The Distance Perspective of Generalized Biadditive Models: Scalings and Transformations

Mark DE ROOIJ

Recently two articles studied scalings in biplot models, and concluded that these have little impact on the interpretation. In this article again scalings are studied for generalized biadditive models and correspondence analysis, that is, special cases of the general biplot family, but from a different perspective. The generalized biadditive models, but also correspondence analysis, are often used for Gaussian ordination. In Gaussian ordination one takes a distance perspective for the interpretation of the relationship between a row and a column category. It is shown that scalings—but also nonsingular transformations—have a major impact on this interpretation. So, depending on the perspective one takes, the inner product or distance perspective, scalings and transformations do have (distance) or do not have (inner-product) impact on the interpretation. If one is willing to go along with the assumption of the author that diagrams are in practice often interpreted by a distance rule, the findings in this article influence all biplot models.

Key Words: Correspondence analysis; Euclidean distance; Gaussian ordination; Multidimensional scaling; Multidimensional unfolding;

1. INTRODUCTION

Recently Gabriel (2002) studied the geometry of biplots, especially the distribution of the singular values over the row and column points (i.e., "scalings"). He concluded that although theoretically these different scalings provide different insights, in practice these have little effect on the interpretation. Gower (2004) extended this work geometrically. We will again look at scalings in biplot models but from a different perspective—the ordination or distance perspective. That is, a perspective where the distances between row and column points are interpreted instead of their inner-product form. We will confine ourselves to the (generalized) biadditive models, since these models can be interpreted from an ordination perspective, that is, these models can explicitly be redefined in terms of distances.

We have two reasons for looking at these models from an ordination perspective:

Mark de Rooij is Assistant Professor at the Leiden University Institute for Psychological Research—Methodology and Statistics Unit, P.O. Box 9555, 2300 RB Leiden, The Netherlands (E-mail: *rooijm@fsw.leidenuniv.nl*).

Journal of Computational and Graphical Statistics, Volume 16, Number 1, Pages 210–227
DOI: 10.1198/106186007X180101

1. The log-biadditive model (defined below), but also correspondence analysis, is often used for Gaussian ordination.

2. The author believes that biadditive models, and more generally all biplot models, are often interpreted in practice from an ordination perspective.

Biadditive models for two-way tables (Kempton 1984; Gauch 1992; Gower and Hand 1996) have been useful tools in biological as well as sociological studies. The parameters of biadditive models are usually estimated assuming that the variable of interest is normally distributed and has constant variance. Several authors generalized biadditive models by including a link function and an error distribution from the exponential family (see De Falguerolles and Francis 1992, 1994; Van Eeuwijk 1995), as is done in generalized linear models (McCullagh and Nelder 1989). This generalized biadditive model (GBM) for an $I \times J$-table is defined by

$$g(\mu_{ij}) = g_{ij} = m + a_i + b_j + \sum_{r=1}^{R} \phi_r c_{ir} d_{jr}, \tag{1.1}$$

where μ_{ij} $(i = 1, \ldots, I, j = 1, \ldots, J)$ are the model values given some error distribution from the exponential family, $g(\cdot)$ is a link function, m is a constant, a_i and b_j are the main effect terms, c_{ir} and d_{jr} $(r = 1, \ldots, R)$ are normalized row and column parameters, respectively, and ϕ_r is sometimes called an "intrinsic association" parameter (Goodman 1991). If $R < \min(I - 1, J - 1)$ then the model provides a reduced rank approximation of the usual interaction term for two-way tables. Often the interest is in small R, that is, $R \leq 2$, such that the model can be graphically represented. Algorithms to fit GBMs were given by De Falguerolles and Francis (1992, 1994) and for the log-biadditive model by Becker (1990). We shall find it convenient to present model (1.1) in matrix form, that is,

$$g(\boldsymbol{\mu}) = \mathbf{G} = m\mathbf{1}\mathbf{1}^T + \mathbf{a}\mathbf{1}^T + \mathbf{1}\mathbf{b}^T + \mathbf{C}\boldsymbol{\Phi}\mathbf{D}^T. \tag{1.2}$$

At first, the usual identification constraints will be applied, that is, $\mathbf{a}^T\mathbf{1} = \mathbf{b}^T\mathbf{1} = 0$, $\mathbf{1}^T\mathbf{C} = \mathbf{1}^T\mathbf{D} = \mathbf{0}$, $\mathbf{C}^T\mathbf{C} = \mathbf{I}$, $\mathbf{D}^T\mathbf{D} = \mathbf{I}$ and $\boldsymbol{\Phi}$ is a diagonal matrix with ordered elements $\phi_1 > \phi_2 > \cdots > \phi_R$. Let $\mathbf{X}_\tau = \mathbf{C}\boldsymbol{\Phi}^\tau$ and $\mathbf{Y}_\kappa = \mathbf{D}\boldsymbol{\Phi}^\kappa$, such that $\tau + \kappa = 1$, which gives

$$g(\boldsymbol{\mu}) = \mathbf{G} = m\mathbf{1}\mathbf{1}^T + \mathbf{a}\mathbf{1}^T + \mathbf{1}\mathbf{b}^T + \mathbf{X}_\tau\mathbf{Y}_\kappa^T. \tag{1.3}$$

The elements of $\mathbf{X}_\tau$ will be denoted by $x_{ir}^{(\tau)}$, and in a similar fashion $y_{jr}^{(\kappa)}$. Model values in (1.3) do not change under nonsingular transformations of the row and column scores, that is, we may rewrite (1.3) as

$$\begin{aligned} g(\boldsymbol{\mu}) = \mathbf{G} &= m\mathbf{1}\mathbf{1}^T + \mathbf{a}\mathbf{1}^T + \mathbf{1}\mathbf{b}^T + \mathbf{X}_\tau\mathbf{T}\mathbf{T}^{-1}\mathbf{Y}_\kappa^T \\ &= m\mathbf{1}\mathbf{1}^T + \mathbf{a}\mathbf{1}^T + \mathbf{1}\mathbf{b}^T + \tilde{\mathbf{X}}_\tau\tilde{\mathbf{Y}}_\kappa^T, \end{aligned} \tag{1.4}$$

for any nonsingular $R \times R$-matrix $\mathbf{T}$.

Ihm and Groenewoud (1984), but also Takane (1987) and De Rooij and Heiser (2005) in another context, show an equivalence relationship between the log-biadditive model and a

Table 1. An Example Dataset

	C1	*C2*	*C3*	*Total*
E1	5	7	2	14
E2	18	46	20	84
E3	19	29	39	87
E4	12	40	49	101
E5	3	7	16	26
Total	57	129	126	312

variant of Gaussian ordination that is attractive in case the sites vary in size. Their Gaussian ordination model is

$$\mu_{ij} = \alpha_i \beta_j \exp\left(-\frac{(c_i - d_j)^2}{2t^2}\right), \tag{1.5}$$

which after a logarithmic transformation and working out the distance terms is equal to model (1.1) where $1/2t^2$ plays the same role as ϕ. Thus, the log-biadditive model has an ordination interpretation, which can easily be generalized to the GBM (see Section 2). Ter Braak (1987, chap. 1) showed that a first-order Taylor approximation to the log-biadditive model leads to correspondence analysis, from which he concluded that correspondence analysis can be used for Gaussian ordination.

Note that in the derivations above the values in $\mathbf{X}_\tau$ and $\mathbf{Y}_\kappa$ did not change. These are the values normally represented in a joint display to interpret the interaction.

In this article special attention will be given to the distance interpretation of the GBM and we will be especially interested in the case $R = 2$, since in that case the model can be represented graphically. The influence of different scalings (τ) (we will denote scalings by τ instead of τ/κ since $\kappa = 1 - \tau$) and the nonsingular transformation ($\mathbf{T}$) on the distance representation will be studied. It is easy to see that the inner-products in all models above are equal, that is, scalings and transformation have no influence on these values. However, distances based on $\mathbf{X}_\tau \mathbf{T}$ and $\mathbf{Y}_\kappa (\mathbf{T}^{-1})^T$ are not equal for different scalings (τ) and transformations ($\mathbf{T}$).

A 5×3 data example will be used throughout to illustrate the methodology. The data are given in Table 1 and were analyzed by Greenacre and Hastie (1987) using correspondence analysis. Here we will use the log-biadditive model for illustration. The sample consists of 312 people classified into five educational groups (E1, some primary school; E2, primary school completed; E3, some secondary school; E4, secondary school completed; E5, some tertiary school) and three categories of readership of the newspaper (C1, glance; C2, fairly thorough; C3, very thorough). The advantage of this example is that the data can be represented without abstraction in a graphical display. We give a detailed discussion of the substantive content of this application. The row categories could as well be sites and the column categories species, or any other example of interest.

Table 2. Parameter Estimates of the Two-Component Model

	Main effects $m = 2.66$	*scores-1* $\phi_1 = 1.56$	*scores-2* $\phi_2 = 0.44$
E1	−1.24	−0.71	−0.21
E2	0.58	−0.29	0.50
E3	0.67	0.12	−0.62
E4	0.70	0.34	0.53
E5	−0.72	0.54	−0.20
C1	−0.45	−0.53	−0.62
C2	0.30	−0.27	0.77
C3	0.15	0.80	−0.15

The parameter estimates of the log-biadditive model (i.e., model (1.1) with log-link) with two components are given in Table 2. The two-component model gives an exact representation of the data, that is, the expected frequencies are equal to the observed frequencies.

The remainder of the article is organized as follows. Section 2 discusses graphical displays for the GBM, with a focus on the distance interpretation. Section 3 studies different scalings and transformations. Two issues will be emphasized:

1. Minimizing the constant to maximally draw a distinction between the distances.
2. Minimizing the correlation between distances and expected data, to increase the interpretability.

For correspondence analysis the first issue is implicit, since one never looks at these main terms and the constant. The second issue, however, is as valid for correspondence analysis as it is for the GBM. This article concludes with discussion of the obtained results.

2. GRAPHICAL REPRESENTATIONS

For the inner-product parameterization an additive decomposition into main and interaction effects exists. For the distance parameterization a similar decomposition exists, but since it differs from the decomposition in the inner-product representation they will be called *unique effects* and *common effects*. These unique and common effects may be represented in two separate graphs, in which case the squared distances of the two plots have to be added to obtain model values.

First the transformation of the GBM towards a distance model is recapitulated and some notation introduced

$$\begin{aligned} g(\boldsymbol{\mu}) = \mathbf{G} &= m\mathbf{1}\mathbf{1}^T + \mathbf{a}\mathbf{1}^T + \mathbf{1}\mathbf{b}^T + \mathbf{X}_\tau \mathbf{Y}_\kappa^T \\ &= m\mathbf{1}\mathbf{1}^T + (\mathbf{a} + \mathbf{s}_x)\mathbf{1}^T + \mathbf{1}(\mathbf{b} + \mathbf{s}_y)^T - \frac{1}{2} d^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa) \\ &= m\mathbf{1}\mathbf{1}^T + \mathbf{m}\mathbf{1}^T + \mathbf{1}\mathbf{n}^T - \frac{1}{2} d^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa), \end{aligned} \tag{2.1}$$

where $\mathbf{s}_x = \frac{1}{2}\text{diag}(\mathbf{X}_\tau \mathbf{X}_\tau^T)$, and diag($\mathbf{Z}$) takes the diagonal elements of $\mathbf{Z}$ and puts them in a vector, $\mathbf{s}_y = \frac{1}{2}\text{diag}(\mathbf{Y}_\kappa \mathbf{Y}_\kappa^T)$, $\mathbf{m} = \mathbf{a} + \mathbf{s}_x$, $\mathbf{n} = \mathbf{b} + \mathbf{s}_y$, and $d^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa)$ is the matrix with squared Euclidean distances given by

$$d^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa) = \text{diag}(\mathbf{X}_\tau \mathbf{X}_\tau^T)\mathbf{1}^T + \mathbf{1}\text{diag}(\mathbf{Y}_\kappa \mathbf{Y}_\kappa^T)^T - 2\mathbf{X}_\tau \mathbf{Y}_\kappa^T, \tag{2.2}$$

whose elements are

$$d_{ij}^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa) = \sum_{r=1}^{R} \left(x_{ir}^{(\tau)} - y_{jr}^{(\kappa)} \right)^2. \tag{2.3}$$

To obtain the unique effect dimensions a vector $\mathbf{u}$ is created such that $\mathbf{u} = [u_1, u_2, \ldots, u_I]^T$ where $u_i = \sqrt{2|m_i - \max_i(m_i)|}$, and a vector $\mathbf{v}$ is created such that $\mathbf{v} = [v_1, v_2, \ldots, v_J]^T$ where $v_j = \sqrt{2|n_j - \max_j(n_j)|}$ and $m^* = m + \max_i(m_i) + \max_j(n_j)$. Equation (2.1) can now be rewritten as (De Rooij and Heiser 2003, 2005)

$$\begin{aligned} g(\boldsymbol{\mu}) = \mathbf{G} &= m\mathbf{1}\mathbf{1}^T + \mathbf{m}\mathbf{1}^T + \mathbf{1}\mathbf{n}^T - \frac{1}{2}d^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa) \\ &= m^*\mathbf{1}\mathbf{1}^T - \frac{1}{2}d^2(\mathbf{u}; \mathbf{0}) - \frac{1}{2}d^2(\mathbf{0}; \mathbf{v}) - \frac{1}{2}d^2(\mathbf{X}_\tau; \mathbf{Y}_\kappa) \\ &= m^*\mathbf{1}\mathbf{1}^T - \frac{1}{2}d^2\left([\mathbf{u}, \mathbf{0}, \mathbf{X}_\tau]; [\mathbf{0}, \mathbf{v}, \mathbf{Y}_\kappa]\right). \end{aligned} \tag{2.4}$$

Thus, the GBM has been transformed to a parameterization in terms of a squared Euclidean multidimensional scaling model. The scaling model consists of unique effect dimensions for the rows ($\mathbf{u}$) and for the columns ($\mathbf{v}$) and common effect dimensions ($\mathbf{X}_\tau$ and $\mathbf{Y}_\kappa$). This model gives a distance approximation to the expected values. The larger the distance the more is subtracted from m^*, which thus indicates a maximum (last line of (2.4)).

2.1 Unique Effect Display

The unique effects can be represented as a two-dimensional plot, where one dimension pertains to the unique effects of the row categories and the other dimension pertains to the unique effects of the column categories. The squared distance of a point towards the origin corresponds to the value that needs to be subtracted from m^* for the category that is represented by that point. Because the model is formulated in squared Euclidean distances, dimensions are additive, so the value that needs to be subtracted from m^* for the combination E1 with C2 is by Pythagoras theorem the squared distance from E1 to C2. A plot of the unique effect dimensions using scaling $\tau = 0.5$ is shown in the left-hand panel of Figure 1. From this figure we see that Educational group E4 is most prevalent as well as readership category C3 (this "prevalence" is a conditional measure; that is, conditional on the distances in the common effect dimensions the distance from the origin to the point refers to the prevalence of that category). Educational groups E1 and E5 are least frequent and few people read the newspaper at glance (C1). We will see in Section 3 that, for different scalings and transformations, the unique effect dimensions differ quite a lot, not only in magnitude but also in ordering.

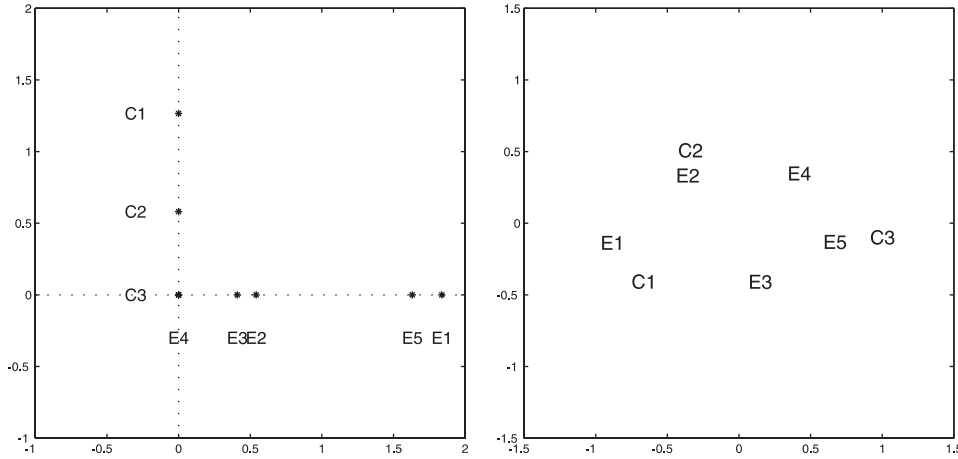

Figure 1. Unique (left) and common (right) effect dimensions using $\tau = 0.5$. The value of m^* equals 4.16. The correlation between squared distances and expected data equals $r = -0.48$.

2.2 COMMON EFFECT DISPLAY

The common effect dimensions are simple Euclidean distance spaces which can be interpreted like any map. The graphical representation with scalings $\tau = 0.5$ is shown in the right-hand panel of Figure 1. In this display it is clear that educational group E1 is close to C1, E2 is closest to C2. Educational group E3 is in between reading very thorough and at glance, while E4 is in between fairly thorough and very thorough. Group E5 is closest to C3.

In the inner-product parameterization, the distances between the points with coordinates $\mathbf{X}_1$ ($= \mathbf{C\Phi}$) are equal to the distances between the rows of $\mathbf{G} - \mathbf{a1}^T$, and the distances between the points with coordinates $\mathbf{Y}_1$ ($= \mathbf{D\Phi}$) are equal to the distances between the columns of $\mathbf{G} - \mathbf{1b}^T$. In the distance parameterization, however, the distances between the points with coordinates $\mathbf{X}_1$ do not equal the distances between the rows of $\mathbf{G} - \mathbf{m1}^T$, and similarly for distances defined by $\mathbf{Y}_1$ and the columns of $\mathbf{G} - \mathbf{1n}^T$.

3. SCALINGS AND TRANSFORMATIONS

This section shows the effect of scalings and transformations on the outcome of an analysis. Scalings and transformations can be used to optimize some criterion. Two criteria will be discussed here: The first is minimizing the constant (m^*) in Equation (2.4), which contributes to maximal differentiation between distances in the display, and so provides maximum insight into the structure of the data. Second, we will minimize the correlation between (expected) data and squared distances in the common dimensions, since from the definition in (2.1) it is clear that a monotone decreasing relation is sought for. Before doing so we will further look into scalings and transformations from a mathematical perspective which will result in four more or less hierarchical conditions for a given criterion. MATLAB programs to make the contour plots and to minimize the functions presented in this article can be obtained from the author.

3.1 Mathematical Analysis

In the previous sections we have seen that the representation of the GBM is not unique: scalings and transformations are possible and do affect the distances. It may seem that scalings and transformations are very different but in fact they are very similar.

The effect of a scaling is stretching/shrinkage of a dimension, which can also be accomplished by a transformation with a diagonal matrix $\mathbf{T}$.

However, a scaling is not equal to a transformation with $\mathbf{T} = t\mathbf{I}$, which can be seen by writing

$$\begin{aligned} \mathbf{c}_i \mathbf{\Phi}^\tau &= [c_{i1}\phi_1^\tau, c_{i2}\phi_2^\tau] \\ &= [c_{i1}\exp(\tau \times \log \phi_1), c_{i2}\exp(\tau \times \log \phi_2)] \\ &= [c_{i1}t_{11}, c_{i2}t_{22}] \\ &= \mathbf{c}_i \mathbf{T}, \end{aligned} \tag{3.1}$$

where $\mathbf{T}$ is a diagonal matrix. In other words, a scaling can be rewritten to a transformation with diagonal $\mathbf{T}$ where the diagonal elements are functionally related to each other: $t_{11} = \exp(\tau \times \log \phi_1)$ and $t_{22} = \exp(\tau \times \log \phi_2)$. This functional relation can be canceled out if the scaling parameter is defined for each dimension, that is, τ_r. Such a differential scaling can also be defined in terms of a diagonal matrix $\mathbf{T}$, where $t_{rr} = \exp(\tau_r \times \log \phi_r)$. This differential scaling can be restricted in two ways, namely by $t_{rr} = t_{r'r'}$ or by $\tau_r = \tau_{r'}$ for all $r \neq r'$. To summarize, any criterion can be minimized under four conditions:

1. $\mathbf{T} = t\mathbf{I}$, which will be called scaling-t;
2. $\mathbf{T}$ is diagonal with $t_{rr} = \exp(\tau \times \log \phi_r)$, which will be called scaling-τ;
3. $\mathbf{T}$ diagonal, which will be called differential scaling; and
4. $\mathbf{T}$ nonsingular, which will be called transformation.

Condition 4 is the most general condition, the other conditions can be obtained by putting constraints on this one. Similarly, Conditions 1 and 2 can be obtained by putting constraints on Condition 3. This hierarchy of conditions shows that it would not make sense to optimize over τ and nonsingular $\mathbf{T}$ simultaneously.

What is the effect of a transformation by $\mathbf{T}$? If we write $\mathbf{T} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ the singular value decomposition of $\mathbf{T}$ with $\mathbf{U}^T\mathbf{U} = \mathbf{I} = \mathbf{V}^T\mathbf{V}$ and decreasing singular values in $\mathbf{\Lambda}$, we see $\mathbf{XT}$ is first a rotation of $\mathbf{X}$ by $\mathbf{U}$, then a stretching/shrinkage of the dimensions by $\mathbf{\Lambda}$ and finally a rotation by $\mathbf{V}$. The latter rotation is redundant since it does not change the distances. We see that such a transformation by $\mathbf{T}$ is in fact a differential scaling, but after a rotation. So, in Conditions 1 to 3 the principal axes are stretched/shrunken, whereas in Condition 4 first a rotation is performed and then the axes are stretched/shrunken.

3.2 Minimizing the Value of m^*

We start this section by looking at the influence of differential scalings, that is, Condition 3 in terms of τ_r, on the value of the constant m^*. Using a contour plot this influence is easily

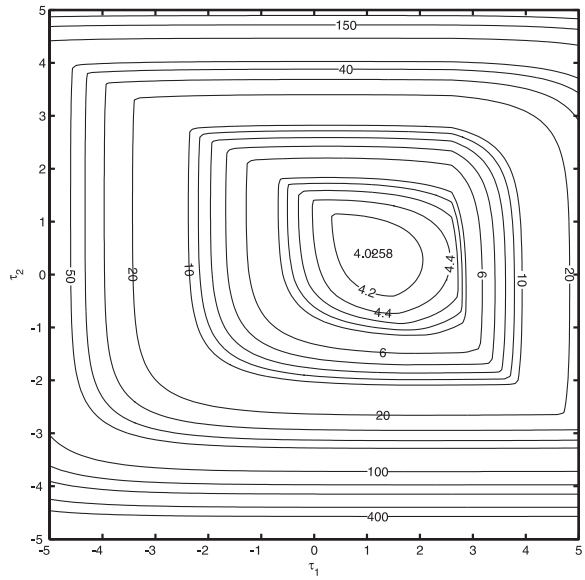

Figure 2. Contour plot of the value of the constant (m^*) as a function of τ_r.

visualized. In Figure 2 we see that the value of m^* varies from about 4 to about 400 for values τ_r in between -5 and 5. Moreover, the scalings in the second dimension τ_2 have a larger impact than the scalings of the first dimension

To obtain the exact minimum value of the constant under the four conditions a minimization function can be defined. To facilitate notation we will use $\mathbf{x}_i$ for $\mathbf{x}_i^{(0.5)}$, and similarly $\mathbf{y}_j$ for $\mathbf{y}_j^{(.5)}$ in the remainder, unless stated otherwise. For Conditions 1, 3, and 4 the loss function is

$$\arg\min_{\mathbf{T}\in\,\Omega}\left\{\max_i\left[a_i+\mathbf{x}_i^T\mathbf{T}\mathbf{T}^T\mathbf{x}_i\right]+\max_j\left[b_j+\mathbf{y}_j^T\mathbf{S}\mathbf{S}^T\mathbf{y}_j\right]\right\}, \tag{3.2}$$

where $\mathbf{S}=(\mathbf{T}^{-1})^T$ and Ω represents the feasible region. For Condition 2 the loss function is

$$\arg\min_{\tau\in\,\Re}\left\{\max_i\left[a_i+\sum_{r=1}^{R}\phi_r^{2\tau}c_{ir}^2\right]+\max_j\left[b_j+\sum_{r=1}^{R}\phi_r^{(2-2\tau)}d_{jr}^2\right]\right\}. \tag{3.3}$$

Both functions can be minimized using a Sequential Quadratic Programming method (see Gill, Murray, and Wright 1981, pp. 237–242) as implemented in, for example, the MATLAB Optimization Toolbox. This works well, and the experience of the author is that there is no trouble with local minima. For Condition 4 the value of $\mathbf{T}$ may differ because of the rotation by $\mathbf{V}$. This can be solved using the singular value decomposition of $\mathbf{T}$ and then redefine $\mathbf{T}=\mathbf{U}\mathbf{\Lambda}$.

3.2.1 Scaling-*t*

The value obtained for m^* is $\hat{m}^*=4.05$ when $\hat{t}=1.36$. The solution is shown in Figure 3 where we see in the unique effect plot (left-hand side) that Educational group E4 is most

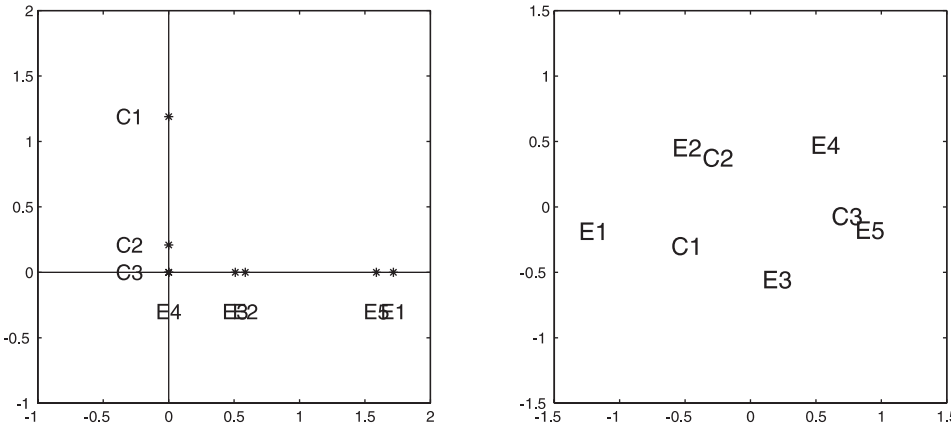

Figure 3. The solution for which m^* is minimized as function of t (Condition 1). The value of m^* equals 4.05. The correlation between squared distances and expected data equals $r = -0.60$.

prevalent and most people read the newspaper very thorough (C3), that is, these categories are in the origin. Few people are in the lowest or highest educational groups, and also few people read the newspaper just by glance. From the common effect plot (right-hand side) we see that E1 is close to C1, that is, this group reads the paper at glance, while educational group E2 reads the paper fairly thorough (C2). Educational group E3 is in between and about equally spaced from C1 and C3. Educational group E4 is about equal far from fairly thorough and very thorough while E5 is very close to very thorough. Readership category C1 is about equally far from E1, E2, and E3. Readership category C3 is closest to educational group E5. The correlation between the data and the squared distances in the common display in this solution equals $r = -0.60$.

3.2.2 Scaling-τ

The value obtained for m^* is $\hat{m}^* = 4.06$ when $\hat{\tau} = 0.78$. The solution is shown in Figure 4. In the unique effect display we see that again E4 is the most prevalent educational group, but now readership categories C2 and C3 are in the origin. In the common effect display we see that educational group E2 is (compared to the previous solution) now more in between C1 and C2, although still closer to C2. Readership category C1 is closest to E1, and readership category C3 is closest to educational group E5. The correlation between the data and the squared distances in the common display in this solution equals $r = -0.56$

3.2.3 Differential Scaling

The value obtained for m^* is $\hat{m}^* = 4.02$ when $\hat{\mathbf{T}} = \text{diag}([1.30, 1.05])$, which corresponds to differential scalings $\tau_1 = 1.09$ and $\tau_2 = 0.44$. The solution is shown in Figure 5. In the unique effect display readership categories C2 and C3 are most prevalent, they lie both in the origin. Educational group E4 is again most prevalent and E1 and E5 least. In the common effect display E1 is close to C1, that is, this group reads the paper at glance,

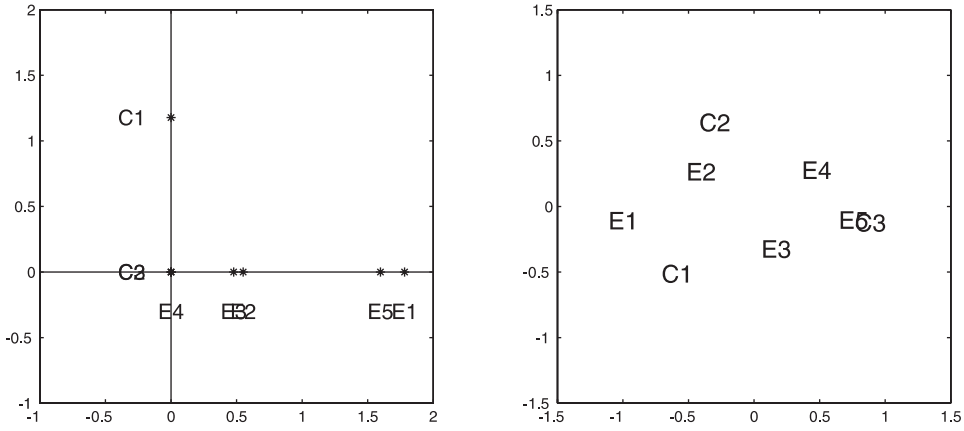

Figure 4. The solution for which m^* is minimized as function of τ (Condition 2). The value of m^* equals 4.06. The correlation between squared distances and expected data equals $r = -0.56$.

while educational group E2 reads the paper fairly thorough (C2). Educational group E3 is in between and about equally spaced from C1 and C3. Educational group E4 is about equal far from fairly thorough and very thorough while E5 is very close to very thorough. Readership category C1 is about equally far from E1, E2, and E3. Readership category C3 is closest to educational group E5. The correlation between the data and the squared distances in the common display in this solution equals $r = -0.60$

3.2.4 Transformation

The minimized value obtained for m^* is $\hat{m}^* = 3.93$ when

$$\hat{\mathbf{T}} = \begin{pmatrix} 1.39 & -0.41 \\ -0.71 & -0.80 \end{pmatrix}.$$

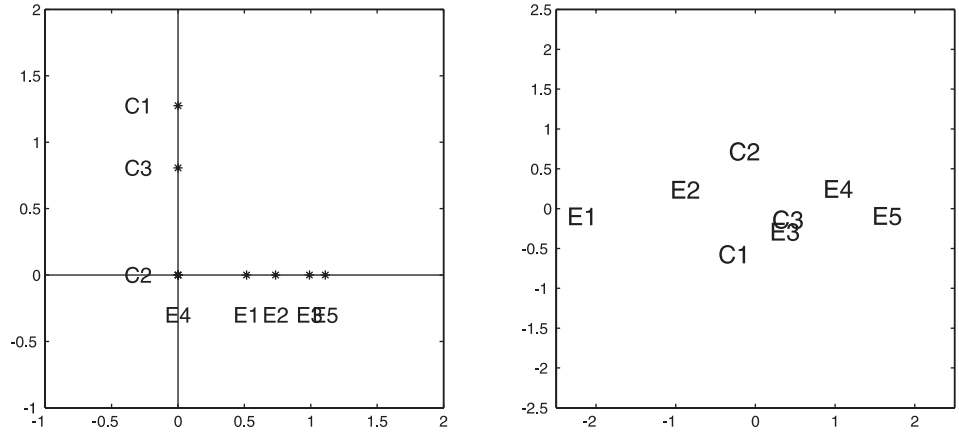

Figure 5. The solution for which m^* is minimized as function of diagonal $\mathbf{T}$ (Condition 3). The value of m^* equals 4.02. The correlation between squared distances and expected data equals $r = -0.60$.

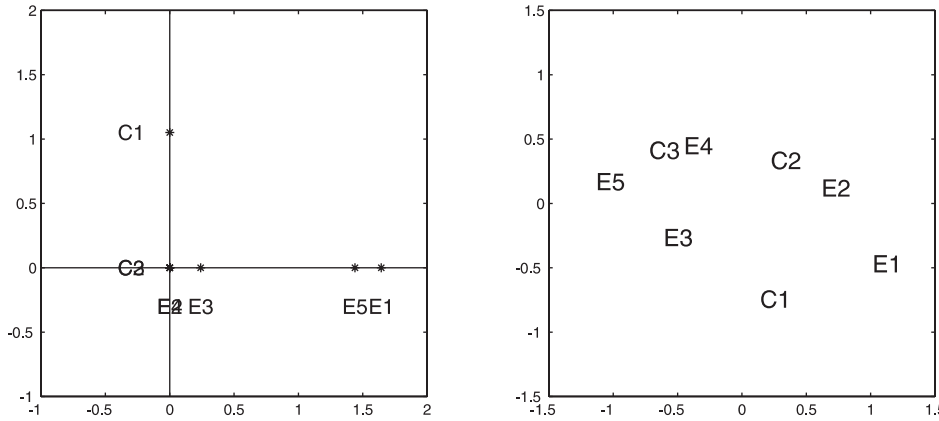

Figure 6. The solution for which m^* is minimized as function of $\mathbf{T}$ *(Condition 4). The value of m^* equals 3.93. The correlation between squared distances and expected data equals $r = -0.71$.*

The solution is shown in Figure 6 where in the unique effect display now Educational groups E4 and E2 are most prevalent together with readership categories C2 and C3. In the common effect display we see that educational group E4 is closest to C3, where in the previous solutions this was E5, also for educational group E3 readership category C3 is now closest. The correlation between the data and the squared distances in the common display in this solution equals $r = -0.71$

3.3 Minimizing the Correlation Between Distances and Expected Data

As can be seen in Figure 7 the correlation between the squared distances and the expected data (μ_{ij}), depends on τ. In Figure 7 it can be seen that the curve for the Pearson correlation and the Spearman rank correlation are about equal, suggesting that the relationship between the frequencies and squared distances is approximately linear. Another way of imposing the identification constraints is to minimize this correlation. From Figure 7 it is clear that for the current dataset this is at about $\hat{\tau} = 1.5$, where the correlation equals -0.68. Moreover, for the standard choices of scaling, that is, $0 \leq \tau_1 = \tau_2 \leq 1$, the correlation ranges from -0.3 ($\tau = 0$) to somewhat below -0.6 ($\tau = 1$).

The contour plot of the correlation as a function of differential scalings τ_r's is given in Figure 8. The approximate values for which the correlation is minimized is $\tau_1 = 2.5$ and $\tau_2 = 0.9$, the value of the correlation in that case equals -0.78.

The loss function for Conditions 1, 3, and 4 is

$$\arg\min_{\mathbf{T}\in\,\Omega} \sum_{i,j} \left(\mu_{ij} - \mu_{..}\right) \left(d_{ij}^2(\mathbf{XT}, \mathbf{YS}) - d_{..}^2(\mathbf{X}_\tau\mathbf{T}, \mathbf{Y}_\kappa\mathbf{S})\right)$$

$$= \arg\min_{\mathbf{T}\in\,\Omega} \sum_{i,j} e_{ij} \left(\mathbf{x}_i^T\mathbf{T}\mathbf{T}^T\mathbf{x}_i + \mathbf{y}_j^T\mathbf{S}\mathbf{S}^T\mathbf{y}_j\right), \quad (3.4)$$

where $e_{ij} = \mu_{ij} - \mu_{..}$, and a dot in the subscript denotes taking the mean over the omitted index. For Condition 2 the loss function is

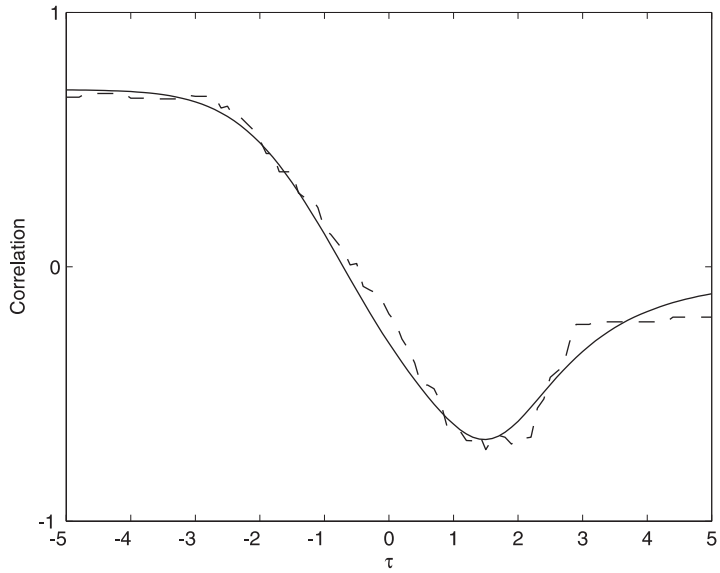

Figure 7. The correlation between the frequencies and the squared distances in the common effect dimensions as a function of τ. The solid line is the Pearson correlation coefficient, and the dashed line is the Spearman rank correlation coefficient.

$$\arg\min_{\tau\in\Re}\sum_{i,j}\left(\mu_{ij}-\mu_{\cdot\cdot}\right)\left(d_{ij}^2(\mathbf{X}_\tau;\mathbf{Y}_\kappa)-d_{\cdot\cdot}^2(\mathbf{X}_\tau;\mathbf{Y}_\kappa)\right)$$

$$=\arg\min_{\tau\in\Re}\sum_{i,j}e_{ij}\left(\sum_r\phi_r^{2\tau}(c_{ir}^2-c_{\cdot r}^2)+\phi_r^{2-2\tau}(d_{jr}^2-d_{\cdot r}^2)\right). \quad (3.5)$$

Both functions can be minimized using a quasi-Newton approach, where the Hessian is approximated by the BFGS formulas (see Gill, Murray, and Wright 1981, p. 119). This procedure is also implemented in MATLAB's Optimization Toolbox. This works good, and the experience of the author is that there is no trouble with local minima. Like for minimizing the constant, for Condition 4 the final value of $\mathbf{T}$ may differ because of the rotation by $\mathbf{V}$. This can be solved using the singular value decomposition of $\mathbf{T}$ and then redefine $\mathbf{T}=\mathbf{U}\mathbf{\Lambda}$.

3.3.1 Scaling-*t*

The minimized value obtained for r is $\hat{r}=-0.76$ when $\hat{t}=2.68$. The solution is shown in Figure 9, where in the unique effect display it can be seen that again educational group E4 is the most prevalent, but readership category C3 is not anymore. Now C2 is the most prevalent. In the common effect display we see the three readership categories clustered in the center with the educational groups surrounding this cluster. Educational group E1 is relatively far from the readership categories, but closest to category C1. Educational group E2 is closest to C2 while E3 is closest to C1 and C3. Comparing the readership categories we see that C3 is closest to E5 and E4. Comparing the Educational groups we see that E3 is closest to both C3 and C1. The value of the constant in this solution is $m^*=4.75$.

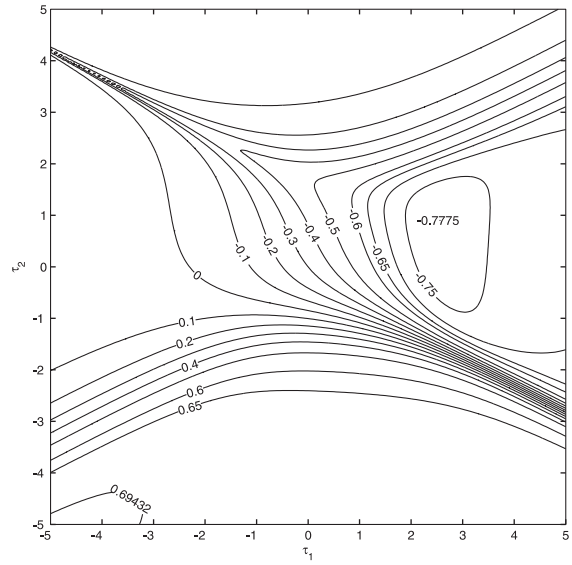

Figure 8. Contour plot of the correlation between the frequencies and the squared distances in the common effect dimensions as a function of τ_r.

3.3.2 Scaling-τ

The minimized value obtained for r is $\hat{r} = -0.68$ when $\hat{\tau} = 1.48$. The solution is shown in Figure 10, where in the unique effect display it can be seen that again educational group E4 is the most prevalent and readership category C2 is the most prevalent. C3 is relatively far from the origin. In the common effect display we see that educational groups E3, E4, and E5 are all close to C3, but that E4 is closest. Educational group E3 is closest to C1. Educational group E1 is far from all readership categories, but C1 is the readership

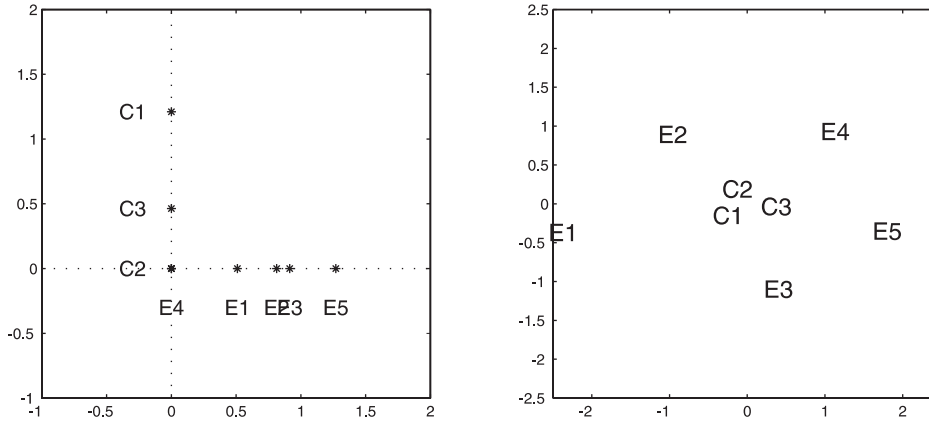

Figure 9. The solution for which the correlation is minimized as function of t (Condition 1). The value of m^* equals 4.75. The correlation between squared distances and expected data equals $r = -0.76$.

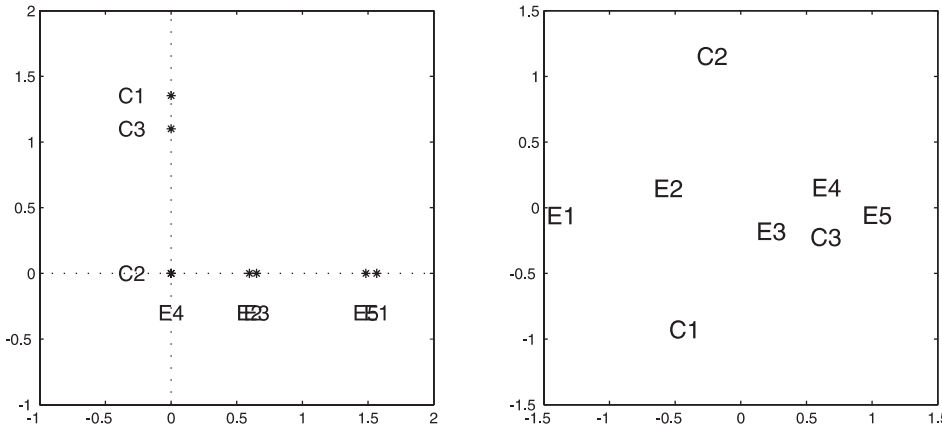

Figure 10. The solution for which the correlation is minimized as function of τ (Condition 2). The value of m^* equals 4.57. The correlation between squared distances and expected data equals $r = -0.68$.

category with the smallest distance from E1. Educational group E2 is in between C1 and C2, somewhat closer to C1. The value of the constant in this solution is $m^* = 4.57$.

3.3.3 Differential Scaling

The minimized value obtained for r is $\hat{r} = -0.78$ when $\hat{\mathbf{T}} = \text{diag}([2.46, 0.71])$, which corresponds to differential scalings $\tau_1 = 2.53$ and $\tau_2 = 0.91$ The solution is shown in Figure 11, where in the unique effects display, like in the previous two solutions, readership category C2 is most prevalent, than C3 and last C1. Also Educational group E4 is most prevalent but here E1 is second closest to the origin. In the common effect display the educational groups are all close to the horizontal axis, ordered from E1 to E5. The readership categories

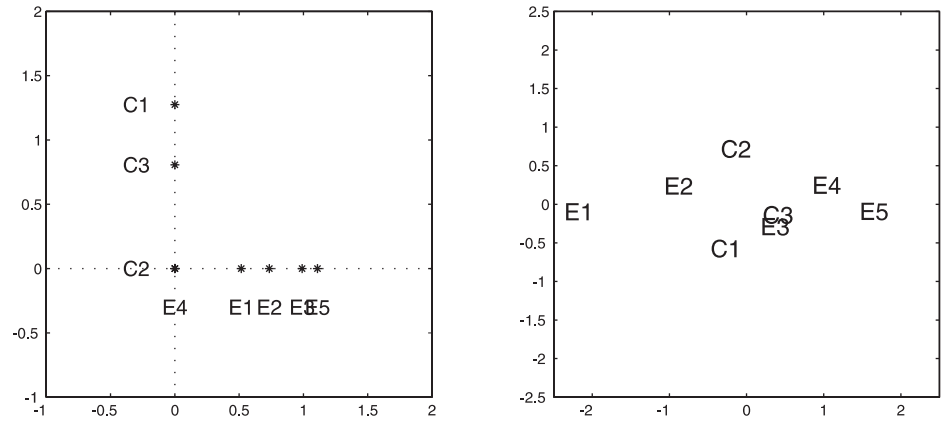

Figure 11. The solution for which the correlation is minimized as function of diagonal $\mathbf{T}$ (Condition 3). The value of m^* equals 4.49. The correlation between squared distances and expected data equals $r = -0.78$.

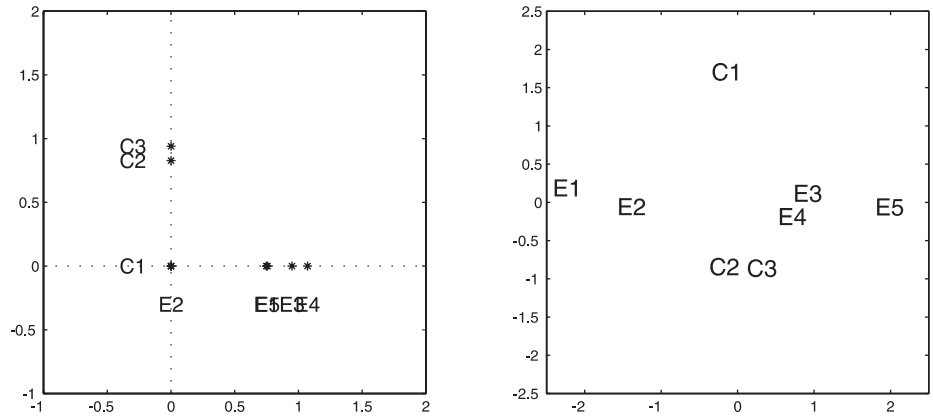

Figure 12. The solution for which the correlation is minimized as function of **T** *(Condition 4). The value of* m^* *equals 5.21. The correlation between squared distances and expected data equals* $r = -0.93$.

are close to the vertical axis. Educational group E3 is closest to readership category C3, but C3 is also the readership category that is closest to E4 and E5. Educational group E1 is far from all readership categories, but closest to C1. The value of the constant in this solution is $m^* = 4.49$.

3.3.4 Transformation

The minimized value of the correlation is $\hat{r} = -0.93$ when

$$\hat{\mathbf{T}} = \begin{pmatrix} 2.72 & -0.16 \\ -1.25 & -0.35 \end{pmatrix}$$

The solution is shown in Figure 12 where in the unique effect display it can be seen that now the most prevalent educational group is E2 and C1 is the most prevalent readership category. In the common effect display we see the Educational groups again situated close to a straight line. Educational group E4 is now closest to readership category C3, although this readership category is also the closest to E3 and E5. Educational group E1 is closer to C2 than to C1 which is also true for E2. The value of the constant in this solution is $m^* = 5.21$.

3.4 Considerations in Choosing an Identification

Eight (more or less) different solutions were shown in Figures 3, 4, 5, 6, 9, 10, 11, and 12. The question is now which solution to choose. Of course, at the end the user has to make a decision for his or her own data. Some further properties of the solutions will be discussed here that can guide us in choosing.

The interpretation of the common dimensions in the solutions for which the constant was minimized is close to the interpretation resulting from the inner-product perspective. That is, the correlations between squared distances in the common dimensions and values of

the inner-products are -0.89, -0.93, -0.89, and -0.86 in Conditions 1 to 4, respectively. *The solutions where the constant is minimized are thus best when one is interested in the deviation from independence.* When the correlation between expected data and distances in the common dimensions is minimized the distances provide a direct interpretation of the data, and the unique dimensions give real unicities of each of the categories. One of the major problems in the interpretation of, for example, correspondence analysis is that people tend to interpret the plot as the data and not as the departure of independence. *By minimizing the correlation between the distances and the (expected) data the distances represent the data itself*, and this problem is resolved.

The problem of which condition to choose is basically the problem of how far one wants to diverge from the standard approach. Normally a scaling is chosen such that τ equals 0, $\frac{1}{2}$, or 1. Here we discussed solutions where τ can be any real value, where τ may differ over the dimensions, and where first a rotation of the axes is sought and then a scaling is performed. The recommendation of the author is to use the transformations with nonsingular $\mathbf{T}$. In this case the chosen criterium is further minimized than in the other three conditions, and interpretation is clearest. If one insists upon preserving the principal axes a differential scaling is advised.

4. CONCLUSIONS AND DISCUSSION

The use of graphical displays is admissible only if the geometry is well understood by the users. Without this knowledge graphical displays are subject to incorrect interpretation and misuse. This article provides more insight into the graphical display of GBMs. The GBM may be interpreted by either inner-products or distances.

In general, the distance interpretation is more intuitive than the inner-product interpretation (see also Gower 2004, p. 713). However, the inner-product representation has some desirable features. Where the distance interpretation is from high to low without an indication where the sign is changing, the change of positive to negative is clear in inner-product representations. Moreover, the main effects in the inner-product representation are not affected by changes in τ and $\mathbf{T}$ whereas in the distance parameterization they are. In this article we used this dependence, to obtain identifications that (in some way) better represent the data. Moreover, it was shown that a standard scaling is not optimal for interpretational purposes.

One important lesson learned from all figures shown above, is that one should never look at the common plot only. The unique plot (or main effects) should always be taken into account. In all shown plots, the distances in the unique *and* the common dimensions give an exact representation of the data, while there are some clear differences between the solutions if one neglects either the unique or the common dimensions. Two ways of identifying the solution were proposed. The first minimizes the constant such that the distances are maximally differentiated; the second minimizes the correlation between expected data and squared distances in the common dimensions. Since graphical representations are often interpreted as if they are representing the data, the second way of choosing a scaling/transformation is recommended. In Section 3.4 further guidelines were given to make a

choice among the two criteria used. Other criteria are possible, for instance using external information. This article dealt with situations in which the identification is based on the expected data only.

A number of authors have shown that correspondence analysis represents distance relationships quite well (Heiser 1981, chap. 4; Ter Braak 1985; Nishisato 1996). However, in these articles scalings of the correspondence analysis solution are typically ignored, and a standard scaling is used. We showed that scalings have a major impact on the relationship between data and distances. This is of utmost importance when correspondence analysis is used as a technique to get results for Gaussian ordination (Ter Braak 1985). Having such an approximation one should be suspicious about the correlation between data and distances obtained, and probably the standard scaling is not sufficient.

To conclude the article: if one is willing to go along with the author in assuming that graphical displays are intuitively and thus in practice more often interpreted by distances than by inner-products, the current study is of major importance for the use of biplot models. Where Gabriel (2002) and Gower (2004) found that scalings hardly influence the interpretation of results in the inner-product perspective, we have shown that these scalings do have a major impact on the interpretation in the distance perspective. If one is not willing to go along with the author in the assumption, still generalized biadditive models and correspondence analysis are often used for Gaussian ordination and one should be careful not to take a standard scaling too easily.

ACKNOWLEDGMENTS

The author would like to thank John Gower and two anonymous reviewers for valuable comments on an earlier draft of this article.

[Received January 2005. Revised July 2006.]

REFERENCES

Becker, M. P. (1990), "Maximum Likelihood Estimation of the RC(M) Association Model," *Applied Statistics*, 39, 152–167.

De Falguerolles, A., and Francis, B. (1992), "Algorithmic Approaches for Fitting Bilinear Models," in *COMPSTAT 92, Computational Statistics* (vol. I), eds. Y. Dodge and J. Whittaker Heidelberg: Physica-Verlag for IASC, pp. 77–82.

——— (1994), "A Algorithmic Approach to Bilinear Models for Two-Way Contingency Tables," in *New Approaches in Classification and Data Analysis*, eds. E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and P. Burtschy, Berlin: Springer-Verlag, pp. 518–524.

De Rooij, M., and Heiser, W. J. (2003), "A Distance Representation of the Quasi-Symmetry Model and Related Distance Models," in *New Developments on Psychometrics: Proceedings of the International Meeting of the Psychometric Society,* eds. H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. J. Meulman, Tokyo: Springer-Verlag.

——— (2005), "Graphical Representations and Odds Ratios in a Distance-Association Model for the Analysis of Cross-Classified Data," *Psychometrika*, 70, 99–122.

Gabriel, K. R. (2002), "Goodness of Fit of Biplots and Correspondence Analysis," *Biometrika*, 89, 423–436.

Gauch Jr., H. G. (1992), *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*, Amsterdam: Elsevier

Gill, P.E., Murray, W., and Wright, M. H. (1981), *Practical Optimization*, London: Academic Press.

Goodman, L. A. (1991), "Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data," *Journal of the American Statistical Society*, 86, 1085–1111.

Gower, J. C. (2004), "The Geometry of Biplot Scaling," *Biometrika*, 91, 705–714.

Gower, J. C., and Hand, D. J. (1996), *Biplots*, London: Chapman and Hall.

Greenacre, M. J., and Hastie, T. (1987), "The Geometric Interpretation of Correspondence Analysis," *Journal of the American Statistical Association*, 82, 437–447.

Heiser, W. J. (1981), "Unfolding Analysis of Proximity Data," unpublished doctoral dissertation, Leiden University.

Ihm, P., and Van Groenewoud, H. (1984), "Correspondence Analysis and Gaussian Ordination," *COMPSTAT Lectures*, 3, 5–60.

Kempton, R. A. (1984), "The Use of Biplots in Interpreting Variety by Environment Interactions," *Journal of Agricultural Science, Cambridge*, 103, 123–135.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Nishisato, S. (1996), "Gleaning in the Field of Dual Scaling," *Psychometrika*, 61, 559–599.

Takane, Y. (1987), "Analysis of Contingency Tables by Ideal Point Discriminant Analysis," *Psychometrika*, 52, 493–513.

Ter Braak, C. J. F. (1985), "Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model," *Biometrics*, 41, 857–873.

——— (1987), "Unimodal Models to Relate Species to Environment," unpublished doctoral dissertation, Wageningen University.

Van Eeuwijk, F. A. (1995), "Multiplicative Interaction in Generalized Linear Models," *Biometrics*, 51, 1017–1032.