PSYCHOMETRIKA—VOL. 74, NO. 2, 317–330
JUNE 2009
DOI: 10.1007/S11336-008-9105-9

IDEAL POINT DISCRIMINANT ANALYSIS REVISITED WITH A SPECIAL EMPHASIS ON VISUALIZATION

MARK DE ROOIJ

LEIDEN UNIVERSITY INSTITUTE FOR PSYCHOLOGICAL RESEARCH

Ideal point discriminant analysis is a classification tool which uses highly intuitive multidimensional scaling procedures. However, in the last paper, Takane wrote about it. He concludes that the interpretation is rather intricate and calls that a weakness of the model. We summarize the conditions that provide an easy interpretation and show that in maximum dimensionality they can be obtained without any loss. For reduced dimensionality, it is conjectured that loss is minor which is examined using several data sets.

Key words: classification, Euclidean distance, (multinomial) logistic regression.

1. Introduction

Ideal Point Discriminant Analysis (IPDA; Takane, Bozdogan, & Shibayama, 1987) was originally proposed as a technique for discriminant analysis with a mixed measurement level of predictor variables. It is a very appealing technique since it uses multidimensional scaling procedures, which are generally thought to be very intuitive, for classification purposes. In maximum dimensionality, IPDA equals multinomial logistic regression (MNL). In other words, IPDA provides a type of biplot (Gower & Hand, 1996) to these logistic regression models. On the other hand, IPDA provides the possibility of dimension reduction as in Canonical Discriminant Analysis (CDA) with the same interpretational facilities. An advantage of IPDA over CDA is that it does not assume normality for the predictor variables, an assumption that in most practical settings is false.

In the last paper of Takane about IPDA (Takane, 1998), he discussed visualization aspects and concludes that there are some weaknesses to IPDA's display. The current paper revisits IPDA, these weaknesses, and the origins thereof. Then it is shown that in maximum dimensionality these can be taken away without any loss, and it is conjectured and empirically illustrated that in reduced space the loss is often minor.

2. IPDA and Visualization

The purpose of classification is to assign subjects ($i = 1, \ldots, n$) to one of several predefined classes ($g = 1, \ldots, G$) based on measurements $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. The explanatory variables $\mathbf{x}_i$ are gathered in a matrix $\mathbf{X}$ as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. In ideal point discriminant analysis, this assignment to classes is based on the following conditional probability model (Takane et al., 1987)

$$\pi_{g|i} = \frac{m_g \exp(-d_{ig}^2)}{\sum_h m_h \exp(-d_{ih}^2)}, \tag{1}$$

This research was conducted while the author was sponsored by the Netherlands Organisation for Scientific Research (NWO), Innovational Grant, no. 452-06-002.

Requests for reprints should be sent to Mark de Rooij, Methodology and Statistics Unit., Leiden University Institute for Psychological Research, P.O. Box 9555, 2300RB Leiden, The Netherlands. E-mail: rooijm@fsw.leidenuniv.nl

where the m_g are bias parameters which can be interpreted as prior probabilities of certain classes or whatever makes a class more or less likely (Takane et al., 1987), and d_{ig}^2 is the squared Euclidean distance in an R-dimensional space between an ideal point for subject i with coordinates y_{ir} and a class point for class g with coordinates z_{gr}, that is,

$$d_{ig}^2 = \sum_{r=1}^{R} (y_{ir} - z_{gr})^2. \tag{2}$$

The ideal points $\mathbf{y}_i = (y_{i1}, \ldots, y_{iR})^T$ that are gathered in a matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ are assumed to be a linear combination of the predictor variables $\mathbf{X}$, i.e.,

$$\mathbf{Y} = \mathbf{XB},$$

with $\mathbf{B}$ the regression weights, which are estimated and from which the ideal points are derived. It is assumed that $\mathbf{X}$ is centered (it might be standardized in order to compare the magnitude of the regression effects). Contrary to standard practice in (generalized) linear models $\mathbf{X}$ does not contain a vector of ones. Such a vector would translate the origin of the Euclidean space and since distances are invariant with respect to such a translation it is omitted. The number of independent parameters in this IPDA model equals $G - 1 + (p + G) \times R - R(R + 1)$ (Takane et al., 1987).

Takane et al. (1987) further restrict the model by placing the class points in the centroids of the ideal points of the subjects observed to be in those classes. Therefore, let $f_{ig} = 1$ if subject i is observed to be in class g, otherwise $f_{ig} = 0$, such that $\sum_g f_{ig} = 1$, and define $\mathbf{F} = \{f_{ig}\}$. Then

$$\mathbf{Z} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}. \tag{3}$$

This centroid restriction will be dropped for the moment, but later on we further comment on this restriction.

Takane (1998) discussed the interpretation of the graphical display, especially the interpretation of the distances between ideal and class points, and concludes they are "rather intricate" and "care should be exercised when they are interpreted in probabilistic terms." Takane's main findings are (Takane, 1998, p. 448):

1. $\pi_{i|g}$ is inversely monotonic with d_{ig} for each class g, so that $d_{ig} > d_{i'g} \Leftrightarrow \pi_{i|g} < \pi_{i'|g}$;
2. π_{ig} is not necessarily inversely monotonic with d_{ig} unless m_g is constant across g;
3. $\pi_{g|i}$ is inversely monotonic with d_{ig} within i for different classes (g) only if the bias parameters (m_g) are constant across classes (g).

In the classification context, the interest is often in the conditional probabilities $\pi_{g|i}$ or the joint probabilities π_{ig} (situation 2 or 3), neither of which are monotonically related to the distances. The bias parameters (m_g) harm this monotonic relationship.

3. The Zero Effect of the Bias Parameters

In dimensionality $G - 1$ (i.e., maximum dimensionality) the effect of the bias parameters on the fit is nil. To show this, we will use dimension augmentation (De Rooij & Heiser, 2005). Therefore, define $a_g = \log m_g$ and rewrite the IPDA model as

$$\pi_{g|i} = \frac{\exp(a_g - d_{ig}^2)}{\sum_h \exp(a_h - d_{ih}^2)}. \tag{4}$$

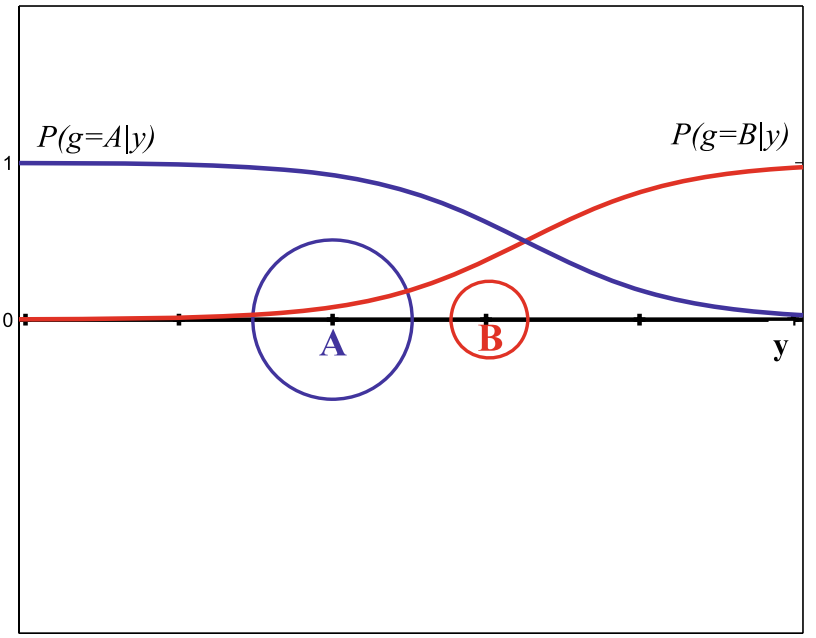

FIGURE 1.
A graphical display of IPDA with two classes. The bias parameters are represented using the area of the *circles*. Conditional probabilities are also shown for the two classes by the *curved lines*. The vertical axis represents the probability scale.

The a_g are identified only up to an additive constant. Due to this indeterminacy, the a_g can be incorporated in the distance part of the model. Define dimension $R+1=G$, with coordinates for the classes $z_{g,R+1}=\sqrt{\max_g(a_g)-a_g}$ and ideal points equal to zero ($y_{i,R+1}=0$). Now the classification model is solely based on distances. It has the following form:

$$\pi_{g|i}=\frac{\exp(-d_{ig}^2)}{\sum_h \exp(-d_{ih}^2)}, \tag{5}$$

where the distances are defined in dimensionality $R+1$ (whereas in earlier definitions the dimensionality was R).

We illustrate this using a two class model in a single dimension, although the same reasoning holds for G class models in $(G-1)$-dimensional space. The solution of an IPDA model is shown in Fig. 1 where the two classes A and B have their location at 0 and 1, respectively. The bias parameters are represented by the area of the circles around the points, i.e., the bias parameter for A is large, while that of B is small. Furthermore, the conditional probabilities of the two classes are also shown. It should be noted that the conditional probability of being in class B at the position of B is smaller than the conditional probability of being in class A. The decision boundary is placed at the crossing of the two probability lines, that is, at the right-hand side of B.

In Fig. 2, we give the same IPDA solution but with augmented dimensionality. Since point A had largest bias parameter, it has a zero coordinate on the vertical axis, for B the coordinate on the second axis is $\Delta=\sqrt{a_A-a_B}$. The two new points A' and B' are shown. Since this augmented space is solely based on Euclidean distances, the decision line is exactly in the middle of A' and B' and it is represented by the dotted line. Note that it indeed crosses the horizontal axis at the point where the two conditional probabilities lines also crossed. Observe that the two points A' and B' still lie in a one dimensional space. The (original) ideal points ($\mathbf{y}$) can be projected

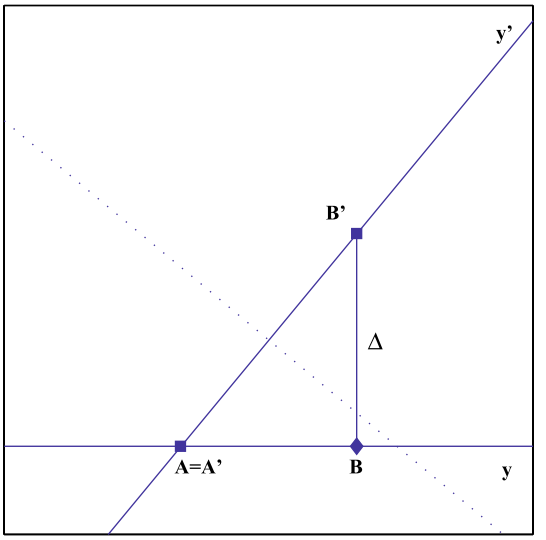

FIGURE 2.
A graphical display of IPDA with the unique dimension (vertical) representing the bias parameters. The horizontal axis **y** is the original one-dimensional space, $\mathbf{y}'$ is the one-dimensional space after translation. Δ represents the square root of the absolute value of the difference between bias parameters.

onto this new one-dimensional space to obtain $\mathbf{y}'$. That is, using the rules of projection $\mathbf{y}'$ can be written as

$$\mathbf{y}' = \mathbf{y} \times \frac{d_{ab}}{\sqrt{d_{ab}^2 + \Delta^2}} \tag{6}$$

where d_{ab} is the distance between the two class points on the **y**-axis (horizontal/original). The multiplication with **y** changes the regression weights

$$\mathbf{y}' = \mathbf{X}\mathbf{b}' = \mathbf{X}\mathbf{b} \times \frac{d_{ab}}{\sqrt{d_{ab}^2 + \Delta^2}}.$$

The new regression weights $\mathbf{b}'$ are equal to $\mathbf{b}' = \mathbf{b} \times \frac{d_{ab}}{\sqrt{d_{ab}^2+\Delta^2}}$. The new coordinates for the class points are $z'_A = \frac{z_A}{\sqrt{d_{ab}^2+\Delta^2}}$ and $z'_B = z'_A + \sqrt{d_{ab}^2 + \Delta^2}$. We thus found a new one-dimensional space with the same classification probabilities ($\pi_{g|i}$), but without the bias terms.

Comparing the distances on $\mathbf{y}'$ with those in the two-dimensional plane, we can say that the effects of this projection are that the distances between ideal points and class points change. These distances change in such a way that the choice probabilities are unaffected since the squared length of a line segment perpendicular to $\mathbf{y}'$ from a point on **y** to a point on $\mathbf{y}'$ has no effect on the classification probabilities, being common to both squared distances from the point on **y** to all the class points on $\mathbf{y}'$. Since the likelihood is a function of the probabilities, the transformation does not change it's value. The distances between ideal points are uniformly shrunk. The distances between the class points remain the same compared to the distances in the two-dimensional plane, but became larger compared to the original one dimensional representation. These latter two sets of distances, however, do not affect the likelihood.

If the class coordinates are subject to the centroid restriction as proposed by (Takane et al., 1987) this combination of changes is not possible, since if the distances between ideal points uniformly shrink by the centroid restriction the distances between class points also uniformly shrink.

Now suppose there are three class points in a two-dimensional plane. The bias parameters can be transformed to coordinates on a third dimension. For one of the three class points (the one with the largest bias parameter), this coordinate is zero, so two of the three points are raised in this third dimension. However, there still exists a two-dimensional plane containing these three points. The ideal points can be projected onto this two-dimensional plane which changes the regression weights, but not the classification probabilities. The coordinates of the class points have to be recomputed relative to this new two-dimensional plane. Further generalizing, suppose there are G classes, which lie in a $(G-1)$-dimensional space. The bias parameters can be transformed to coordinates on a Gth dimension, but still the G points lie in a $(G-1)$-dimensional subspace. The original space can be projected onto this new space which changes the regression weights, but not the classification probabilities and the class points have to be recomputed with respect to this new plane, but preserving their distances.

In summary, in maximum dimensionality the bias parameters have no effect on the fit of the model, they complicate the interpretation in terms of distances, but this complication can be circumvented as discussed above.

When the dimensionality decreases the reasoning presented above does not hold anymore. As an example, consider three points and a single dimension. The bias parameters can be transformed to coordinates on a second dimension. However, by doing so, the three points are not necessarily on a line anymore (this can easily be seen when the middle category has the largest bias parameter). However, we conjecture that in most practical cases the effect of the bias parameters is minor. This conjecture will be further studied in Section 5.

4. Degrees of Freedom and Identification for the Model without Bias Parameters

Before we turn our attention to verifying our conjecture, we will discuss the number of independent parameters in the model without bias terms. This is important, since we are going to compare fit measures which must be related to the difference of independent parameters. For the model with bias parameters, the number of independent parameters equals $G-1+(p+G)\times R-R(R+1)$ (Takane et al., 1987).

For the model without bias parameters, the number of parameters is $(p+G)R$. However, there are a number of indeterminacies, such as rotational freedom that do not change the probabilities. Therefore, first consider our probability model (5). The probabilities remain the same when a constant is added for each subject, i.e.,

$$\pi_{g|i}=\frac{\exp(-d_{ig}^2)}{\sum_h \exp(-d_{ih}^2)}=\frac{\exp(-d_{ig}^2+c_i)}{\sum_h \exp(-d_{ih}^2+c_i)}. \tag{7}$$

Since the probabilities in our model are solely based on squared Euclidean distances, we have that a model based on any squared distance matrix $\mathbf{D}_*$ defined as $\mathbf{D}_*=\mathbf{D}+\mathbf{c}\mathbf{1}^T$ provides the same probabilities as the model defined with squared distances $\mathbf{D}$. Suppose $\mathbf{B}$ and $\mathbf{Z}$ give $\mathbf{D}$ and $\mathbf{B}_*$ and $\mathbf{Z}_*$ give $\mathbf{D}_*$, such that $\mathbf{D}_*=\mathbf{D}+\mathbf{c}\mathbf{1}^T$. How are these related? The squared distance matrices can be written as

$$\mathbf{D}=\operatorname{diag}\left(\mathbf{X}\mathbf{B}\mathbf{B}^T\mathbf{X}^T\right)\mathbf{1}^T+\mathbf{1}\left(\operatorname{diag}\left(\mathbf{Z}\mathbf{Z}^T\right)\right)^T-2\mathbf{X}\mathbf{B}\mathbf{Z}^T,$$

where diag$(\cdot)$ takes the diagonal of a matrix and puts it in a vector, and

$$\mathbf{D}_* = \text{diag}\big(\mathbf{X}\mathbf{B}_*\mathbf{B}_*^T\mathbf{X}^T\big)\mathbf{1}^T + \mathbf{1}\big(\text{diag}\big(\mathbf{Z}_*\mathbf{Z}_*^T\big)\big)^T - 2\mathbf{X}\mathbf{B}_*\mathbf{Z}_*^T.$$

Since these two must be equal up to an additive row constant, it follows that

1. diag$(\mathbf{X}\mathbf{B}\mathbf{B}^T\mathbf{X}^T)$ may change without restrictions, since this will be captured in $\mathbf{c}$;
2. diag$(\mathbf{Z}_*\mathbf{Z}_*^T)$ and diag$(\mathbf{Z}\mathbf{Z}^T)$ must be equal up to a constant q, i.e., diag$(\mathbf{Z}_*\mathbf{Z}_*^T) =$ diag$(\mathbf{Z}\mathbf{Z}^T) + q\mathbf{1}$, otherwise it cannot be captured in the $\mathbf{c}\mathbf{1}^T$ term;
3. $\mathbf{X}\mathbf{B}_*\mathbf{Z}_*^T$ must be equal to $\mathbf{X}\mathbf{B}\mathbf{Z}^T + \tilde{\mathbf{c}}\mathbf{1}^T$, then again changes are captured in the $\mathbf{c}\mathbf{1}^T$ term.

If we transform $\mathbf{Z}$ and $\mathbf{B}$ to

$$\begin{aligned}\mathbf{Z}_* &= \mathbf{1}\mathbf{v}^T + \mathbf{Z}\mathbf{T},\\ \mathbf{B}_* &= \mathbf{B}\big(\mathbf{T}^{-1}\big)^T\end{aligned}$$

with $\mathbf{T}$ an $R \times R$ matrix and $\mathbf{v}$ an $R \times 1$ vector, we have

$$\begin{aligned}\mathbf{X}\mathbf{B}_*\mathbf{Z}_*^T &= \mathbf{X}\mathbf{B}\big(\mathbf{T}^{-1}\big)^T\big(\mathbf{1}\mathbf{v}^T + \mathbf{Z}\mathbf{T}\big)^T\\ &= \mathbf{X}\mathbf{B}\mathbf{Z}^T + \mathbf{X}\mathbf{B}\big(\mathbf{T}^{-1}\big)^T\mathbf{v}\mathbf{1}^T\\ &= \mathbf{X}\mathbf{B}\mathbf{Z}^T + \tilde{\mathbf{c}}\mathbf{1}^T\end{aligned}$$

such that (1) and (3) are fulfilled. However, diag$(\mathbf{Z}_*\mathbf{Z}_*^T) =$ diag$(\mathbf{Z}\mathbf{Z}^T) + q\mathbf{1}$ is not necessarily fulfilled by these transformations. We have to explicitly impose these $G - 1$ restrictions on the transformations. From the two transformation equations, we see that

- there are $R(R + 1)$ indeterminacies with $G - 1$ restrictions;
- a rotation is always possible, in that case $\mathbf{v} = \mathbf{0}$ and diag$(\mathbf{Z}\mathbf{Z}^T)$ does not change, such that the minimum number of indeterminacies is $R(R - 1)/2$;
- in dimensionality $R = G - 1$, the number of indeterminacies is $R(R+1) - (G-1) = R^2$: any nonsingular $\mathbf{T}$ can be used and this can be solved by finding an appropriate vector $\mathbf{v}$ such that the restrictions are true

Summarizing, we have R^2 unknowns in $\mathbf{T}$, R in $\mathbf{v}$, but the transformations have $G-1$ restrictions. The number of indeterminacies thus equals $\max(R(R-1)/2, R(R+1) - (G-1))$.

In order to obtain an identified solution, we observe the following. If $\mathbf{J}$ is defined as $\mathbf{J} = \mathbf{I}_G - \mathbf{1}_G\mathbf{1}_G^T/G$, $\mathbf{\Pi}$ as $\mathbf{\Pi} = \{\pi_g(\mathbf{x}_i)\}$, and $\mathbf{\Delta}$ as $\mathbf{\Delta} = \log \mathbf{\Pi}$, we have $-\mathbf{\Delta}\mathbf{J} = \mathbf{D}\mathbf{J} = \mathbf{D}_*\mathbf{J}$, i.e., row wise centering makes solutions equal, and so identifies the solution if parameters can be obtained from these centered distance matrices. In order to do so, metric unfolding with single centering as described in Heiser (1981) can be used. This procedure works as follows. $\mathbf{D}\mathbf{J}$ is a matrix of rank $R + 1$, and can be written as

$$\mathbf{D}\mathbf{J} = \mathbf{1}\boldsymbol{\beta}^T - 2\mathbf{Y}\mathbf{Z}^T \tag{8}$$

where $\boldsymbol{\beta}$ is the sum of squares of the rows of $\mathbf{Z}$ in deviation from their mean. A singular value decomposition of $\mathbf{D}\mathbf{J} = \mathbf{U}_*\mathbf{\Lambda}\mathbf{V}_*^T = \mathbf{U}\mathbf{V}^T$ can be computed where the $R + 1$ nonzero singular values and corresponding vectors are retained. It does matter how the singular values are distributed over $\mathbf{U}_*$ and $\mathbf{V}_*$, we use $\mathbf{U} = \mathbf{U}_*\mathbf{\Lambda}^{1/2}$ and $\mathbf{V} = \mathbf{V}_*\mathbf{\Lambda}^{1/2}$. To obtain an identified solution $\mathbf{U}$ and $\mathbf{V}$ should be transformed such that (8) is true. Therefore, define the $(R + 1) \times (R + 1)$ nonsingular

matrix $\mathbf{R}$ and solve the following system of equations simultaneously

$$\mathbf{UR} = \left[\begin{array}{c|c} \mathbf{1} & -2\mathbf{Y} \end{array}\right],$$

$$\mathbf{VS}^T = \left[\begin{array}{c|c} \boldsymbol{\beta} & \mathbf{Z} \end{array}\right],$$

with $\mathbf{S} = \mathbf{R}^{-1}$. This gives an identified solution. This procedure works fine, except in the situation of maximum dimensionality, i.e., $R = G - 1$ due to the fact that $R + 1$ singular values and vectors have to be retained. In this case (i.e., $R = G - 1$), we identify the solution by a transformation of $\mathbf{Y}$ such that $\mathbf{Y}^T\mathbf{Y} = n\mathbf{I}$ (which can be obtained using a singular value decomposition), and solve for $\mathbf{v}$. In both cases (if $R = G - 1$ and $R < G - 1$), we thereafter solve the rotational indeterminacy by requiring that $b_{jr} = 0$ for $r > j$.

In order to fit the model, the first step is to fit the unidentified model using a quasi-Newton algorithm where the Hessian is computed using a finite difference method. Then the obtained distance matrix is row wise centered and the system of equations is solved. Finally, the solution is rotated. The procedure is implemented in MATLAB (Mathworks, 2006), and uses the MATLAB optimization toolbox for optimization of the likelihood and solving the system of equations. The programs can be obtained from the author upon request.

5. Empirical Verification of the Zero Effect in Lower Dimensionality

In this section, empirical evidence of the conjecture posed in Section 3 is provided. Several empirical data sets will be discussed. The first three data sets have 3, 4, and 5 response classes, respectively; the fourth data set has many (12) response classes. The fifth example has one response class that is very large, while the sixth data set has one class that is really small. For each data set, we discuss the observed marginal proportions and the predictor variables. Then for the models with and without bias terms the deviance value is shown and the difference thereof, given dimensionality 1 through $G - 1$. One should act with caution, however, to use these likelihood ratio statistics for dimensionality selection, since there are indications that these statistics are not chi-squared distributed (Takane, van der Heijden, & Browne, 2003). For the third, fourth, and fifth example, the explanatory variables are categorical. In these cases, the data can be grouped (see Agresti, 2002, Sect. 4.5.3) which results in a different deviance measure. This latter deviance measure can also be used for checking model fit. For the other examples, deviance is based on the individual records (ungrouped) and can only be used to compare nested models. Results for all data sets are shown in Table 1.

The first data set comes from Tabachnik and Fidell (2007, Chap. 9) and includes 465 women who were role-dissatisfied housewives, role-satisfied housewives, or working women with proportions 0.1763, 0.2946, and 0.5290, respectively. There are four explanatory variables: locus of control, satisfaction with marital status, attitude towards women's role, and attitude toward housework. In both, the one-dimensional and two-dimensional solution there is no discernible difference in attained deviance.

The second data set comes from the book by Lattin, Carroll, and Green (2003) and contains information from 141 households from a suburban panel in a Midwestern US market. Each household subscribed to one and only one of the following magazines: Better Homes & Gardens, Reader's Digest, TV Guide, and Newsweek. The proportions of the four magazines in the sample are 0.1844, 0.3475, 0.2766, 0.1915, respectively. Explanatory variables are family size, income,

TABLE 1.

Deviances of models with and without bias terms and their difference. The number of parameters are given in parenthesis, which are for the model with bias terms $G-1+(p+G)R-R(R+1)$ and without $(p+G)R-\max(R(R-1)/2, R(R+1)-(G-1))$. Data sets: [Role] Female role satisfaction data from Tabachnik and Fidell (2007), ($p=4$; $G=3$); [Magazines] Choice of magazine data from Lattin et al. (2003), ($p=10$; $G=4$); [Alligator] Primary food choice of alligators from Agresti (2002), ($p=5$; $G=5$); [Bitterling] Reproductive behavior of the male bitterling from Wiepkema (1961), ($p=11$; $G=12$); [Seat-belt] Car crash data from Agresti (2002), ($p=6$; $G=5$); [DPES] Dutch parliamentary election studies, from Irwin et al. (2003), ($p=5$; $G=4$).

Data	Model	Dimensionality										
		1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D
Role	With	900.95	882.80									
		(7)	(10)									
	Without	900.95	882.80									
		(7)	(10)									
	Difference	0.00	0									
Magazines	With	341.48	321.25	306.98								
		(15)	(25)	(33)								
	Without	341.55	321.25	306.98								
		(14)	(25)	(33)								
	Difference	0.07	0.00	0								
Alligator	With	77.26	59.14	51.71	50.26							
		(12)	(18)	(22)	(24)							
	Without	77.94	59.21	51.71	50.26							
		(10)	(18)	(22)	(24)							
	Difference	0.68	0.08	0.00	0							

TABLE 1.
(*Continued.*)

Data	Model	Dimensionality										
		1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D
Bitterling	With	3290.16	1056.12	405.04	137.388	48.94	27.86	4.26	1.37	0.06	0.03	0
		(32)	(51)	(68)	(83)	(96)	(107)	(116)	(123)	(128)	(131)	(132)
	Without	4896.45	1234.40	426.50	138.94	49.16	27.86	4.28	1.38	0.06	0.03	0
		(23)	(45)	(66)	(83)	(96)	(107)	(116)	(123)	(128)	(131)	(132)
	Difference	1606.28	178.28	21.45	1.56	0.22	0.00	0.01	0.01	0.00	0.00	0
Seat-belt	With	101.31	31.88	15.38	10.84							
		(13)	(20)	(25)	(28)							
	Without	192.00	33.16	15.41	10.84							
		(11)	(20)	(25)	(28)							
	Difference	90.69	1.28	0.03	0							
DPES	With	910.10	801.92	793.72								
		(10)	(15)	(18)								
	Without	910.28	801.92	793.72								
		(9)	(15)	(18)								
	Difference	0.18	0.00	0								

race, number of TV sets, newspaper subscription, missing male or female head of household, children, age, and education (see Lattin et al., 2003, Table 12.6). Since there are four magazines, the dimensionality runs from one to three. Table 1 shows the difference of deviances in each dimensionality, which support our conjecture.

The third data set comes from Chapter 7 of Agresti (2002) and considers the primary food choice of alligators. This response variable has five categories: fish, invertebrate, reptile, bird, and other with proportions are 0.4292, 0.2785, 0.0868, 0.0594, and 0.1461, respectively. There are three categorical explanatory variables, lake (4 categories), gender, and size (two categories). Table 1 shows the difference of deviances in each dimensionality, which again support our conjecture. For the two-dimensional model, we would expect the deviances to be the same since both models gave the same number of independent parameters. We might have ended in a local optimum. Fifty random starts, a start from a correspondence analysis, and a start using the regression parameters and class points from the two-dimensional model with bias terms did not yield a better deviance, however.

The fourth data set was analyzed by De Rooij and Heiser (2005) and is a transition frequency table of 12×12 describing reproductive behavior of male bitterlings (Wiepkema, 1961). Behavioral categories (proportions) are jerking (0.1661), turning beats (0.0525), head butting (0.0978), chasing (0.0663), fleeing (0.0724), quivering (0.1977), leading (0.0451), head down posture (0.1292), skimming (0.0533), snapping (0.0749), chafing (0.0258), and finflickering (0.0188). The previous behaviors (rows of the transition frequency table) serve here as categories of a single explanatory variable for the current behavior (columns) and were transformed using dummy-coding. Results can be found in Table 1 where it can be seen that for low dimensional models (one to three) the difference between the models with and without bias terms is substantial, in higher dimensionalities, the difference is ignorable. Notice that for the dimensionalities where the models differ in fit statistics, the deviance points out that neither the model with nor the model without bias terms fits the data adequately (degrees of freedom are 144 minus the number of independent parameters). For the four and five-dimensional models, the same comment as for the two-dimensional alligator solutions applies: we expected the same deviance for the models with and without bias terms here, but many analyses did not yield them. It seems that the model without bias parameters has some difficulties in finding the global optimum of the likelihood function.

The fifth data set comes from Agresti (2002) and has one very large class. The data deals with injuries after a car crash having five categories: Not injured (0.9087); injured but not transported by emergency medical services (0.0131); injured and transported by emergency medical services (0.0649); injured and hospitalized, but did not die (0.0113); injured and died (0.0020). Although the response variable has ordered categories, we do not use that information here. Explanatory variables are gender, location (urban/rural), and seat belt use (yes/no). We included also the pairwise interactions between the explanatory variables as predictors. In Table 1, we see that in all dimensionalities except the one-dimensional model the bias parameters can be removed without considerable loss. Looking at the deviances of both one-dimensional models, we see that they do not fit the data, however (degrees of freedom equal 40 minus the number of independent parameters). For the two-dimensional model, we expected the same deviance for the model with and without bias terms.

In order to also show a data set with one very small response class, we created a data set from the Dutch parliamentary election studies 2002–2003 (Irwin, Van Holsteyn, & Den Ridder, 2003). We created a data set of 629 subjects that either voted in 2003 for one of the three large political parties in the Netherlands PvdA, CDA, and VVD (proportions in data set 0.3911, 0.3831, and 0.2162, respectively) or a very small party SGP, with proportion 0.0095. There are five explanatory variables self left-right scaling, age, sex, religious denomination, and social class. Table 1 shows that in all dimensionalities there is no considerable difference in fit.

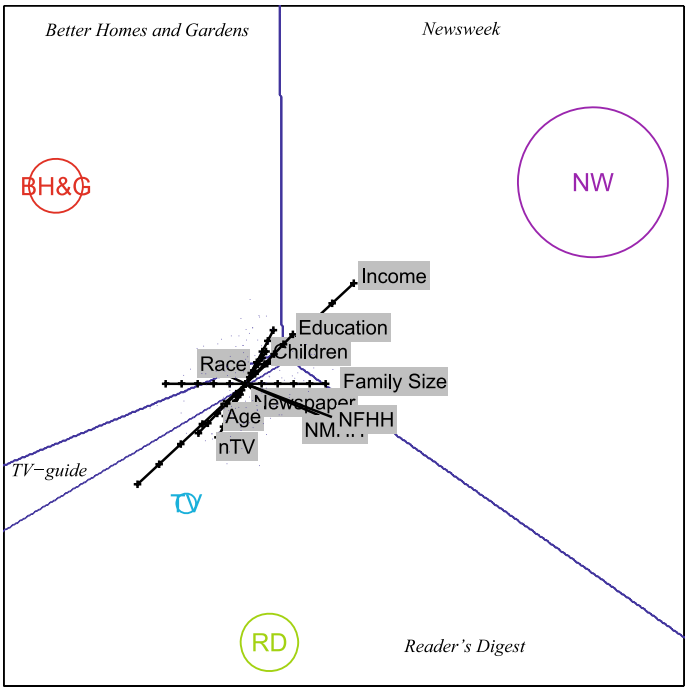

FIGURE 3.
Result of the model with bias parameters for the magazines data. The response categories are labeled by BH&G (Better Homes and Gardens), RD (Reader's Digest), TV (TV-Guide), and NW (Newsweek). Bias terms are represented by the area of the circle. Also shown are lines between prediction regions with in the regions the name of which category has the highest odds. Notice that TV-Guide is outside it's own prediction region. Explanatory variables are family size, income, race, number of TV sets (nTV), newspaper subscription, missing male (NMHH) or female head of household (NFHH), children, age, and education.

In some cases, as noted above, the models with and without bias terms differ in the deviance values, but have the same number of independent parameters. In all cases, the deviance is smaller for the model with bias parameters. This difference is probably due to suboptimal solutions for the model without bias terms. In all cases, we did a smart start using correspondence analysis, fifty random starts, and a start from the solution of the model with bias parameters. It seems that the model without bias terms and with only categorical explanatory variables is somewhat more difficult to fit. In all cases, the differences are not very large.

Now we have showed that the models with and without bias parameters differ not much in fit, we will show some graphical results. In Figures 3 and 4, we show the results for the magazines data in two dimensions with and without bias terms. In Figures 5 and 6, we show the results of the alligator data in two dimensions. The figures show class points, explanatory variables, and prediction regions. Prediction regions are areas in which the predicted odds are in favor of a specific class.

Comparing the representations of the models with and without bias parameters, it can be seen that for the models without bias parameters the class points always lie in the interior of their own prediction region and decision boundaries are exactly in the middle of two class points, i.e., $\pi_{g|i}$ is inversely monotonic with d_{ig}^2. This is not true for the model with bias terms. In the model with bias terms, a subject can have an ideal point right on top of a class point and still have a higher odds for another class.

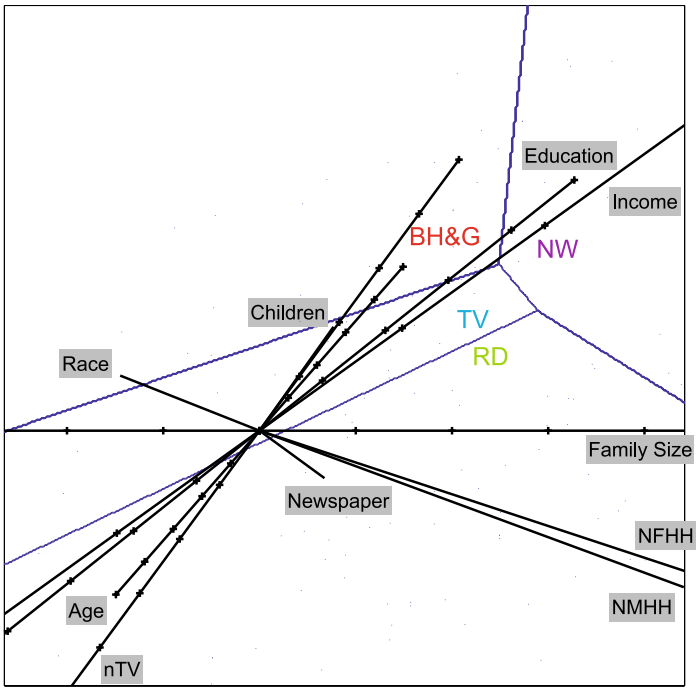

FIGURE 4.
Result of the model without bias parameters for the magazines data. The response categories are labeled by BH&G (Better Homes and Gardens), RD (Reader's Digest), TV (TV-Guide), and NW (Newsweek). Also shown are lines between prediction regions. Explanatory variables are Family size, income, race, number of TV sets (nTV), newspaper subscription, missing male (NMHH) or female head of household (NFHH), children, age, and education.

More specifically, comparing Figures 3 and 4, the deviances of the models underlying these figures are equal as well as the number of independent parameters. The interpretation of 4 is, however, much simpler since for every subject the highest probability of a certain magazine is given by the closest class point. Contrarily, in Figure 3, a subject can be very close to TV-Guide, but has the highest probability for Readers Digest.

Similar remarks apply to Figures 5 and 6. In Figure 5, the problem is even stronger: the "other" category is nowhere predicted and "bird" is only predicted at the boundary of the display. This concurs with the discrepancy as noted by Takane (1998), the conditional probabilities $\pi_{g|i}$ are not inversely monotonic with d_{ig} when the bias parameters are unequal. In Figure 6, this cannot occur: if an alligator is on top of the "other" class, then it has the highest probability for this class.

In Section 3, it was shown that in maximum dimensionality the distances between the class points in the model without bias terms is larger than those distances in the model with bias terms. In Figures 3, 4 and 5, 6, it can be seen that this is not necessarily true for models in lower dimensionality. For example, for the magazines data in the model with bias terms the class points are well spread with the variables (and thus the ideal points) in between, while for the model without bias terms the class points and variables are better mixed. For the alligator data, it is the other way around. In the results for the model with bias terms, the variables and class points are well mixed, while in the model without bias terms, the class points are somewhat on the boundary.

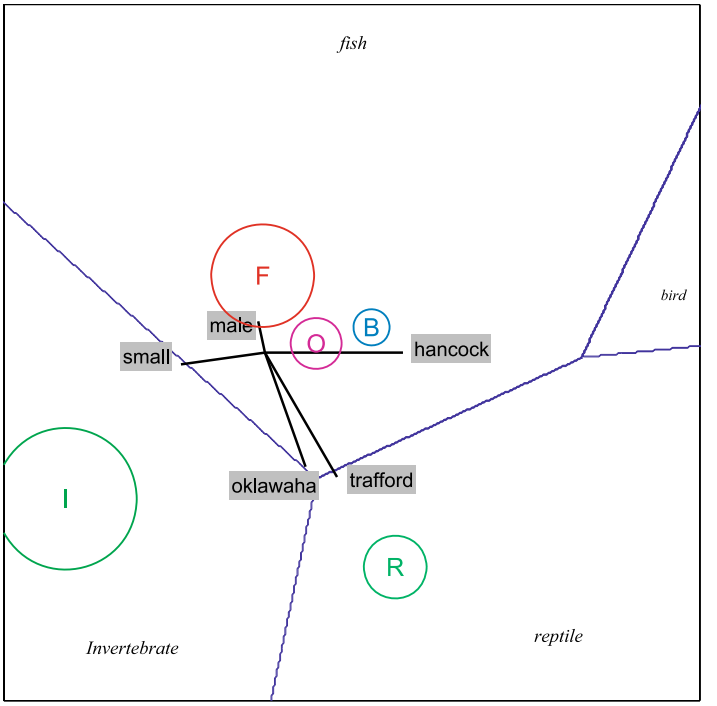

FIGURE 5.
Result of the model with bias parameters for the Alligator data. The response categories are labeled by F (fish), I (invertebrates), R (reptiles), B (bird), and O (other). Bias terms are represented by the area of the circle. The origin refers to large female alligators in Lake George. The variables give the displacement for being small, male, or living in another lake. Also shown are lines between prediction regions with in the regions the name of which category has the highest odds. Notice that there is no place where "other" gets predicted.

6. Conclusion

Ideal point discriminant analysis is a classification tool based on multidimensional scaling techniques. The model looks very much the same as the canonical discriminant analysis tool, however, it does not assume multivariate normality of the explanatory variables. However, as discussed in Takane (1998), the interpretation of IPDA is hampered by the bias terms in this model. The model without bias parameters has a much clearer interpretation, since the decision boundaries are based on distances only, and are thus orthogonal to the line joining two class points and through their centroid, while in the case the model includes bias parameters the decision boundaries shift away from the class with largest bias term.

We showed that in maximum dimensionality the bias terms have a zero effect in case the class point are estimated, i.e., the model without bias effect provides the same fit to the data. This is an important finding since the model without bias terms has an easier interpretation. Moreover, both the model with and the model without bias terms provide the same fit to the data as the multinomial logit model when fitted using the maximum dimensionality.

For reduced dimensionality, it was conjectured and illustrated that in general the effect of the bias parameters is small. There are a few exceptions to this rule. The first is when the response variable has many categories, in that case, it pays of to use bias parameters in low dimensional models. The second case is when the response variable has a category that dominates the other categories, i.e., a category that takes the vast majority of the responses. If the bias parameters are important for a one- or two-dimensional model, these bias parameters could be represented

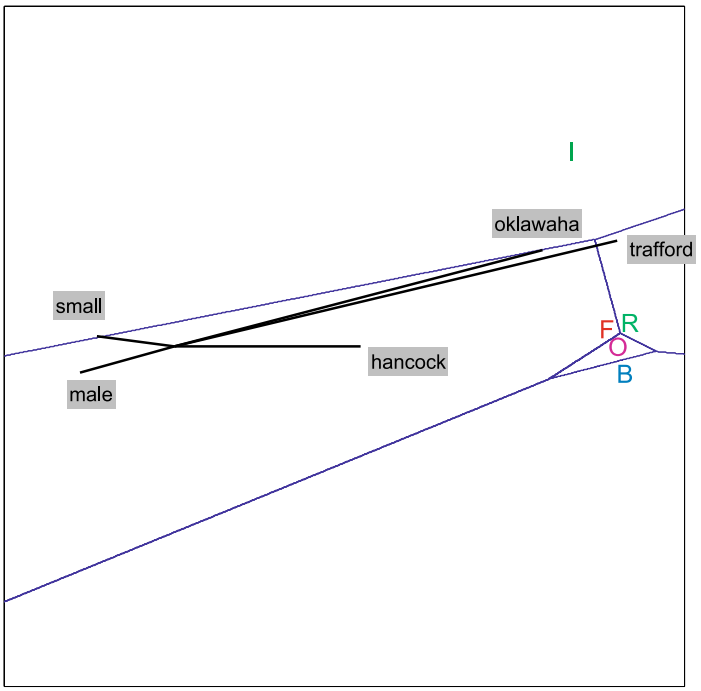

FIGURE 6.
Result of the model without bias parameters for the Alligator data. The response categories are labeled by F (fish), I (invertebrates), R (reptiles), B (bird), and O (other). The origin refers to large female alligators in Lake George. The variables give the displacement for being small, male, or living in another lake. Also shown are lines between prediction regions.

as an extra dimension (as shown in Section 3). The graphical display in that case has a clear interpretation solely based on distances again.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.
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*, 99–123.
Gower, J.C., & Hand, D.J. (1996). *Biplots*. London: Chapman and Hall.
Heiser, W.J. (1981). *Unfolding analysis of proximity data*. Unpublished doctoral dissertation, Leiden University.
Irwin, G.A., Van Holsteyn, J.J.M., & Den Ridder, J.M. (2003). *Dutch parliamentary election study 2002–2003: An enterprise of the foundation for electoral research in the Netherlands (skon)*. Steinmetz Archives, Amsterdam (P1628).
Lattin, J., Carroll, J.D., & Green, P.E. (2003). *Analyzing multivariate data*. Toronto: Thomson learning: Brooks/Cole.
Mathworks (2006). *Matlab: the language of technical computing*. Natick: Mathworks.
Tabachnik, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.
Takane, Y. (1998). Visualization in ideal point discriminant analysis. In J. Blasius, & M.J. Greenacre (Eds.), *Visualization of categorical data* (pp. 441–459). New York: Academic Press.
Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point discriminant analysis. *Psychometrika*, *52*, 371–392.
Takane, Y., van der Heijden, P., & Browne, M. (2003). On likelihood ratio tests for dimensionality selection. In T. Higuchi, Y. Iba, & M. Ishiguro (Eds.), *Proceedings of science of modeling: The 30th anniversary meeting of the information criterion (AIC)* (pp. 348–349). Tokyo: The Institute of Statistical Mathematics.
Wiepkema, P.R. (1961). An ethological analysis of the reproductive behavior of the bitterling (Rhodeus amarus Bloch). *Archives Neerlandais Zoologique*, *14*, 103–199.

Manuscript Received: 12 APR 2007
Final Version Received: 24 NOV 2008
Published Online Date: 14 JAN 2009