I}$, is the geometric mean of the local odds ratio involving $B_j$, $j = 1,\ \ldots,\ (J-1)$. Then $\tilde{\mathbf{D}} = \dot{\mathbf{D}}\tilde{\mathbf{W}}$, where $\tilde{\mathbf{W}}$ is the diagonal matrix of elements $\tilde{w}_j$, $j = 1,\ \ldots,\ J-1$. Therefore, in terms of distances, the two coding schemes are related through the geometric means of the local odds ratio, and the two configurations would be equivalent in the case of independence.

## 4. Related odds ratio representations

The relationship between log-linear and logistic regression models is well known. For the IPDA and DA models, this relationship is shown in similar additive expressions for the odds ratio in terms of the estimated distances (de Rooij & Heiser, 2005). As observed by Takane (1987), the IPDA model can also be formulated in terms of joint probabilities rather than conditional probabilities, like the DA model.

For the DA model, the corresponding logistic model is given by the $I \times (J-1)$ equations,

$$\log\left(\frac{m_{ij}}{m_{iJ}}\right) = \left(\lambda_j - \lambda_J\right) + \left(\dot{d}^2_{iJ} - \dot{d}^2_{ij}\right), \qquad \forall i = 1,\ \ldots,\ I, \quad \forall j = 1,\ \ldots,\ J-1. \tag{11}$$

where the $\lambda_j$ are the column effect parameters, and the $\dot{d}^2_{ij}$ denote the estimated squared Euclidean distances in the DA model. The logit for the IPDA model is defined by a similar

expression, with the corresponding distances, in which $\lambda_j$ is replaced by $\log(w_j)$, with $\sum_j w_j = 1$. In terms of the traditional logistic regression model, for reference cell coding it should be $\dot{d}^2_{Ij} = \dot{d}^2_{IJ}$, $\forall j = 1, \ldots, J-1$, while for deviation from the mean coding, $\sum_i \dot{d}^2_{ij} = \sum_i \dot{d}^2_{iJ}$, $\forall j = 1, \ldots, J-1$. Furthermore, after defining $w_j = \exp(\lambda_j)$, the response probabilities derived from the above relation adopt the expression

$$\dot{\pi}_{j/i} = \frac{w_j e^{-\dot{d}^2_{ij}}}{\sum_{j=1}^{J} w_j e^{-\dot{d}^2_{ij}}}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J, \tag{12}$$

where $\dot{d}^2_{iJ} = \lambda_J$, $i = 1, \ldots, I$, which also defines the conditional probabilities for the IPDA model assuming $\sum_j w_j = 1$. These constraints on distances mean that the direct estimation of the configuration using these models is not feasible in logistic regression.The representation of the odds ratio with the DA model adopts the same expression as with the IPDA model, each with respect to its corresponding estimated probabilities (de Rooij & Heiser, 2005), and the local odds ratios are given by

$$\frac{\pi_{j/i}\pi_{J/I}}{\pi_{J/i}\pi_{j/I}} = \exp\left(-\dot{d}^2_{ij} - \dot{d}^2_{IJ} + \dot{d}^2_{jJ} + \dot{d}^2_{iJ}\right). \tag{13}$$

This expression is difficult to interpret. For example, even in the simple situation of a variable that is supposedly measured twice, and the two categories denoted *a* and *b*, de Rooij and Heiser (2005, p. 103) indicate that the log-odds of staying in either of the categories versus making a transition is $\dot{d}^2_{a_1b_2} + \dot{d}^2_{b_1a_2} - \dot{d}^2_{a_1a_2} - \dot{d}^2_{b_1b_2}$, where $\dot{d}^2_{a_1b_2}$ denotes the squared distance between the point representing category *a* at the first time point, and the point representing category *b* at the second in the DA model. Therefore, the log-odds of staying versus moving is equal to the sum of the squared intercategory distances $\left(\dot{d}^2_{a_1b_2} + \dot{d}^2_{b_1a_2}\right)$ minus the sum of the squared intracategory distances $\left(\dot{d}^2_{a_1a_2} + \dot{d}^2_{b_1b_2}\right)$. Hence, the greater the distances between categories *a* and *b*, the stronger the chance of staying; the greater the distances between the categories of the same variable $\left(\dot{d}^2_{a_1a_2},\ \dot{d}^2_{b_1b_2}\right)$ at the first and second time points, the weaker the chance.

For the DBL model, the odds ratio in this simple situation has a simpler interpretation, derived by comparing two distances: that between category *b* at the first time point and category *a* at the second time point, and that between category *a* at the first and second time points,

$$\Theta = \frac{d^2_{b1a2}}{d^2_{a1a2}}. \tag{14}$$

## 5. Parameter estimation, indeterminacies, and model selection

In multinomial logistic regression, parameter estimation is usually performed unconditionally, and parameter values in each coding scheme are obtained by simple reparameterization. With the DBL model, on the other hand, parameter estimation is performed *ad hoc* for each coding scheme using constrained maximum likelihood in a multinomial

baseline-category logit framework, which aims to provide optimal results in terms of the Euclidean distances in relation to the coding of interest.

Only the estimation of $(I-1)+(J-1)$ points in dimension $M \leq \min(I,\ J)-2$ is required, together with the values of the $J-1$ intercept coefficients. The log-likelihood function to be maximized is given by

$$\log L = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij}\log(\pi_{j/i}). \tag{15}$$

In addition to the usual constraints for response probabilities $\pi_{J/i} = 1 - \sum_{j=1}^{J-1}\pi_{j/i}$, for reference cell coding $d_{Ij}^2 = 1$, $j = 1,\ \ldots,\ J-1$, and the log-likelihood function can be written as

$$\begin{aligned}\log L_0 \quad &= \sum_{j=1}^{J-1}\alpha_j f_{\cdot j} - \sum_{i=1}^{I-1}\sum_{j=1}^{J-1} f_{ij}\log\left(d_{ij}^2\right) - \sum_{i=1}^{I-1} f_{i\cdot}\log\left(1+\sum_{j=1}^{J-1}\exp(\alpha_j)\frac{1}{d_{ij}^2}\right)\\ &\quad - f_{I\cdot}\log\left(1+\sum_{j=1}^{J-1}\exp(\alpha_j)\right),\end{aligned} \tag{16}$$

while for deviation from the mean coding $d_{Ij}^2 = -\sum_{i=1}^{I-1} d_{ij}^2$, $j = 1,\ \ldots,\ J-1$, and the log-likelihood function can be expressed as

$$\begin{aligned}\log L_1 \quad &= \sum_{j=1}^{J-1}\alpha_j f_{\cdot j} - \sum_{i=1}^{I-1}\sum_{j=1}^{J-1} f_{ij}\log\left(d_{ij}^2\right) + \sum_{j=1}^{J-1} f_{Ij}\sum_{i=1}^{I-1}\log\left(d_{ij}^2\right)\\ &\quad - \sum_{i=1}^{I-1} f_{i\cdot}\log\left(1+\sum_{j=1}^{J-1}\exp(\alpha_j)\frac{1}{d_{ij}^2}\right) - f_{I\cdot}\log\left(1+\sum_{j=1}^{J-1}\exp(\alpha_j)\prod_{i=1}^{I-1} d_{ij}^2\right).\end{aligned} \tag{17}$$

Since Euclidean distances remain invariant under isometries, as usual, for identification problems regarding translational and rotational invariance the centroid of each dimension can be set to zero, and the solution in the orientation of the principal axis is considered. This entails estimating $n_p = (J-1)+(I+J-2)M - M(M+1)/2$ parameters in the model, after imposing constraints. Accordingly, let $q(\mathbf{X},\ \mathbf{Y})$ be a scalar function of the configuration matrices $\mathbf{X}$ and $\mathbf{Y}$ to be maximized, written in terms of the overall $(I+J-2)\times M$ configuration $\mathbf{Z} = [\mathbf{X}';\mathbf{Y}']'$ as

$$q(\boldsymbol{X},\ \boldsymbol{Y}) = -\frac{1}{2}\sum_{m=1}^{M}\left(\sum_{k=1}^{I+J-2} z_{km}\right)^2 - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=m+1}^{M}\left(\sum_{k=1}^{I+J-2} z_{km}z_{kn}\right)^2. \tag{18}$$

This is the simultaneous maximization of $\log L(\mathbf{X},\ \mathbf{Y},\ \boldsymbol{\alpha})$ and $q(\mathbf{X},\ \mathbf{Y})$, which is assumed to define the maximum likelihood estimate (see, e.g., Ramsay, 1982). Different iterative optimization procedures can be used to solve this system of non-linear equations. In the present case, we use a general optimization procedure based on a quasi-Newton algorithm together with the BFGS method (Fletcher, 1970) implemented in the *fminunc* function of MATLAB (see the Appendix 1 for further details).

Initial values for the iterative algorithm are given as follows. First, initial values for the logits, $L_{j/i}^{(0)} = \log\left(f_{ij}\right) - \log\left(f_{iJ}\right)$, for $i = 1\ldots,\ I-1$, $j = 1,\ \ldots,\ J-1$, are

considered, and initial values for $\alpha_j^{(0)}$ are given according to the corresponding coding scheme. The $(I-1)\times(J-1)$ matrix $\Delta^2$ with elements $\delta_{ij}^2 = \exp\left(\alpha_j^{(0)} - L_{j/i}\right)$ is calculated, and as is usual in the unfolding framework the matrix $\mathbf{U} = -0.5\mathbf{H}\Delta^2\mathbf{J}$ is considered, where $\mathbf{H}$ and $\mathbf{J}$ are the centring matrices in dimensions $(I-1)$ and $(J-1)$ respectively. The initial values for the configurations in dimension $M$ are thus estimated using the eigenvectors associated with the $M$ largest (positive) eigenvalues of the singular value decomposition of $\mathbf{U} = \mathbf{K}\Gamma\mathbf{G}'$, given by $\mathbf{X}^{(0)} = \mathbf{K}\Gamma^{1/2}$, and $\mathbf{Y}^{(0)} = \mathbf{G}\Gamma^{1/2}$.

### *5.1. Model selection*

In general, there is no guarantee that a perfect low-dimensional solution will be found in metric unfolding (Gold, 1973; Schönemann, 1970). Like the IPDA model, the DBL model can be considered a multinomial logistic regression model in full dimension because, in full dimension, the baseline-category logit model can be reparameterized algebraically in terms of a DBL model. The same equivalence in terms of distances is obtained for the DA model in the framework of a saturated log-linear model but in any dimension, since the linear relationship is preserved when all the effects are included in the model, which is one of the main advantages of this method.

The traditional baseline-category logit model can be parameterized in the form $\tau_{ij} = -\log\left(\delta_{ij}^2 + c\right)$, where $\delta_{ij}^2 = \exp(\alpha_j)\frac{\pi_{j/i}}{\pi_{j/i}}$, and $0 < c < \min\delta_{ij}^2$ is a known additive constant, which for the DBL model is set to zero for an efficient approximation in low dimension (see, e.g., Mardia, 1978; Vera & Macías, 2021).

For the estimated parameter values in traditional logistic regression, a suitable value of $c$ can be set such that $d_{ij}^2 = \delta_{ij}^2 + c$ are squared Euclidean distances (Lingoes, 1971; Vera & Macías, 2021). Thus, in full dimension $M \leq (I+J-3)$, $\mathbf{X}$ and $\mathbf{Y}$ are identified using classical scaling in unfolding (Busing, 2010). Furthermore, both configurations will be of full rank, that is, $M = \min(I,\ J) - 2$, if the conditions for the exact estimation procedure for metric unfolding given by Gold (1973) are met (see also Schönemann, 1970).

In general, the chi-square ($\chi^2$) statistic, given by

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(f_{ij} - f_{i.}\pi_{j/i}\right)^2}{f_{i.}\pi_{j/i}}, \tag{19}$$

can be used to determine the goodness of fit of the model, and as for the IPDA model to determine the number of dimensions. In general, in selecting a model from a set of candidates in this framework, an information criterion can be employed (Agresti, 2013, Ch. 6; Takane, 1987). We consider the Bayesian information criterion (BIC; Schwarz, 1978) under the sample size adjustment suggested by Rissanen (1978), where the number of individuals in this model is adjusted by $b = (N+2)/24$ (Raftery, 1995). Thus, the adjusted BIC criterion adopts the expression

$$BIC = -2\log L + n_p \log b, \tag{20}$$

where the model with the lowest BIC value is usually preferred.

## 6. Experimental results

The behaviour of the estimation procedure is first analysed with a Monte Carlo experiment. The interpretation of the model and its performance for empirical data is then illustrated, and compared with that of a two-step estimation procedure based on traditional logistic regression and unfolding.

### 6.1. Monte Carlo experiment

The performance of the estimation procedure was tested in a Monte Carlo experiment. For each coding scheme and sample size of $N = 500$, 1,000, 1,500, 20 matrices were simulated for the values of $I = 8$ and $J = 7$, and analysed with the DBL and the traditional logistic regression models. For each combination of design factors, two sets of $I - 1$ and $J - 1$ points were simulated in two dimensions with an intermixedness index score (Busing, 2010) below 0.05 to avoid degenerate solutions (see, e.g., D'Ambrosio, Vera, & Heiser, 2021), and then the $(I - 1) \times (J - 1)$ matrix of squared Euclidean distances was calculated. The $\tau_{ij} = -\log\left(d_{ij}^2\right)$ values were obtained, and the values of $\tau_{Ij}$ were set in consonance with the corresponding coding constraints, thus establishing the reference category. Hence, the $\pi_{j/i}$ probabilities were calculated using (2), taking the values of $\alpha_j = 0$. For each matrix simulated, the marginal frequencies were randomly set under the condition $\sum_i f_{i.} = N$. To minimize the presence of zeros in the table, only values of $f_{i.} \geq N/4$ were considered, and the value of $f_{I.}$ was adjusted so that the sum of the marginal frequencies was $N$.

With the probabilities thus obtained, each contingency table was simulated according to a multinomial sampling scheme, where $\left(B_1^A, \ \ldots, \ B_I^A\right)$ are independent multinomials with $E[B_i^A(j)] = E[B = B_j/A = A_i] = n_{i.}\pi_{j/i}$ (see, e.g., Agresti, 2013). In summary, two sets of 60 contingency tables were analysed in two and three dimensions to facilitate their visualization, and the results were then compared with those given in full dimension.

In practice, there is no guarantee that a certain logistic regression model will produce a good fit to the data, and the likelihood ratio test can be used to compare one model with other more complex ones that may also contain a non-linear effect. Since the DBL model can also be viewed as a low-dimensional approximation to the traditional baseline-category logit model, the likelihood-ratio test can be used to determine whether the estimated values differ significantly from those given with the model in full dimension. The statistic $\text{LR} = -2(\log L_{\text{DBL}} - \log L_{\text{FULL}})$ is considered, where $\log L_{\text{DBL}}$ and $\log L_{\text{FULL}}$ are, respectively, the log-likelihood values for the DBL model in low and in full dimensions. Table 1 shows the averaged results obtained in each coding scheme. Each configuration estimated in two dimensions was compared with the simulated one after Procrustes. As can be seen in terms of the averaged values, in all situations Procrustes error values close to zero were obtained, which indicates that the original configuration was well recovered by the DBL model in all situations. The average least squares error was obtained for the estimated probabilities with respect to the original ones, and was normalized by applying the sum of the squared values of the original probabilities, in a similar way to the normalized stress function in unfolding. In addition, averaged values are shown for the BIC statistic, and for the *p*-values of both the chi-square and LR statistics. In full dimension, the average least squares error between the estimated association coefficient $\tau$ and the simulated values $-\log\left(d_{ij}^2\right)$ is also shown.

For both coding schemes, the goodness-of-fit test suggests that the model achieves a good fit in all situations, and as expected the BIC statistic suggests that the two-

**Table 1.** Averaged results for simulated data sets

| | | Dim 2 | | | | Dim 3 | | | |
|---|---|---|---|---|---|---|---|---|---|
| Size | pr-err | π-err | BIC | $p(\chi^2)$ | $p$(LR) | π-err | BIC | $p(\chi^2)$ | $p$(LR) |
| Reference cell coding | | | | | | | | | |
| 500 | 0.0031 | 0.00043 | 1,691.11 | 1 | 1 | 0.00051 | 1,721.22 | 1 | 1 |
| 1,000 | 0.0058 | 0.00017 | 3,166.71 | 1 | 1 | 0.00014 | 3,203.62 | 1 | 1 |
| 1,500 | 0.0011 | 0.00004 | 3,863.51 | 1 | 1 | 0.00005 | 3,904.71 | 0.9999 | 1 |
| Deviation from the mean coding | | | | | | | | | |
| 500 | 0.0096 | 0.00071 | 1,572.39 | 1 | 1 | 0.00063 | 1,602.37 | 1 | 1 |
| 1,000 | 0.0085 | 0.00013 | 2,688.10 | 1 | 1 | 0.00015 | 2,725.30 | 1 | 1 |
| 1,500 | 0.0036 | 0.00006 | 3,630.35 | 1 | 1 | 0.00007 | 3,671.23 | 0.9998 | 1 |

| | Reference cell coding | | | Deviation from the mean coding | | |
|---|---|---|---|---|---|---|
| Size | π-err | BIC | τ-err | π-err | BIC | τ-err |
| Full dimension | | | | | | |
| 500 | 0.00062 | 1,766.48 | 0.0186 | 0.00071 | 1,647.76 | 0.0393 |
| 1,000 | 0.00015 | 3,259.40 | 0.0061 | 0.00015 | 2,780.90 | 0.0486 |
| 1,500 | 0.00005 | 3,966.23 | 0.0014 | 0.00006 | 3,732.68 | 0.0018 |

Note. The Procrustes error (pr-err) and the normalized least squares error for the probabilities ($\pi$-err), and in full dimension for the association coefficient ($\tau$-err), are considered. The BIC statistic, and the $p$-values for the $\chi^2$ and for the LR statistic are also shown.

dimensional solution is preferable. Furthermore, the likelihood ratio test revealed no significant differences with respect to the results in full dimension in any situation. Together, these findings show that the model performs well. The normalized least squares error in terms of the estimated probabilities is negligible in all situations and decreases as the sample size increases. Furthermore, the values are similar to those obtained in full dimension, which correspond to the response probabilities estimated using the multinomial logistic model.

### *6.2. German Longitudinal Election Study*

To illustrate the performance of the model, we considered a data set consisting of a sample of 1,000 respondents from the 2017 German Longitudinal Election Study (GLES, 2019). The *gles* data set is available in the R package MNLpred (Neumann, 2020), in which the influence of the ego-position towards immigration (*egoposition-immigration*) on the vote decision (*vote*) for Germany's political parties was analysed. To prevent the existence of zero-value entries in the table while enhancing the categorical nature of the variable (Ramsay, 1973), the *egoposition-immigration* responses, ranging from 0 = *very open* to 10 = *very restrictive*, were recoded into five categories from a to e by merging every two consecutive response categories, except for the last one, in which the remaining three responses were merged. The political parties in the variable *vote* were *AfD* (Alternative for Germany), *FDP* (Free Democratic Party), *Gruene* (Green party), *LINKE* (The Left), *SPD* (Social Democratic Party), and *CDU/CSU* (Christian Democratic–Social Union). *CDU/CSU* was taken as the baseline category. Table 2 shows the resulting contingency table, with a chi-square value of 251.22 and 20 degrees of freedom, in which

**Table 2.** Vote decision data set for 1,000 respondents

| | AfD | FDP | Gruene | LINKE | SPD | CDU/CSU |
|---|---|---|---|---|---|---|
| a | 2 | 6 | 27 | 21 | 40 | 21 |
| b | 4 | 18 | 62 | 48 | 82 | 50 |
| d | 21 | 30 | 9 | 10 | 31 | 72 |
| e | 38 | 15 | 1 | 8 | 20 | 30 |
| c | 4 | 52 | 44 | 36 | 82 | 116 |

Note. The rows represent the ego-position towards immigration with the categories from a = *very open*, b = *moderately open*, c = *neutral*, d = *moderately restrictive*, and e = *very restrictive*, while the columns represent the five main political parties in Germany.

the reference categories for the set of local odds ratios are located in the final position. A log likelihood value of $-1,\ 557.49$ was also obtained in SPSS when the multinomial logit model was fitted.

This data set was first analysed using the DBL model. Deviation from the mean coding was considered for the predictor variable, and the lowest BIC value was obtained in two dimensions (3,193.2) with respect to three dimensions (3,213.7). A log-likelihood value of $-1,\ 559.27$ was obtained in two dimensions in relation to the configuration shown in Figure 1. A value of LR = 3.5564 was obtained for the likelihood ratio statistic, which is related to a $p$-value of $p_{\text{LR}} = 0.9651$ with five degrees of freedom, while a value of $\chi^2 = 3.0712$ was obtained with a $p$-value of $p_{\chi^2} = 0.6890$. Both of these results suggest that the DBL model obtains a good fit. Table 3 shows the local odds ratio values together with the corresponding distances. For instance, *Gruene* is almost as close to a as to b. It is also somewhat closer to e, a position very restrictive towards immigration ($d^2_{4,3} = 3.3056$), than to d, a moderately restrictive position ($d^2_{3,3} = 3.7886$), which means that the value of $\Theta_{3,3}$ is somewhat lower than the value of $\Theta_{4,3}$, as can be appreciated in Table 3.

This data set has also been analysed using the generalized multinomial logistic model under deviation from the mean coding for the predictor variable with the *multinom()* function of the R package *nnet* (Ripley Venables, 2020). Table 4 shows the estimated parameter values obtained, for a log-likelihood value of $-1,\ 557.50$. The first row (Intercept) corresponds to the $\alpha_j$ parameters, while the remaining values in the table are the $\tau_{ij}$ coefficient values for $i = 1,\ \ldots,\ I-1$ and $j = 1,\ \ldots,\ J-1$, where the $I$ th category is $c$ = neutral and the $J$ th category is the political party *CDU/CSU*. The values $\delta^2_{ij} = \exp(-\tau_{ij})$ were obtained, and the *smacof* package (Mair, de Leeuw, Groenen, & Borg, 2020) of R was used to perform metric unfolding in two dimensions for the $(I-1)\times(J-1)$ matrix $\Delta = (\delta_{ij})$. Figure 2 shows the resulting configuration, which is related to a log-likelihood value of $-1,\ 650.877$, significantly lower than that obtained with the DBL model. Therefore, the DBL model outperforms the solution given by a two-step procedure in which first a general multinomial logistic model is fitted and then the configuration is estimated by unfolding. This is more apparent when the local odds ratios are calculated from the unfolding distances estimated using this approach, as shown in Table 5. If we look again at the *Gruene* category, the distance is now smaller with respect to d ($d^2_{3,3} = 1.97$), than it is with respect to e ($d^2_{4,3} = 4.37$), in contrast to the result obtained with the DBL model according to the odds ratio values.

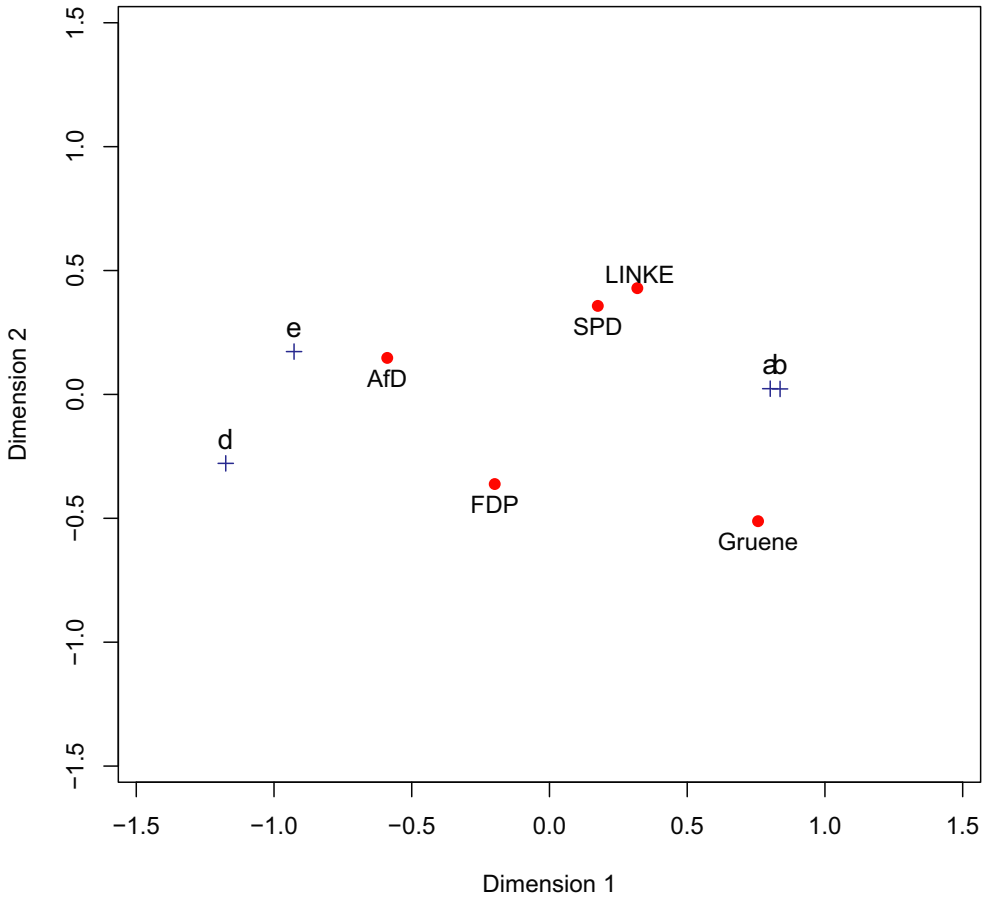


**Figure 1.** Representation of the row and column categories related to the set of local odds ratios for the vote data set using deviation from the mean coding. The DBL model was used taking a neutral ego-position towards immigration (*c*) and the *CDU/CSU* party as the reference categories for the basic set of odds ratios.

### *6.3. Personal profiles*

We also illustrate the performance of the model in a person-oriented approach to data analysis, focusing on the personal profiles of the categorical variables (Bergman & Magnusson, 1997). To do so, we considered a personality data set (Spinhoven, de Rooij, Heiser, Penninx, & Smit, 2009) that was previously analysed, taking a person-oriented approach, by Vera and de Rooij (2020) for sparse tables. In this paper, we analyse a non-sparse data set in which the row profiles are based on the personality variables of Agreeableness and Conscientiousness, while the column profiles are cross-classifications of the four mental disorders Major Depressive Disorder, Generalized Anxiety Disorder, Social Phobia, and Panic Disorder. The personality variables are categorized as High, Moderate, or Low which results in $3^2 = 9$ row profiles. The subjects are diagnosed as being with or without the disorder, which produces $2^4 = 16$ different column profiles, all in a sample consisting of 2, 938 subjects. Table 6 shows the resulting frequencies related to the profiles. Regarding independence, a chi-square value of 537.8 with 120 degrees of

**Table 3.** Local odds ratio and squared distances for the vote data set

| | AfD | FDP | Gruene | LINKE | SPD |
|---|---|---|---|---|---|
| Local odds ratio | | | | | |
| a | 2.12 | 0.79 | 3.30 | 3.30 | 2.57 |
| b | 2.02 | 0.74 | 3.26 | 3.02 | 2.35 |
| d | 7.90 | 0.94 | 0.25 | 0.48 | 0.58 |
| e | 36.03 | 1.11 | 0.28 | 0.81 | 1.03 |
| Squared distances | | | | | |
| a | 1.94 | 1.14 | 0.28 | 0.397 | 0.50 |
| b | 2.04 | 1.22 | 0.29 | 0.431 | 0.55 |
| d | 0.52 | 0.96 | 3.78 | 2.73 | 2.22 |
| e | 0.11 | 0.81 | 3.30 | 1.61 | 1.24 |
| c | 4.14 | 0.90 | 0.95 | 1.31 | 1.29 |

Note. The reference categories are c for the ego-position towards the immigration variable and *CDU/CSU* for the political party variable.

**Table 4.** Estimated parameter values for the vote data sets using multinomial logistic regression with deviation from the mean coding

| | AfD | FDP | Gruene | LINKE | SPD |
|---|---|---|---|---|---|
| (Intercept) | −1.85 | −0.93 | −1.20 | −0.90 | −0.09 |
| a | −0.50 | −0.32 | 1.45 | 0.90 | 0.74 |
| b | −0.68 | −0.09 | 1.41 | 0.86 | 0.59 |
| d | 0.62 | 0.05 | −0.88 | −1.07 | −0.75 |
| e | 2.08 | 0.24 | −2.21 | −0.42 | −0.31 |

freedom was obtained, and a log-likelihood value of $-5,\ 786.2$ was obtained for the multinomial logistic model using SPSS.

The DBL model was run in two dimensions, taking the last column (no disorder) as the baseline category in the response variable. A log-likelihood value of $-5,\ 815.03$ was found taking the last row as the reference category in a reference cell coding scheme for the personality profiles. A value of LR = 57.6175 was obtained for the likelihood ratio statistics, which is related to a $p$-value of $p_{\text{LR}} = 0.9981$ with 92 degrees of freedom, while a value of $\chi^2 = 56.4315$ was obtained with a $p$-value of $p_{\chi^2} = 0.9623$. Both of these results suggest that the DBL model obtains a good fit. Table 7 shows the values obtained for the odds ratio with the DBL model, together with the confidence intervals given with SPSS, when the last category is taken as the reference category in both variables. The estimated squared distances are shown at the top of Table 8.

Figure 3 shows the estimated configuration, which makes it easier to interpret the association patterns in terms of the odds ratio than just looking at the table. Various patterns can be clearly appreciated. For the personality variables, the set {31, 21, 12}, represents profiles with a moderate-low degree of Agreeableness and a high one of Conscientiousness, or high and moderate degrees respectively. A second set consisting of {11, 22} represents, in general, profiles with a low or moderate degree of both variables. Finally, the set {32, 13, 23} corresponds to those profiles without high degree of any variable except when a high degree of Agreeableness but a low one of Conscientiousness

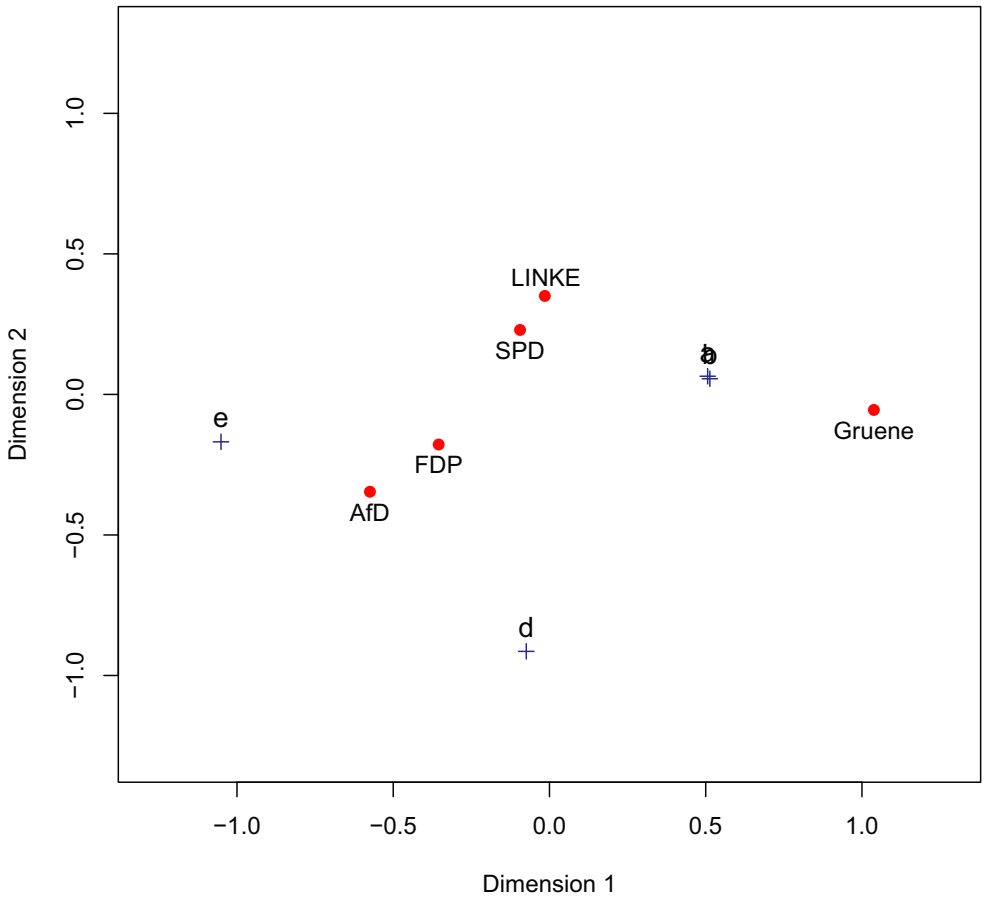


**Figure 2.** Representation of the unfolding configuration for the vote data set after performing generalized multinomial logistic regression in a two-step estimation procedure.

**Table 5.** Basic set of odds ratio and square distances for the vote data set obtained using a two-step procedure of logistic regression and unfolding

| | AfD | FDP | Gruene | LINKE | SPD |
|---|---|---|---|---|---|
| Local odds ratio | | | | | |
| a | 2.81 | 6.42 | 4.55 | 10.18 | 11.87 |
| b | 2.79 | 6.36 | 4.70 | 9.83 | 11.53 |
| d | 6.55 | 8.29 | 0.68 | 2.24 | 3.52 |
| e | 14.50 | 10.59 | 0.31 | 2.68 | 4.28 |
| Unfolding squared distances | | | | | |
| a | 1.34 | 0.80 | 0.30 | 0.35 | 0.39 |
| b | 1.34 | 0.81 | 0.29 | 0.37 | 0.40 |
| d | 0.57 | 0.62 | 1.98 | 1.60 | 1.31 |
| e | 0.26 | 0.49 | 4.38 | 1.34 | 1.07 |

Note. The reference categories are *c* for the ego-position towards immigration and *CDU/CSU* for the political parties.

**Table 6.** Personality data set

| | 1111 | 1112 | 1121 | 1122 | 1211 | 1212 | 1221 | 1222 | 2111 | 2112 | 2121 | 2122 | 2211 | 2212 | 2221 | 2222 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | 366 | 33 | 12 | 15 | 5 | 3 | 1 | 4 | 57 | 20 | 7 | 13 | 7 | 8 | 0 | 6 |
| 12 | 177 | 33 | 17 | 14 | 3 | 3 | 1 | 0 | 33 | 17 | 9 | 14 | 7 | 7 | 2 | 7 |
| 13 | 64 | 14 | 11 | 10 | 6 | 2 | 4 | 2 | 52 | 11 | 13 | 16 | 14 | 5 | 9 | 13 |
| 21 | 184 | 29 | 13 | 10 | 5 | 1 | 3 | 0 | 28 | 14 | 3 | 8 | 5 | 6 | 1 | 4 |
| 22 | 136 | 18 | 13 | 11 | 3 | 2 | 0 | 4 | 29 | 22 | 7 | 14 | 8 | 6 | 2 | 8 |
| 23 | 62 | 19 | 12 | 10 | 5 | 3 | 4 | 2 | 41 | 19 | 12 | 16 | 13 | 9 | 9 | 15 |
| 31 | 136 | 24 | 12 | 4 | 9 | 5 | 2 | 0 | 31 | 12 | 8 | 11 | 8 | 5 | 2 | 6 |
| 32 | 96 | 27 | 20 | 11 | 6 | 4 | 1 | 5 | 34 | 12 | 10 | 12 | 11 | 11 | 6 | 14 |
| 33 | 59 | 16 | 27 | 11 | 7 | 5 | 3 | 9 | 72 | 32 | 21 | 32 | 26 | 18 | 15 | 35 |

Note. The row categories are profiles formed by the personality variables Agreeableness and Conscientiousness categorized as 1 = *high*, 2 = *moderate*, or 3 = *low*. The column profiles are the diagnosis of Major Depressive, Generalized Anxiety, Social Phobia, and Panic Disorders, coded as 1 = *with disorder*, 2 = *without disorder*.

**Table 7.** Local odds ratio for the personality data set using the DBL model

| | 1111 | 1112 | 1121 | 1122 | 1211 | 1212 | 1221 | 1222 |
|---|---|---|---|---|---|---|---|---|
| 11 | 37.85 (14.5, 89.7) | 13.18 (4.2, 34.4) | 3.40 (0.8, 7.7) | 7.93 (2.4, 25.4) | 3.49 (0.9, 17.5) | 2.76 (0.6, 18.6) | 2.26 (0.1, 21.9) | 1.97 (0.6, 11.1) |
| 12 | 16.58 (6.3, 35.5) | 10.01 (3.7, 28.2) | 3.13 (1.1, 8.6) | 4.81 (2.0, 19.7) | 4.06 (0.4, 10.3) | 3.35 (0.5, 15.5) | 2.31 (0.1, 18.4) | 1.28 (0.0, 5.3) |
| 13 | 2.57 (1.4, 6.0) | 2.46 (0.9, 6.1) | 1.11 (0.4, 2.8) | 1.71 (0.8, 7.1) | 1.60 (0.6, 8.1) | 1.40 (0.1, 6.2) | 1.48 (0.7, 18.2) | 1.25 (0.1, 3.1) |
| 21 | 22.69 (9.3, 79.9) | 12.75 (4.7, 52.6) | 2.99 (1.2, 4.3) | 4.78 (2.0, 30.4) | 4.57 (1.3, 29.2) | 3.70 (0.1, 18.9) | 2.52 (1.3, 58.8) | 1.28 (0.0, 10.0) |
| 22 | 14.82 (4.4, 23.0) | 8.67 (1.7, 13.6) | 2.75 (0.7, 5.8) | 5.96 (1.4, 13.6) | 2.94 (0.4, 8.8) | 2.36 (0.2, 10.6) | 2.09 (0.0, 16.0) | 2.04 (0.4, 7.9) |
| 23 | 2.52 (1.2, 4.9) | 2.46 (1.1, 6.8) | 1.05 (0.4, 2.5) | 1.60 (0.7, 6.0) | 1.64 (0.4, 6.1) | 1.44 (0.3, 6.6) | 1.53 (0.6, 15.6) | 1.17 (0.1, 2.6) |
| 31 | 11.13 (5.3, 33.6) | 9.50 (2.9, 25.5) | 2.20 (0.8, 7.7) | 3.06 (0.5, 8.9) | 5.11 (2.0, 27.8) | 4.31 (1.2, 26.4) | 2.75 (0.5, 28.4) | 1.00 (0.0, 6.3) |
| 32 | 4.78 (2.0, 8.1) | 4.66 (1.7, 10.1) | 1.34 (0.7, 4.3) | 2.21 (0.8, 7.0) | 2.58 (0.6, 7.5) | 2.15 (0.4, 8.5) | 2.22 (0.0, 87.2) | 1.29 (0.4, 4.8) |

| | 2111 | 2112 | 2121 | 2122 | 2211 | 2212 | 2221 |
|---|---|---|---|---|---|---|---|
| 11 | 3.90 (1.8, 11.7) | 4.34 (1.3, 10.2) | 2.32 (0.5, 6.5) | 2.68 (0.8, 6.9) | 1.96 (0.4, 5.2) | 2.68 (0.7, 8.6) | 0.66 (0.0, 3.7) |
| 12 | 2.74 (0.9, 5.6) | 3.31 (0.9, 7.2) | 1.64 (0.6, 6.6) | 2.12 (0.7, 6.1) | 1.38 (0.4, 4.3) | 2.17 (0.5, 6.4) | 0.49 (0.1, 3.5) |
| 13 | 1.36 (0.9, 4.0) | 1.33 (0.3, 2.3) | 1.17 (0.6, 4.2) | 1.11 (0.5, 3.2) | 1.10 (0.5, 3.6) | 1.09 (0.2, 2.4) | 1.26 (0.5, 4.5) |
| 21 | 2.69 (1.1, 10.4) | 3.23 (1.1, 12.8) | 1.62 (0.2, 6.1) | 2.07 (0.6, 7.9) | 1.37 (0.4, 6.8) | 2.11 (0.7, 11.6) | 0.52 (0.0, 5.6) |
| 22 | 3.39 (0.7, 4.2) | 3.56 (1.1, 7.7) | 2.23 (0.4, 4.6) | 2.39 (0.7, 5.1) | 1.93 (0.4, 4.0) | 2.36 (0.4, 4.8) | 0.75 (0.1, 3.0) |
| 23 | 1.28 (0.6, 2.7) | 1.26 (0.6, 3.1) | 1.10 (0.5, 3.3) | 1.05 (0.5, 2.7) | 1.04 (0.4, 2.8) | 1.03 (0.4, 3.1) | 1.35 (0.5, 3.9) |
| 31 | 1.94 (0.9, 6.5) | 2.28 (0.7, 6.5) | 1.25 (0.6, 7.2) | 1.57 (0.6, 6.0) | 1.07 (0.5, 5.8) | 1.60 (0.4, 6.0) | 0.49 (0.1, 4.3) |
| 32 | 1.61 (0.5, 2.4) | 1.62 (0.3, 2.3) | 1.27 (0.4, 3.1) | 1.27 (0.3, 2.3) | 1.17 (0.4, 2.7) | 1.25 (0.5, 4.0) | 1.22 (0.3, 3.1) |

Note. Reference categories are 33 for personality profiles and 2222 for mental disorder profiles. The corresponding confidence intervals obtained by SPSS are also shown (in parentheses).

**Table 8.** Squared Euclidean distances between row and column categories for the personality data

| | Reference cell coding | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1111 | 1112 | 1121 | 1122 | 1211 | 1212 | 1221 | 1222 | 2111 | 2112 | 2121 | 2122 | 2211 | 2212 | 2221 |
| 11 | 0.03 | 0.08 | 0.29 | 0.13 | 0.29 | 0.36 | 0.44 | 0.51 | 0.26 | 0.23 | 0.43 | 0.37 | 0.51 | 0.37 | 1.51 |
| 12 | 0.06 | 0.10 | 0.32 | 0.21 | 0.25 | 0.30 | 0.43 | 0.78 | 0.36 | 0.30 | 0.61 | 0.47 | 0.72 | 0.46 | 2.04 |
| 13 | 0.39 | 0.41 | 0.90 | 0.59 | 0.62 | 0.72 | 0.68 | 0.80 | 0.73 | 0.75 | 0.86 | 0.90 | 0.91 | 0.92 | 0.80 |
| 21 | 0.04 | 0.08 | 0.33 | 0.21 | 0.22 | 0.27 | 0.40 | 0.78 | 0.37 | 0.31 | 0.62 | 0.48 | 0.73 | 0.47 | 1.93 |
| 22 | 0.07 | 0.12 | 0.36 | 0.17 | 0.34 | 0.42 | 0.48 | 0.49 | 0.29 | 0.28 | 0.45 | 0.42 | 0.52 | 0.42 | 1.33 |
| 23 | 0.40 | 0.41 | 0.95 | 0.62 | 0.61 | 0.70 | 0.65 | 0.86 | 0.78 | 0.80 | 0.91 | 0.95 | 0.96 | 0.97 | 0.74 |
| 31 | 0.09 | 0.11 | 0.45 | 0.33 | 0.20 | 0.23 | 0.36 | 1.00 | 0.52 | 0.44 | 0.80 | 0.64 | 0.93 | 0.62 | 2.02 |
| 32 | 0.21 | 0.21 | 0.75 | 0.45 | 0.39 | 0.46 | 0.45 | 0.78 | 0.62 | 0.62 | 0.79 | 0.79 | 0.85 | 0.80 | 0.82 |
| | Deviation from the mean coding | | | | | | | | | | | | | | |
| | 1111 | 1112 | 1121 | 1122 | 1211 | 1212 | 1221 | 1222 | 2111 | 2112 | 2121 | 2122 | 2211 | 2212 | 2221 |
| 11 | 0.23 | 0.46 | 0.61 | 0.47 | 0.76 | 0.66 | 1.30 | 0.43 | 0.64 | 0.53 | 0.87 | 0.73 | 0.88 | 0.62 | 1.42 |
| 12 | 0.50 | 0.55 | 0.81 | 0.71 | 0.65 | 0.75 | 0.63 | 1.52 | 0.81 | 0.84 | 0.87 | 0.87 | 0.94 | 0.94 | 1.22 |
| 13 | 2.11 | 1.44 | 1.13 | 1.36 | 0.99 | 1.08 | 0.61 | 1.73 | 1.09 | 1.25 | 0.83 | 0.99 | 0.83 | 1.13 | 0.44 |
| 21 | 0.26 | 0.33 | 0.55 | 0.46 | 0.44 | 0.51 | 0.52 | 1.10 | 0.56 | 0.57 | 0.64 | 0.62 | 0.69 | 0.65 | 1.00 |
| 22 | 0.76 | 0.96 | 0.99 | 0.87 | 1.25 | 1.07 | 1.95 | 0.44 | 1.02 | 0.87 | 1.27 | 1.08 | 1.23 | 0.92 | 1.77 |
| 23 | 2.44 | 1.70 | 1.33 | 1.59 | 1.21 | 1.28 | 0.84 | 1.85 | 1.29 | 1.45 | 1.01 | 1.17 | 0.99 | 1.30 | 0.55 |
| 31 | 0.57 | 0.55 | 0.78 | 0.71 | 0.60 | 0.71 | 0.52 | 1.54 | 0.77 | 0.82 | 0.81 | 0.82 | 0.87 | 0.90 | 1.09 |
| 32 | 2.10 | 1.86 | 1.52 | 1.60 | 1.84 | 1.62 | 2.39 | 0.77 | 1.53 | 1.44 | 1.60 | 1.50 | 1.52 | 1.36 | 1.66 |

Note. The values are shown at the settings obtained using reference cell coding (top) and deviation from the mean coding (bottom).

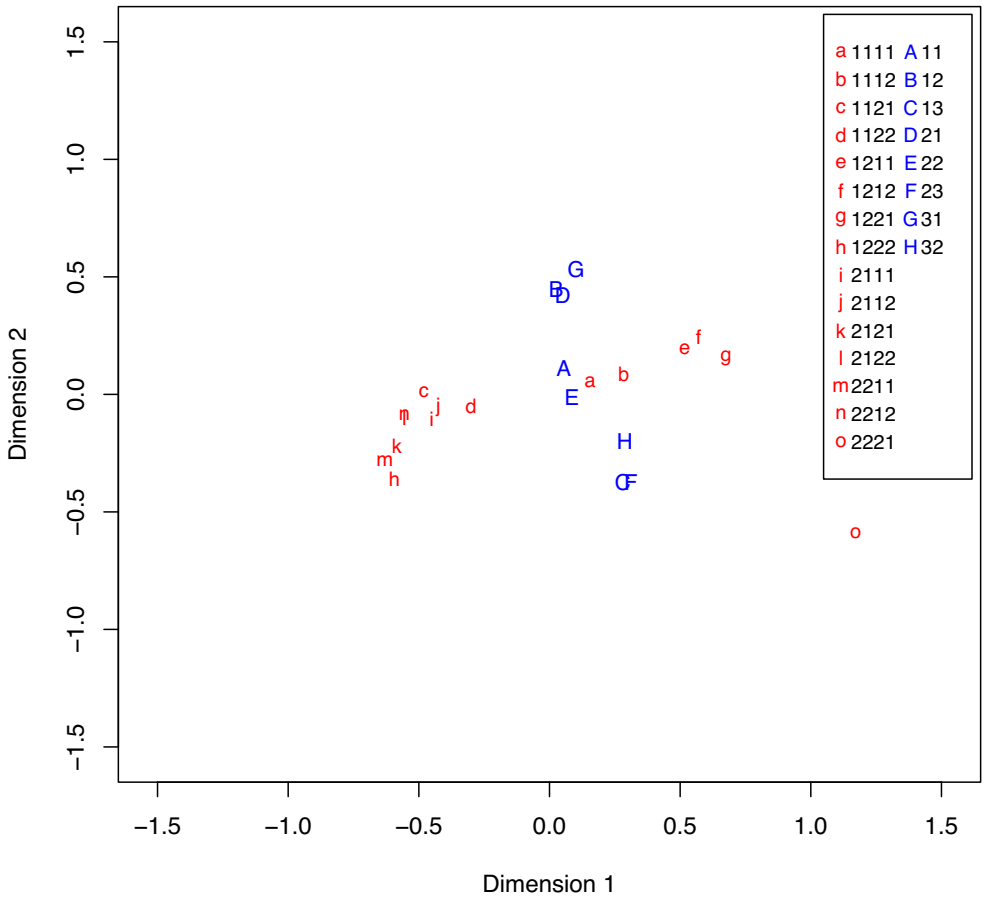


**Figure 3.** Representation of the row and column categories related to the set of local odds ratios for the profiles data set taking (33) as the reference cell for personality. The DBL model was used taking (2222) as the baseline category.

occurs simultaneously. With respect to mental disorder, three sets of profiles with similar characteristics can be seen, together with a single profile corresponding to individuals in whom only Panic Disorder is present. For instance, categories {1211, 1212, 1221} are closer to {31, 21, 12}, followed by {11, 22} and {32, 13, 23}. This indicates a reverse ordering for the odds ratios involving the column profiles in this set. Thus, compared with the profile of low degree of personality in both variables (33), the odds of presenting simultaneously Major Depressive Disorder, no Generalized Anxiety Disorder, and one or both of the remaining two disorders, with respect to having no disorder, are higher for personality profiles with a moderate–low degree of Agreeableness and a high degree of Conscientiousness, or high and moderate degrees, respectively. These sets are followed by those related to the profiles presenting a low or moderate degree in both variables, and finally by those in which there is no high degree in any variable except when high Agreeableness is accompanied by a low degree of Conscientiousness. Similarly, profiles {1111, 1112} are closer to 11 and 22, which means that the odds of presenting at least the first three disorders, compared with not having any, are greater for profiles with a high or

moderate degree in both personality variables, with respect to a low degree in both variables, as could be expected.

This data set was also analysed using the deviation from the mean coding for the personality profiles. A log likelihood value of $-5,\ 813.80$ was obtained in two dimensions when the last column (no disorder) was set as the baseline category in the response variable. The small values of LR $= 55.1606$ and of $\chi^2 = 54.1063$ indicate that there is no evidence of ill fitting ($p$-values of $p_{\text{LR}} = 0.9992$ and of $p_{\chi^2} = 0.9779$ respectively). The squared Euclidean distances between the row and column points for this coding scheme are shown at the bottom of Table 8, and Figure 4 shows the configuration related to the ratio of the odds to the geometric mean of the odds. Here, the set {31, 21, 12}, representing profiles with a moderate-low degree of Agreeableness and a high degree of Conscientiousness, or high and moderate degrees, respectively, again presents a similar pattern, as is also the case with profiles {13, 23}, while profile 32 is now isolated. This is also apparent for profiles 11 and 22. In general, the locations of the disorder profiles do not reveal well defined sets of different patterns with respect to those of personality.

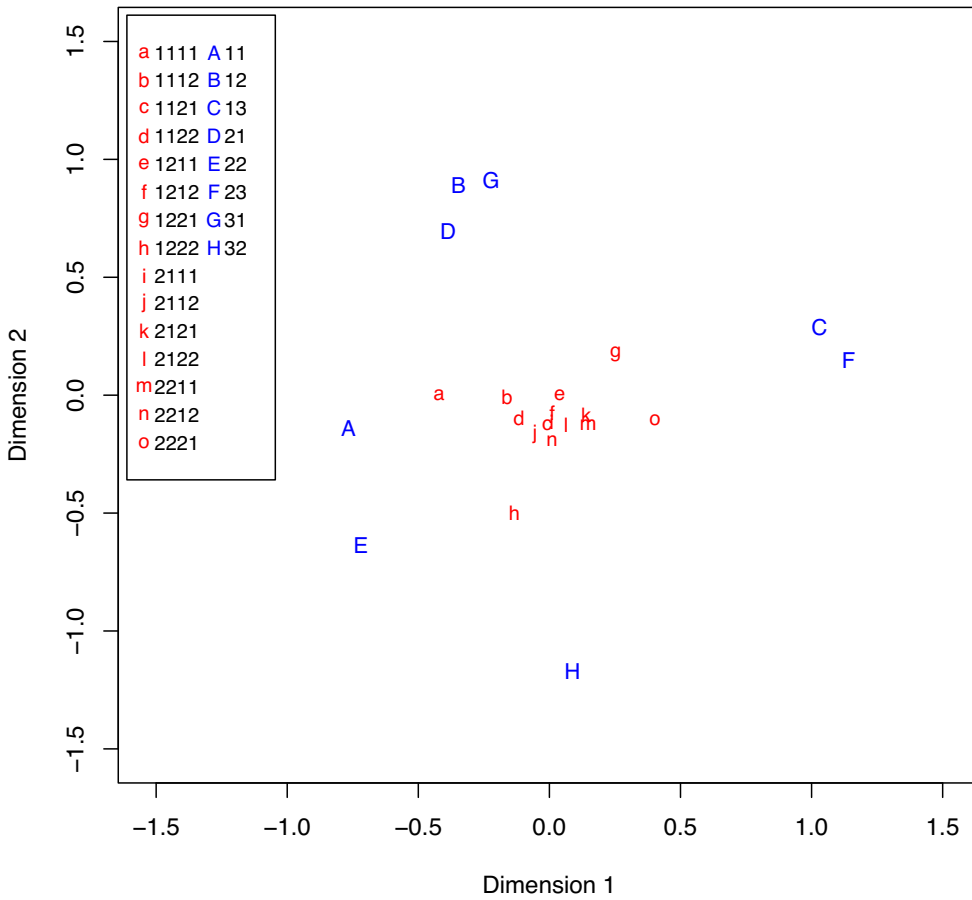


**Figure 4.** Representation of the row and column categories related to the set of local odds ratio for the profiles data set using deviation from the mean coding. The DBL model was used taking (2222) and (33) as the reference categories for the basic set of odds ratios.

For instance, in this context 1111 is near to 11, followed by 21, 12, 31, and somewhat further from 22, and finally to 32, 13, and 23. This means that, when the average (geometric mean) across all personality profiles is considered for the odds of each disorder profile with respect to 2222 (no disorder), for individuals in the category 1111 the odds for the profile featuring a high degree of Agreeableness and Conscientiousness versus no disorder 33, compared with the corresponding average odds across all personality profiles, are higher than the odds of profiles with a low degree of Agreeableness and a high degree of Conscientiousness, or a high–moderate degree of Agreeableness and a low one of Conscientiousness, versus no disorder 33, compared with the corresponding average, followed by the odds of the profile with a moderate degree for both personality of variables 22 versus no disorder 33 compared with its average, and successively by the odds of profiles 32, 13, and 23 versus 33, compared with the average in each case.

Finally, this model was analysed using the two-step procedure described above, in which first a general multinomial logistic model is fitted and then the configuration is estimated by unfolding. The analysis obtained a log-likelihood value of $-6,\ 141.22$ using reference cell coding and $-5,\ 868.70$ for the deviation from the mean coding, which shows that the DBL model outperforms the solution given by a two-step procedure.

## 7. Discussion

This paper presents a DBL model for the analysis of cross-classified data for a polytomous predictor variable and a multinomial response variable. As well as estimating response probabilities based on Euclidean distances, the model represents associations in terms of a basic set of odds ratios, facilitating their interpretation in a baseline-category multinomial logit model framework.

In this approach, we consider the two most common coding schemes for a polytomous predictor variable, each of which generates a configuration that facilitates the interpretation of the associations. In general, Euclidean distances inversely represent the local odds ratio; for the deviation from the mean coding, Euclidean distances also inversely represent the ratio between the corresponding odds and the geometric mean of the odds, thus facilitating the interpretation of the model parameters. The simplicity of the relationship between the odds ratio and the distances is contrasted with the more complicated relationships apparent between the estimated distances in other related models, such as the DA and the IPDA.

Considering the main effect for each category without referring to a fixed one, the smaller the distance between an observed category and a response category, the greater the deviation of the corresponding odds from the overall (geometric) mean of the odds for this response. For reference cell coding, the relationship between the local odds ratios can be directly obtained as the inverse relationship between the distances in the estimated configuration. The relationship between the two configurations is also discussed in terms of their corresponding Euclidean distances in full dimension.

Since it is impossible to assure a perfect low-dimensional solution in metric unfolding, the DBL model can be used as a multinomial logistic regression model in full dimension, in a similar way to the IPDA model. Parameter estimation in the DBL model is performed *ad hoc* for each coding scheme using constrained maximum likelihood in a multinomial baseline-category logit framework. The results thus obtained are optimal for the coding of interest in terms of Euclidean distances. The values estimated in different dimensions and coding schemes are compared and analysed with a Monte Carlo experiment.

The likelihood ratio test is used to determine whether the estimated values in the DBL model differ significantly from those derived using the logistic regression model, and the performance of the model is analysed using the traditional goodness of fit test. The BIC is applied to select the appropriate dimension for the rows and column representation. The ease of interpretation provided by the DBL model is illustrated for two real data sets. The model results are also compared with those of a two-step procedure in which a traditional baseline-categorical multinomial logistic model is fitted for a polytomous predictor, after which a related configuration is estimated using unfolding. In general, the results obtained show that the model we propose recovers the simulated values in the Monte Carlo experiment almost perfectly, and that as expected in all the data sets considered there are no statistically significant differences between the estimated values and those given in full dimension.

The model presented in this paper can be extended to other experimental situations. For example, several categorical explanatory variables combined to conform profiles in a sparse table is an interesting case. In addition, we are currently investigating the application of the DBL model to an ordinal response variable, and the extension of this model, as well as related ones, for the analysis of multi-way contingency tables.

## Acknowledgments

This work has been partially supported by Grant RTI2018-099723-B-I00 (Ministry of Science and Innovation – State Research Agency/10.13039/501100011033/Spain) by 'ERDF A way of making Europe'. Funding for open access charge: Universidad de Granada/CBUA.

## Conflicts of interest

All authors declare no conflict of interest.

## Data availability statement

Supplementary Material is available at the address: http://www.ugr.es/local/jfvera/DBL_BJMSP.zip

## References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York, NY: Wiley.

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, *54*, 280–288. https://doi.org/10.1080/00031305.2000.10474560

Agresti, A., & Coull, B. A. (1998). Approximate is better than exact' for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126. https://doi.org/10.2307/2685469

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, *9*, 291–319. https://doi.org/10.1017/s095457949700206x

Busing, F. M. T. A. (2010). *Advances in multidimensional unfolding* (Unpublished PhD Thesis). Leiden University, The Netherlands.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematics, statistics and probability*, Vol. *4* (pp. 135–148). Berkeley, CA: University of California Press.

D'Ambrosio, A., Vera, J. F., & Heiser, W. J. (2021). Avoiding degeneracies in ordinal unfolding using Kemeny-equivalent dissimilarities for two-way two-mode preference rank data. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2021.1899892

De Rooij, M. (2007). The distance perspective of generalized biadditive models: Scalings and transformations. *Journal of Computational and Graphical Statistics*, *16*, 210–227. https://doi.org/10.1198/106186007x180101

De Rooij, M. (2008). The analysis of change, Newton's law of gravity, and association models. *Journal of the Royal Statistical Society A*, *171*, 137–157. https://doi.org/10.1111/j.1467-985x.2007.00498.x

De Rooij, M. (2009). Ideal point discriminant analysis revisited with an emphasis on visualization. *Psychometrika*, *74*, 317–330. https://doi.org/10.1007/s11336-008-9105-9

De Rooij, M., & Heiser, W. J. (2005). Graphical representations and odds ratios in a distance association model for the analysis of cross-classified data. *Psychometrika*, *70*, 99–122. https://doi.org/10.1007/s11336-000-0848-1

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, *13*, 317–322. https://doi.org/10.1093/comjnl/13.3.317

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, *89*, 190–200. https://doi.org/10.1080/01621459.1994.10476460

GLES. (2019). *Rolling Cross-Section Campaign Survey with Post-election Panel Wave (GLES. 2017)*. Cologne, Germany: GESIS Data Archive. ZA6803 Data file Version 4.0.1. doi:https://doi.org/10.4232/1.13213

Gold, E. M. (1973). Metric unfolding: Data requirement for unique solution and clarification of Schönemann's algorithm. *Psychometrika*, *38*, 555–569. https://doi.org/10.1007/BF02291494

Good, I. J. (1956). On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society B*, *18*, 113–124. https://doi.org/10.1111/j.2517-6161.1956.tb00216.x

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetric models for contingency tables with or without missing entries. *The Annals of Statistics*, *13*, 10–69. https://doi.org/10.1214/aos/1176346576

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764. https://doi.org/10.2307/2281536

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Wiley Series in Probability and Statistics, 3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118548387

Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, *36*, 195–203. https://doi.org/10.1007/BF02291398

Mair, P., de Leeuw, J., Groenen, P., & Borg, I. (2020). *smacof: Multidimensional Scaling: 2.1-1*. Retrieved from https://CRAN.R-project.org/package=smacof

Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. *Communications in Statistics – Theory and Methods*, *7*, 1233–1241. https://doi.org/10.1080/03610927808827707

Neumann, M. (2020). *MNLpred: Simulated predicted probabilities for multinomial logit models. 0.04*. https://CRAN.R-project.org/package=MNLpred

Pearson, K. (1904). *Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. Draper's Co. Research Memoirs, Biometric Series, no. 1*. Reprinted in E. S. Pearson (Ed.), Karl Pearson's early papers. Cambridge, UK: Cambridge University Press, 1948.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*, 513–532. https://doi.org/10.1007/BF02291492

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society A*, *145*, 285–312. https://doi.org/10.2307/2981865

Ripley, B., & Venables, W. (2020). *nnet: Feed-forward neural networks and multinomial log-linear models: R package version 7.3-14*. https://CRAN.R-project.org/package=nnet

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471. https://doi.org/10.1016/0005-1098(78)90005-5

Schönemann, P. H. (1970). On metric multidimensional unfolding. *Psychometrika*, *35*, 349–366. https://doi.org/10.1007/BF02310794

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Spinhoven, P., De Rooij, M., Heiser, W. J., Penninx, B., & Smit, J. (2009). The role of personality in comorbidity among anxiety and depressive disorders in primary care and specialty care: A cross sectional analysis. *General Hospital Psychiatry*, *31*, 470–477. https://doi.org/10.1016/j.genhosppsych.2009.05.002

Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika*, *52*, 493–513. https://doi.org/10.1007/BF02294815

Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika*, *52*, 371–392. https://doi.org/10.1007/BF02294362

Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, *18*, 342–353. https://doi.org/10.2307/2527476

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, *76*, 103–154. https://doi.org/10.1086/224909

Vera, J. F., & de Rooij, M. (2020). A latent block distance-association model for cross-classified categorical data. *Multivariate Behavioral Research*, *55*, 329–343. https://doi.org/10.1080/00273171.2019.1634995

Vera, J. F., de Rooij, M., & Heiser, W. J. (2014). A latent class distance association model for cross-classified data with a categorical response variable. *British Journal of Mathematical and Statistical Psychology*, *67*, 514–540. https://doi.org/10.1111/bmsp.12038

Vera, J. F., & Macías, R. (2021). On the behaviour of k-means clustering of a dissimilarity matrix by means of full multidimensional scaling. *Psychometrika*, *86*, 489–513. https://doi.org/10.1007/s11336-021-09757-2

Vera, J. F., Macías, R., & Angulo, J. M. (2009). A latent class MDS model with spatial constraints for non-stationary spatial covariance estimation. *Stochastic Environmental Research and Risk Assessment*, *23*, 769–779. https://doi.org/10.1007/s00477-008-0257-z

Vera, J. F., Macías, R., & Heiser, W. J. (2009a). A latent class multidimensional scaling model for two-way one-mode continuous rating dissimilarity data. *Psychometrika*, *74*, 297–315. https://doi.org/10.1007/s11336-008-9104-x

Vera, J. F., Macías, R., & Heiser, W. J. (2009b). A dual latent class unfolding model for two-way two-mode preference rating data. *Computational Statistics and Data Analysis*, *53*, 3231–3244. https://doi.org/10.1016/j.csda.2008.07.019

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, *19*, 251–253. https://doi.org/10.1111/j.1469-1809.1955.tb01348.x

*Received 17 May 2021; revised version received 6 December 2021*

## Supporting Information

The following supporting information may be found in the online edition of the article:

Supinfo S1 MATLAB code

## Appendix 1:

### *Gradient vector for maximizing the constrained log-likelihood*

Given the initial values for the parameters, the unconstrained function to be minimized is given by

$$Lq_i(\boldsymbol{\alpha},\ \mathbf{X},\ \mathbf{Y}) = -\log L_i(\boldsymbol{\alpha},\ \mathbf{X},\ \mathbf{Y}) + q(\mathbf{X},\ \mathbf{Y}) \tag{21}$$

where $\boldsymbol{\alpha} = \left(\alpha_1,\ \ldots,\ \alpha_{J-1}\right)'$, and $\log L_i$, $i = 0,\ 1$, is the log-likelihood for the corresponding coding scheme (16) or (17). This function is minimized using a quasi-Newton algorithm together with the BFGS method (Fletcher, 1970) implemented in the *fminunc* function of MATLAB. The partial derivatives conforming the gradient vector in each coding scheme are given by the following expressions:

$$\frac{\partial}{\alpha_r} Lq = f_{.r} - \sum_{i=1}^{I} f_{i.}\pi_{r/i} \tag{22}$$

where the expression of $\pi_{j/i}$ is given by (6) and the corresponding distances for each coding scheme. Hence, for reference cell coding the partial derivatives of the log-likelihood (16) adopt the expression

$$\frac{\partial}{x_{sm}} \log L_0 = \sum_{j=1}^{J-1} \left( f_{s.}\pi_{j/s} - f_{sj} \right) \frac{2\left(x_{sm} - y_{jm}\right)}{d_{sj}^2}, \tag{23}$$

$$\frac{\partial}{y_{rm}} \log L_0 = -\sum_{i=1}^{I-1} \left( f_{i.}\pi_{r/i} - f_{ir} \right) \frac{2(x_{im} - y_{rm})}{d_{ir}^2}, \tag{24}$$

while for deviation from the mean coding the partial derivatives for the log-likelihood (17) have the expression

$$\frac{\partial}{x_{sm}} \log L_1 = \sum_{j=1}^{J-1} \left( \left( f_{s.}\pi_{j/s} - f_{I.}\pi_{j/I} \right) - \left( f_{sj} - f_{Ij} \right) \right) \frac{2\left(x_{sm} - y_{jm}\right)}{d_{sj}^2}, \tag{25}$$

$$\frac{\partial}{y_{rm}} \log L_1 = -\sum_{i=1}^{I-1} \left( \left( f_{i.}\pi_{r/i} - f_{I.}\pi_{r/I} \right) - \left( f_{ir} - f_{Ir} \right) \right) \frac{2(x_{im} - y_{rm})}{d_{ir}^2}. \tag{26}$$

The penalty function $q(\mathbf{Z})$ can be written in terms of $\mathbf{X}$ and $\mathbf{Y}$ as

$$q(X,\ Y) = -\frac{1}{2} \sum_{m=1}^{M} \left( \sum_{i=1}^{I-1} x_{im} + \sum_{j=1}^{J-1} y_{jm} \right)^2 - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=m+1}^{M} \left( \sum_{i=1}^{I-1} x_{im}x_{in} + \sum_{j=1}^{J-1} y_{im}y_{in} \right)^2 \tag{27}$$

and therefore the partial derivatives with respect to $x_{sm}$ and $y_{rm}$ are given by

$$\frac{\partial}{x_{sm}} q = -\sum_{l=1}^{M}\left(\sum_{i=1}^{I-1} x_{il} + \sum_{j=1}^{J-1} y_{jl}\right) - \left[\left(\sum_{l<n}^{M}\left(\sum_{i=1}^{I-1} x_{il}x_{in} + \sum_{j=1}^{J-1} y_{jl}y_{jn}\right)\right)\sum_{l=1}^{M} x_{sl}\right] \tag{28}$$

$$\frac{\partial}{y_{rm}} q = -\sum_{l=1}^{M}\left(\sum_{i=1}^{I-1} x_{il} + \sum_{j=1}^{J-1} y_{jl}\right) - \left[\left(\sum_{l<m}^{M}\left(\sum_{i=1}^{I-1} x_{il}x_{in} + \sum_{j=1}^{J-1} y_{jl}y_{jn}\right)\right)\sum_{l=1}^{M} y_{rl}\right]. \tag{29}$$