Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 587–603
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Homogeneity analysis using absolute deviations

George Michailidis[a,*], Jan De Leeuw[b]

[a] *Department of Statistics, The University of Michigan, 439 West Hall, 550 East University, Ann Arbor 481091092, USA*

[b] *Department of Statistics, University of California, Los Angeles, USA*

Received April 2003; received in revised form 12 March 2004; accepted 13 March 2004

Abstract

Homogeneity analysis is a technique for making graphical representations of categorical multivariate data sets. Such data sets can also be represented by the adjacency matrix of a bipartite graph. Homogeneity analysis optimizes a weighted least-squares criterion and the optimal graph layout is computed by an alternating least squares algorithm. Heiser Comput. Statist. Data Anal. (1987) 337, looked at homogeneity analysis under a more robust to outliers criterion, namely a weighted least absolute deviations criterion. In this paper, we take an in-depth look at the mathematical structure of this problem and show that the graph drawings are created by reciprocal computation of multivariate medians. Several algorithms for computing the solution are investigated and applications to actual data suggest that the resulting p-dimensional drawings ($p \geqslant 2$) are degenerate, in the sense that all object points are clustered in $p+1$ locations. We also examine some variations of the criterion used and conclude that the generate solutions observed are a consequence of the normalization constraint employed in this class of problems.

Keywords: Multivariate categorical data; Visualization; Optimal layout; Mazorization algorithms; Loss functions

1. Introduction

Homogeneity Analysis (also known as Multiple Correspondence Analysis (MCA)) is a well-known technique to make graphical representations of categorical multivariate data (Gifi, 1990). It can also be presented as a technique to produce informative layouts of *bipartite* graphs (Michailidis and de Leeuw, 1998; De Leeuw and Michailidis, 2000a,b).

* Corresponding author. Tel.: +1 7347633498; fax: +1 7347634676.
E-mail addresses: gmichail@umich.edu (G. Michailidis), deleeuw@stat.ucla.edu (J. De Leeuw).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.03.007

The setting is as follows: data have been collected for N *objects* on J *categorical* variables with k_j categories per variable. Let $K = \sum_{j=1}^{J} k_j$ be the total number of categories in the data set. Then, a graph $\mathcal{G}$ with nodes (vertices) corresponding to the N objects and the K categories and with edges linking the object nodes to the category nodes, and thus reflecting which objects belong to which categories, contains the same information as the original data set. The latter information is usually represented in matrix form through a binary (0-1) matrix $W = \{w_{ij}|,\ i = 1,\ldots,N,\ j = 1,\ldots,K\}$. It can be easily shown that the $(N + k) \times (N + k)$ matrix

$$A = \begin{bmatrix} 0 & W \\ W' & 0 \end{bmatrix}$$

corresponds to the *adjacency* matrix of our graph. The above defined *multivariate data graph* $\mathcal{G}$ with vertex set V and edge set E has a special structure, namely that the N nodes corresponding to the objects are not connected between themselves and similarly for the K category nodes. This can also be seen by the two zero submatrices in the adjacency matrix A of $\mathcal{G}$. Thus, we are dealing with a bipartite graph.

A *drawing* of the graph $\mathcal{G}$ is a mapping of its vertex set V into p-dimensional space. Adjacent points in the graph are connected by lines in the drawing. This goes in the direction of making a picture of the data, and when things work out well, a picture is worth a lot of numbers, especially when these numbers are just zeros and ones as several examples in the literature have shown (Gifi, 1990; Michailidis and de Leeuw, 1998).

The quality of the drawing is measured by the loss function

$$\mathbf{pull}_2(X, Y) = \sum_{i=1}^{N} \sum_{j=K}^{m} w_{ij} d^2(x_i, y_j), \tag{1.1}$$

where the x_i's contain the coordinates of the N objects and the y_j the coordinates of the K categories of all the variables in the p-dimensional space, and d denotes the Euclidean distance. The objective is to arrange the vertices (objects and categories) of the graph in such a way, so that the loss would be small. Thus points which are connected by lines should be close, i.e. the lines in the drawing should be short.

If we design algorithms to minimize $\mathbf{pull}_2(X, Y)$, then we must make sure that the perfect, but trivial, solution $X = Y = 0$ is excluded. This is done by imposing *normalization* constraints. For example, in MCA drawings are normalized by requiring that $X'X = I$. Under this normalization the solution to problem (1.1) is characterized by the *centroid* principle (Gifi, 1990), namely that the category points are located in the center of gravity of the objects they belong to. An additional advantage of this normalization is that the optimal solution is given by an eigenvalue problem (Gifi, 1990). The $p = 2$-dimensional solution for the Guttman–Bell and sleeping bags data sets (for their description see Section 4) that illustrate the centroid principle are given in Fig. 1.

However, MCA has a few drawbacks; the major ones are: (i) the influence of objects with 'rare' profiles that tend to dominate the solution (Michailidis and de Leeuw, 1998), as can be seen on the left part of the picture for the Guttman–Bell drawing and (ii) the presence of *horseshoes* (De Leeuw et al., 1980).

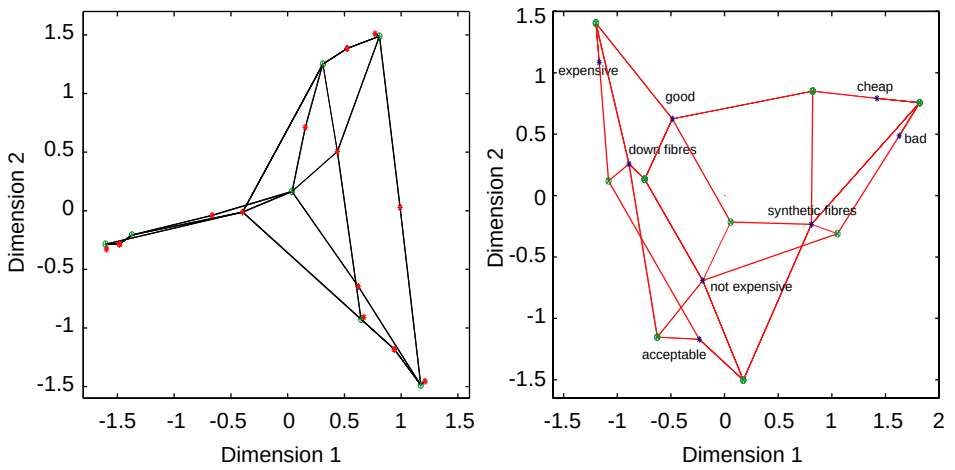

Fig. 1. Left panel: Guttman–Bell graph drawing, with $*$ denoting the object points and $\diamondsuit$ the category points. Right panel: Sleeping bags graph drawing. Both examples illustrate quite clearly the centroid principle.

One possible solution to these is to use a more 'robust' loss function, such as

$$\mathbf{pull}_1(X, Y) = \sum_{i=1}^{N}\sum_{j=1}^{K} w_{ij} d(x_i, y_j), \tag{1.2}$$

i.e. it is the same loss function as (1.1), but without squaring the distance. The same normalization is used as before, requiring that $X'X = I$. This is a special case of a very general framework introduced in Michailidis and de Leeuw (2001), where the square of the distance in the definition of the loss function (1.1) is replaced by a general function $\phi(d)$. Robust estimation has a very long history in statistics (Huber, 1981). The case (1.2) was discussed earlier in Heiser (1987) in the context of correspondence analysis (graphical representation of a two-way table) who gave an algorithm and an example that corresponded to our framework. The example showed clustering, in the sense that many of the objects and categories in the optimal drawing on the plane were collapsed into single points, and only very few distinct points were left. Heiser (1987, p. 349) made the following comments regarding this clustering phenomenon.

> How should we appreciate this result? There are perhaps two views. One is that in the process of mapping the original table into a spatial configuration too much of the fine detail is lost, and that the approach leads to a dead end. The other is that it appears to be possible to devise a class of clustering techniques that is smoothly related to a more continuous representation, and that seems to avoid the usual combinatorial complications.

In Fig. 2, the optimal graph drawings of the Guttman–Bell and sleeping bags data sets under loss function (1.2) are shown. In both cases a very strong clustering pattern

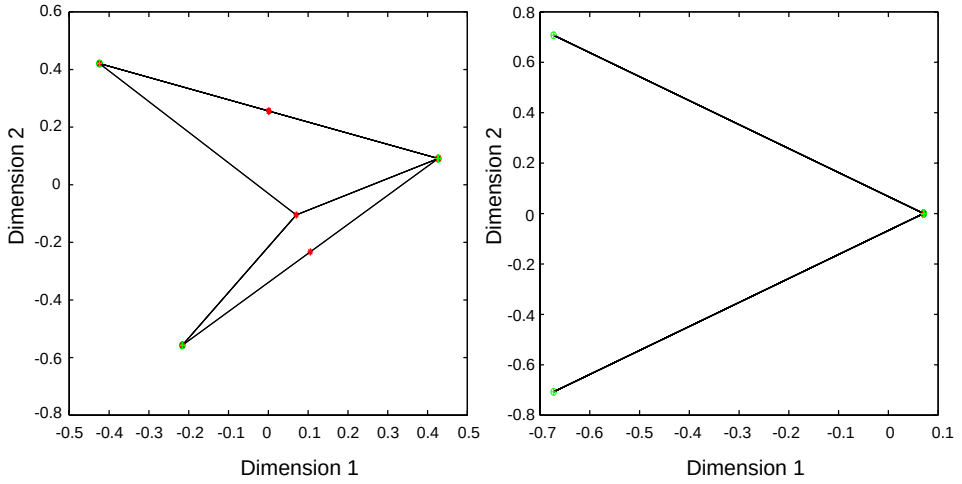

Fig. 2. Left panel: Guttman–Bell graph drawing under loss function (1.2); Right panel: sleeping bags graph drawing.

emerges for the object points; i.e. all of the object nodes occupy only three locations. On the other hand, the category points still seem to obey some form of the centroid principle for the Guttman–Bell example.

Experience with many other categorical data sets with varying numbers of objects, variables and categories per variable confirm the above empirical finding; namely, that the optimal 2-dimensional layout consists of three object nodes Michailidis and de Leeuw (2000). Analogously, the three-dimensional layouts consist of four object nodes. Finally, for $p=1$ the result also holds, namely that the optimal solution consists of two points only, and is *rigorously* proven in De Leeuw and Michailidis (2003). Obviously, such solutions become totally *uninteresting* from a data analysis point of view, since they are unable to uncover interesting patterns in the data. Hence, it is of great interest to gain insight into the origins of this phenomenon and examine possible alternatives that overcome the problem.

The paper is organized as follows: Section 2 discusses the structure of the loss function (1.2) and presents several optimization algorithms for computing the optimal solution. In Section 3, the structure of the optimal solution is investigated and in Section 4 the performance of the various algorithms is examined. Finally, in Section 5 we look into other loss functions and present some potential solutions to the strong clustering problem observed.

2. The loss function and its optimization

Our objective is to minimize loss function (1.2) over all $N \times p$ matrices X satisfying $X'X=I$ and over all $K \times p$ matrices Y. The $N \times K$ matrix $W=\{w_{ij}\}$ is the off-diagonal part of the adjacency matrix A of the bipartite multivariate data graph $\mathscr{G}$.

For purposes of regularization, to avoid problems with differentiability and division by zero, we actually define

$$d_\varepsilon^2(x_i, y_j) \triangleq (x_i - y_j)'(x_i - y_j) + \varepsilon^2$$

throughout, where ε is small, and we minimize

$$\mathbf{pull}_{(1,\varepsilon)}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} d_\varepsilon(x_i, y_j).$$

In the remainder of the paper we will omit the subscripts in **pull**, because we will be dealing exclusively with $\mathbf{pull}_{(1,\varepsilon)}$.

2.1. A matrix expression for the loss function

If we use unit matrices $E_{ij} = e_i e_j'$, where $e_i(e_j)$ are column vectors with a one in the ith (jth) position and zeros everywhere else, we can write

$$d^2(x_i, y_j) = (x_i - y_j)'(x_i - y_j) = \mathbf{tr}\, X'E_{ii}X + \mathbf{tr}\, Y'E_{jj}Y - 2\,\mathbf{tr}\, X'E_{ij}Y, \tag{2.1}$$

and thus

$$\mathbf{pull}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)}\{\mathbf{tr}\, X'E_{ii}X + \mathbf{tr}\, Y'E_{jj}Y - 2\,\mathbf{tr}\, X'E_{ij}Y + \varepsilon^2\}. \tag{2.2}$$

In matrix notation we can write

$$\mathbf{pull}(X, Y) = \mathbf{tr}\, X'A_\varepsilon(X, Y)X + \mathbf{tr}\, Y'B_\varepsilon(X, Y)Y - 2\,\mathbf{tr}\, X'C_\varepsilon(X, Y)Y + \varepsilon^2\tau_\varepsilon(X, Y), \tag{2.3a}$$

with

$$A_\varepsilon(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)} E_{ii} = \sum_{i=1}^{n} E_{ii} \sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)}, \tag{2.3b}$$

$$B_\varepsilon(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)} E_{jj} = \sum_{j=1}^{m} E_{jj} \sum_{i=1}^{n} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)}, \tag{2.3c}$$

$$C_\varepsilon(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)} E_{ij}. \tag{2.3d}$$

and

$$\tau_\varepsilon(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i, y_j)}. \tag{2.3e}$$

Observe that $A_\varepsilon(X, Y)$ and $B_\varepsilon(X, Y)$ are both diagonal and contain the row and column sums of C_ε respectively.

2.2. Influence of the smoothing parameter

We briefly examine the influence of the smoothing parameter ε, next. Let

$$\mathbf{pull}(\varepsilon) \triangleq \min_{X'X=I} \min_{Y} \mathbf{pull}_{(1,\varepsilon)}(X, Y).$$

and denote by $X(\varepsilon)$ and $Y(\varepsilon)$ its minimizers.

Proposition 2.1.

(1) *The objective function* $\mathbf{pull}(\varepsilon)$ *is increasing in the parameter* ε.
(2) $\lim_{\varepsilon\to 0} \mathbf{pull}(\varepsilon) = \mathbf{pull}(0)$.

Proof. The first part follows by differentiating the objective function with respect to ε

$$\frac{\partial \mathbf{pull}(\varepsilon)}{\partial \varepsilon} = \varepsilon \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{w_{ij}}{d_\varepsilon(x_i(\varepsilon), y_j(\varepsilon))} \geqslant 0,$$

which implies that it is increasing with larger values of ε.

For the second part it suffices to examine a single term. It is easy then to see that for the (i,j)th term we have that $|\mathbf{pull}(\varepsilon) - \mathbf{pull}(0)| = \sqrt{\varepsilon}$ and the result follows. □

Experience has shown that for values of $\varepsilon < 10^{-5}$ its effect on the loss function is truly marginal.

2.3. Optimization algorithms

The minimization problem of the **pull** function has the special property that there are two blocks of variables X and Y, which are treated in an asymmetric way. We normalize X by $X'X = I$ and we leave Y free. This makes it natural to use optimization methods, which take this block structure into account De Leeuw and Michailidis (2000a,b).

We briefly present one approach that sheds light into the structure of the problem under consideration and then introduce another algorithm which proves very attractive from a *programming* point of view. Finally, we present a third algorithm that avoids the computationally expensive eigenvalue decompositions present in the second algorithm.

The first approach is based on *block relaxation*, which alternates minimization over the variables in block X, while keeping Y fixed, and minimization over Y, with block X fixed. We alternate minimization of $\mathbf{pull}(X, Y)$ over Y with X fixed at its current value and over X satisfying $X'X = I$ with Y fixed. More precisely, we start with $X^{(0)}$. Then we alternate, for $k = 0, 1, \ldots$

$$Y^{(k)} = \operatorname{argmin}_Y(\mathbf{pull}(X^{(k)}, Y)),$$

$$X^{(k+1)} = \operatorname{argmin}_{X'X=I}(\mathbf{pull}(X, Y^{(k)})).$$

The first subproblem, updating Y, due to the Euclidean distance function used, amounts to solving K separate Weber problems (Vardi and Zhang, 2001). To find the coordinates

in $\mathbb{R}^p$ of category point y_j we minimize

$$\mathbf{pull}(y_j) = \sum_{i=1}^{n} w_{ij} d_\varepsilon(x_i, y_j).$$

The solution to this problem corresponds to determining in p-dimensional space the coordinates of a *multivariate median*. An enormous body of literature has emerged over the years for solving the Weber problem, also known in the optimization literature as the problem of minimizing a sum of Euclidean norms (Kuhn, 1967). The classical algorithm is the one by Weiszfeld (1937), which is a linearly convergent majorization method (Vardi and Zhang, 2001; Voss and Eckhardt, 1980). The second subproblem, updating X for fixed Y, is considerably more complicated because of the normalization constraint $X'X = I$, which defines a Stiefel manifold. The general methodology of optimizing functions over the Stiefel manifold proposed by Edelman et al. (1999) could then be used.

A second approach can be based on the concept of majorization (Lange et al., 2000; De Leeuw and Michailidis, 2000a,b). By the Arithmetic Mean/Geometric Mean inequality we have for points x_i, y_i, $\tilde{x}_i$, $\tilde{y}_i \in \mathbb{R}^p$ that

$$\sqrt{d^2(x_i, y_j) d^2(\tilde{x}_i, \tilde{y}_j)} \leqslant \tfrac{1}{2}\{d^2(x_i, y_j) + d^2(\tilde{x}_i, \tilde{y}_j)\},$$

and thus

$$d(x_i, y_j) \leqslant \frac{1}{2d(\tilde{x}_i, \tilde{y}_j)} \{d^2(x_i, y_j) + d^2(\tilde{x}_i, \tilde{y}_j)\}.$$

This implies

$$\mathbf{pull}(X, Y) \leqslant \tfrac{1}{2}\{\mathbf{pull}(\tilde{X}, \tilde{Y}) + \mathbf{pull}(X, Y|\tilde{X}, \tilde{Y})\},$$

where

$$\mathbf{pull}(X, Y|\tilde{X}, \tilde{Y}) \triangleq \mathbf{tr}\, X' A_\varepsilon(\tilde{X}, \tilde{Y}) X + \mathbf{tr}\, Y' B_\varepsilon(\tilde{X}, \tilde{Y}) Y - 2\,\mathbf{tr}\, X' C_\varepsilon(\tilde{X}, \tilde{Y}) Y,$$

with $\tilde{X}, \tilde{Y}$ the optimal values from a previous iteration of the algorithm. The last expression further implies that we can construct a convergent algorithm by using the current best solution for $(\tilde{X}, \tilde{Y})$ and finding the next best solution by minimizing $\mathbf{pull}(X, Y|\tilde{X}, \tilde{Y})$. The solution $(\hat{X}, \hat{Y})$ for the latter problem is given by

$$\hat{Y} = B_\varepsilon^{-1}(\tilde{X}, \tilde{Y}) C_\varepsilon'(\tilde{X}, \tilde{Y}) \hat{X},$$

where $\hat{X}$ solves the eigenvalue problem

$$D_\varepsilon(\tilde{X}, \tilde{Y}) \hat{X} = \hat{X} \Lambda, \tag{2.4}$$

with

$$D_\varepsilon(\tilde{X}, \tilde{Y}) \triangleq A_\varepsilon(\tilde{X}, \tilde{Y}) - C_\varepsilon(\tilde{X}, \tilde{Y}) B_\varepsilon^{-1}(\tilde{X}, \tilde{Y}) C_\varepsilon'(\tilde{X}, \tilde{Y})$$

and with Λ a diagonal matrix containing the p smallest eigenvalues of the matrix $D_\varepsilon(\tilde{X}, \tilde{Y})$. It is worth noting that the smallest eigenvalue is 0, since both the rows and the columns of $D_\varepsilon(\tilde{X}, \tilde{Y})$ add up to zero as a weighted sum of matrices of the form $(e_i - e_j)(e_i - e_j)'$.

It follows that at the optimal solution (in fact, at any stationary point of the algorithm) $\mathbf{pull}(X,Y)$ is equal to the sum of the p smallest eigenvalues of $D_\varepsilon(X,Y)$, while X is the corresponding set of eigenvectors. The matrix Y contains the weighted centroid $B_\varepsilon^{-1}(X,Y)C_\varepsilon'(X,Y)X$, which means that at the same time the y_js solve the corresponding Weber problems, previously discussed.

Observe that this also implies that we cannot use the normalization $\mathbf{tr}(X'X)=p$. By the argument above, all columns of X would be equal to the eigenvector corresponding to the smallest eigenvalue of $D_\varepsilon(X,Y)$, which gives an interesting solution only if the smallest eigenvalue has multiplicity of at least p.

In order to avoid solving a sequence of eigenvalue problems we can resort to a second level of majorization. This can be done by a second majorization, this time of $\mathbf{pull}(X,Y|\tilde{X},\tilde{Y})$. Write $X=\tilde{X}+(X-\tilde{X})$. Then

$$\mathbf{pull}(X,Y|\tilde{X},\tilde{Y})=\mathbf{pull}(\tilde{X},Y|\tilde{X},\tilde{Y})+2\,\mathbf{tr}\,(X-\tilde{X})'\{A_\varepsilon(\tilde{X},\tilde{Y})Y-C_\varepsilon(\tilde{X},\tilde{Y})\tilde{X}\}$$
$$+\mathbf{tr}\,(X-\tilde{X})'A_\varepsilon(\tilde{X},\tilde{Y})(X-\tilde{X}).$$

Suppose $\alpha(\tilde{X},\tilde{Y})$ is the largest diagonal element of $A_\varepsilon(\tilde{X},\tilde{Y})$. Also, let

$$\bar{X}=\tilde{X}-\frac{1}{\alpha(\tilde{X},\tilde{Y})}V_\varepsilon(\tilde{X},\tilde{Y}),$$

where $V_\varepsilon(\tilde{X},\tilde{Y})\triangleq A_\varepsilon(\tilde{X},\tilde{Y})Y-C_\varepsilon(\tilde{X},\tilde{Y})\tilde{X}$. Then, the second term above can be written (using the definition of $\bar{X}$) as

$$2\,\mathbf{tr}(X-\bar{X})'V_\varepsilon(\tilde{X},\tilde{Y})-\frac{2}{\alpha(\tilde{X},\tilde{Y})}\,\mathbf{tr}\,V_\varepsilon'(\tilde{X},\tilde{Y})V_\varepsilon(\tilde{X},\tilde{Y}),$$

and the third term as where $V_\varepsilon(\tilde{X},\tilde{Y})\triangleq A_\varepsilon(\tilde{X},\tilde{Y})Y-C_\varepsilon(\tilde{X},\tilde{Y})\tilde{X}$. Then, the second term above can be written (using the definition of $\bar{X}$) as

$$2\,\mathbf{tr}(X-\bar{X})'V_\varepsilon(\tilde{X},\tilde{Y})-\frac{2}{\alpha(\tilde{X},\tilde{Y})}\,\mathbf{tr}\,V_\varepsilon'(\tilde{X},\tilde{Y})V_\varepsilon(\tilde{X},\tilde{Y}),$$

and the third term as

$$\frac{1}{\alpha(\tilde{X},\tilde{Y})^2}V_\varepsilon'(\tilde{X},\tilde{Y})A(\tilde{X},\tilde{Y})V_\varepsilon(\tilde{X},\tilde{Y})-\frac{2}{\alpha(\tilde{X},\tilde{Y})}(X-\bar{X})'A_\varepsilon(\tilde{X},\tilde{Y})V_\varepsilon(\tilde{X},\tilde{Y}).$$

Collecting terms, using the definition of $\alpha(\tilde{X},\tilde{Y})$ and some algebra show that

$$\mathbf{pull}(X,Y|\tilde{X},\tilde{Y})\leqslant\mathbf{pull}(\tilde{X},Y|\tilde{X},\tilde{Y})+\alpha(\tilde{X},\tilde{Y})\mathbf{tr}\,(X-\bar{X})'(X-\bar{X})$$
$$-\frac{1}{\alpha(\tilde{X},\tilde{Y})}\,\mathbf{tr}\{V_\varepsilon'(\tilde{X},\tilde{Y})V_\varepsilon(\tilde{X},\tilde{Y})\}.$$

Minimizing this second majorization over X is the same as minimizing $\mathbf{tr}\ (X-\bar{X})'(X-\bar{X})$, which is a so-called *orthogonal procrustes* problem, whose solution is classical. If $\bar{X}=K\Gamma L'$ is the singular value decomposition of $\bar{X}$, then the solution is $\hat{X}=KL'$. This is the algorithm proposed by Heiser (1987), compare also Kiers (1997).

3. Performance assessment of the algorithms through real examples

3.1. Guttman–Bell dataset

This small dataset dealing with attitudes of social groups (also analyzed in Guttman (1968) and in Gifi (1990)) consists of 7 objects and 5 variables with a total of $K = 17$ categories. In Fig. 3, the homogeneity analysis solution under the $\mathbf{pull}_2$ and the $\mathbf{pull}_1$ loss functions are given. The lines indicate which objects under the $\mathbf{pull}_2$ solution are mapped to three points that describe the $\mathbf{pull}_1$ solution. This mapping of the object points to one of the 3 locations is completely determined by the solution of the K Weber problems, as shown in the next Section. In Table 1 the correspondence between the 17 category points and the 3 object points in the solution is given. It can be seen that all the objects belonging to category A1 are mapped to the same location. On the other hand the two objects belonging to category B2 are mapped to two different locations, while one of the objects in category E2 is mapped to the first point and the remaining 3 objects to the second point. The boxed entries indicate where, according to Witzgall's majority theorem (see Section 4), the category point should be located. Notice that all the contributions to the loss function come from categories whose objects are not mapped to a single location.

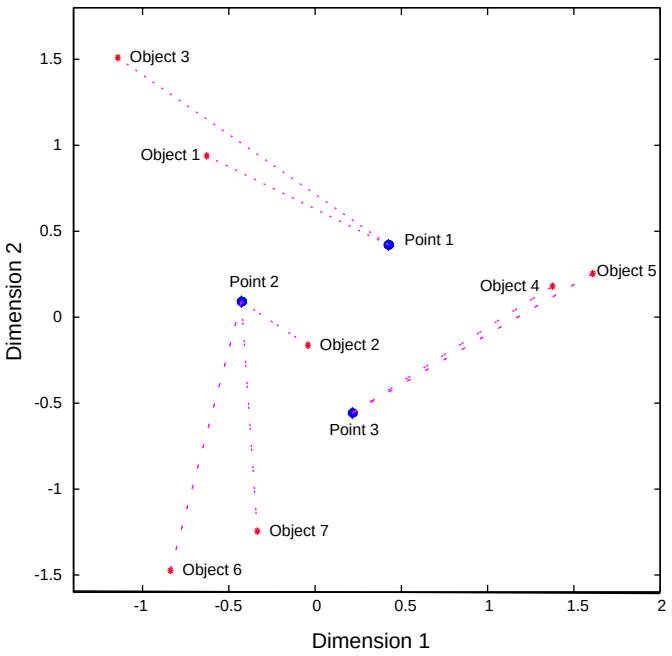

Fig. 3. The homogeneity analysis solution for the Guttman–Bell data set under (1.1) (red * points) and the solution under (1.2) (blue diamond-shape points) and the correspondence between the two solutions.

Table 1
Decomposition of the total loss for the Guttman–Bell data set

Categories	Point 1	Point 2	Point 3	Loss
A1	[2]	0	0	0
A2	0	[2]	0	0
A3	0	[1]	0	0
A4	0	0	[2]	0
B1	[2]	0	0	0
B2	0	[1]	[1]	0.91
B3	0	[2]	0	0
B4	0	0	[1]	0
C1	[1]	0	0	0
C2	[1]	[1]	0	0.91
C3	0	[2]	0	0
C4	0	0	[2]	0
D1	[1]	[1]	0	0.91
D2	1	[2]	[2]	2.61
E1	[1]	0	0	0
E2	1	[3]	0	0.91
E3	0	0	[2]	0
Total	10	15	10	6.26

For example, the two objects that belong to category A1 are located in the optimal $\mathbf{pull}_1$ solution at point 1 and hence no loss is incurred. On the other hand, the two objects that belong to category B2 are located at points 2 and 3 respectively, and therefore a loss is incurred.

3.2. Sleeping bags

This data set is taken from De Leeuw and Michailidis (2000a,b) and describes 21 sleeping bags in terms of three variables (price, filling and quality) with a total of 8 categories. Thus, its structure is different that the Guttman–Bell data set, since there are more objects than categories. In Fig. 4 the homogeneity analysis solution under (1.1) together with the one under absolute deviations are given. The multiple lines that originate from the points of the first solution is due to the fact that several objects, exhibiting identical patterns, have been mapped to the same location (a well known property of that solution; see Michailidis and de Leeuw (1998)). In Table 2 the decomposition of the total loss for the optimal solution is given, along with the correspondence between the 8 category points and the 3 object points. It can be seen again that losses occur when all the objects belonging to a particular category are not mapped to the same location.

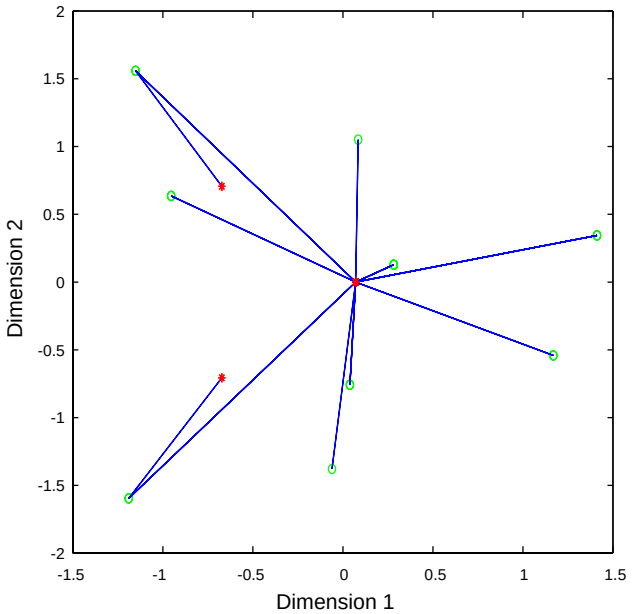

Fig. 4. The homogeneity analysis solution for the sleeping bags data set under (1.1) (red * points) and the corresponding absolute deviations solution (blue diamond-shape points) and the correspondence between the two solutions.

Table 2
Decomposition of the total loss for the sleeping bags data set

Categories	Point 1	Point 2	Point 3	Loss
A1	4	1	0	1.03
A2	11	0	1	1.03
A3	4	0	0	0
B1	7	0	1	1.03
B2	12	1	0	1.03
C1	10	0	0	0
C2	6	0	1	1.03
C3	3	1	0	1.03
Total	57	3	3	6.16

3.3. Performance assessment of the optimization algorithms

In this Section we briefly examine the performance of the three algorithms presented in Section 2; namely the block relaxation algorithm (A1), the majorization algorithm

Table 3

Data Set	A1	A2	A3
Guttman–Bell	72	88	84
Sleeping bags	86	96	94

(A2) and the double majorization algorithm (A3). The final configuration of the object and category points and the value of the loss function were calculated for 99 random starts of the object points (kept fixed for the three algorithms) and for the solution provided by the $\mathbf{pull}_2$ solution. In Table 3 the number of times the various algorithms found the minimal configuration is shown. It is worth noting that for the sleeping bags data set all three algorithms did not converge to the minimal solution when the starting point was the homogeneity analysis solution, a finding that has been observed with other data sets as well. The results indicate that the majorization algorithm outperforms its competitors. However, its down side is that for large problems a somewhat expensive eigenvalue problem needs to be solved a fairly large number of times.

4. The structure of the optimal solution

The block relaxation algorithms has provided insight into the structure of the optimal solution with respect to the category points. Since the y_js must correspond to multivariate medians, their position in the optimal graph layout is completely determined by this requirement.

Moreover, on the basis of extensive numerical experience (see previous section and also Michailidis and de Leeuw (2000)) we make the following conjecture.

Conjecture 4.1. *The p-dimensional optimal solution X that minimizes the* $\mathbf{pull}(X, Y)$ *function subject to the normalization constraint $X'X = I$, has exactly $p + 1$ distinct points.*

Knowledge of the location of the $p + 1$ points in X makes it simple to determine the location of the points in Y due to the following result.

Proposition 4.1 (Destination optimality). *Suppose $\hat{y}$ minimizes* $\mathbf{pull}(y) = \sum_{i=1}^{m} w_i d(x_i, y)$, *where the x_i are distinct. Then $\hat{y} = x_k$ if and only if*

$$w_k \geqslant \left\| \sum_{i \neq k} w_i \frac{x_i - x_k}{d(x_i, x_k)} \right\|$$

Proof. See Kuhn (1967), Theorem 4.2. □

A useful corollary that explains the decomposition of the total loss presented in the tables of Section 3 is

Corollary 4.2 (Witzgall's Majority Theorem (Witzgall, 1964)). *If* $w_k \geqslant \sum_{i \neq k} w_i$ *then* $\hat{y} = x_k$.

Proof.

$$\left\| \sum_{i \neq k} w_i \frac{x_i - x_k}{d(x_i, x_k)} \right\| \leqslant \sum_{i \neq k} w_i \left\| \frac{x_i - x_k}{d(x_i, x_k)} \right\| = \sum_{i \neq k} w_i. \qquad \square$$

Assuming that the conjecture is true and given the above results, there is an alternative algorithm worth mentioning. Suppose S is an *assignment* matrix, i.e. an $n \times (p+1)$ binary indicator matrix, which assigns each object i to one of the $p+1$ points. The column sums of S are the occupancies of the points, and the occupancies together with the normalization constraint $X'X = I$ determine the location of the points up to a rotation. Then we can fit in the Y points by solving the corresponding Weber problem. It follows that the solution is completely determined by the assignment S, and thus we can consider our loss function **pull** to be a function of assignments only. Optimizing over assignments obviously is a combinatorial optimization problem. For data sets with a large number of categories per variable we can establish the following result.

Corollary 4.3. *If the* $p+1$ *points conjecture holds and* $k_j \geqslant 3$, $j = 1, \ldots, J$, *that is the frequencies for all categories of all the variables are larger than* 3, *then the minimum loss is given by*

$$\min_X \mathbf{pull}_1(X, Y) = L \sum_{i=2}^{p+1} d(x_1, x_i), \tag{4.1}$$

where L *is the number of* $w_{ij} \neq 0$ *corresponding to the points located at* $x_j \neq x_1$.

Proof. Given the conjecture, without loss of generality the last $N-(p+1)$ points can be collapsed to point x_1. The Witzgall's Majority Theorem together with the assumption regarding the category frequencies show that $y_j = x_1$ for all j. Hence, $d(x_i, y_j) \equiv d(x_1, x_i)$ for $i = 2, \ldots, p+1$ and given that points $x_2, \ldots, x_{p+1}$ have L nonzero w_{ij}'s, the result follows. $\square$

5. Discussion: other loss functions and potential solutions

The $\mathbf{pull}_1(X, Y)$ function used so far is a special case of the more general class of functions defined by

$$\mathbf{pull}_\beta(X, Y) = \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} d(x_i, y_j)^\beta, \beta \in [1, 2]. \tag{5.1}$$

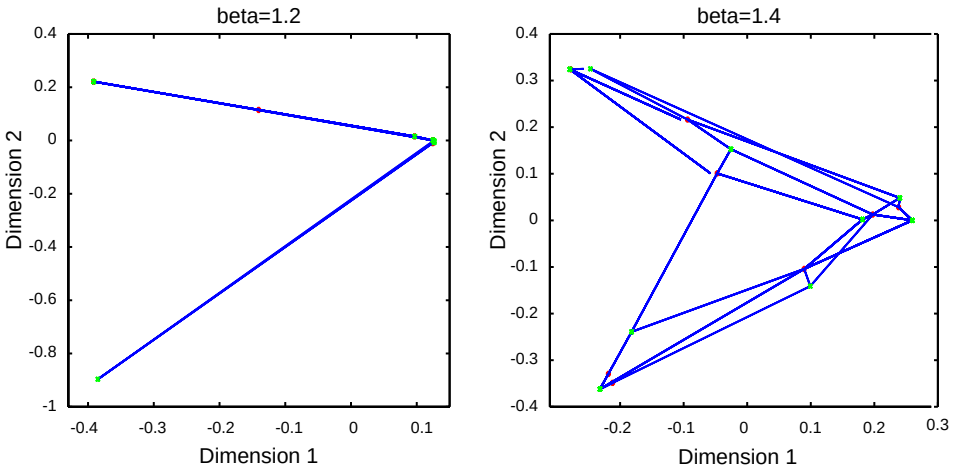

Fig. 5. Left panel: $\beta = 1.2$; Right panel: $\beta = 1.4$.

This is a family of convex functions with growth rates slower than the quadratic. The class contains as extreme cases both the $\mathbf{pull}_2$ and the $\mathbf{pull}_1$ functions. An application of Young's inequality shows that we can construct a majorization algorithm to minimize members of this class under the $X'X = I$ minimization constraint. Specifically we have that

$$d(x_i, y_j)^\beta \leqslant \frac{2-\beta}{2}\, d(\tilde{x}_i, \tilde{y}_j)^\beta + \frac{2}{\beta d(\tilde{x}_i, \tilde{y}_j)^{2-\beta}}\, d(x_i, y_j)^2, \tag{5.2}$$

which implies that we can construct a *quadratic* majorizing function and thus in one iteration we solve an eigenvalue problem similar to the one given in (2.4). The resulting graph layouts for the sleeping bags data for values of $\beta = 1.2, 1.4, 1.6$ and 1.8 are shown in Figs. 5 and 6. It can be seen that for values of β around 1.4 there seems to occur a 'phase transition', since for larger values the result is essentially identical to the one obtained in homogeneity analysis, while for smaller values identical to those from the $\mathbf{pull}_1$ loss function. For data sets involving a larger number of objects and categories experience indicates that the 'critical' value for the parameter β is around 1.5.

We have also examined a variety of other loss functions that employ the logarithm of the distances, or the logarithm of the squared distances, or the logistic function of the distances, or Huber's and biweight functions (Verboon, 1994) with analogous results. It should be noted that a similar algorithm as above, based on the concept of majorization works for these other loss functions. The results emphasize the very special nature of the $\mathbf{pull}_2$ function, which in conjunction with the $X'X = I$ normalization, is the only one that produces interesting from a data analysis point of view results; i.e. a layout of the object and category points in p-dimensional space that allows the data analyst to obtain insight about patterns in the multivariate categorical data set under consideration.

For loss functions that attempt to robustify the distances involved, the normalization constraint becomes highly problematic. The message of our investigations is that

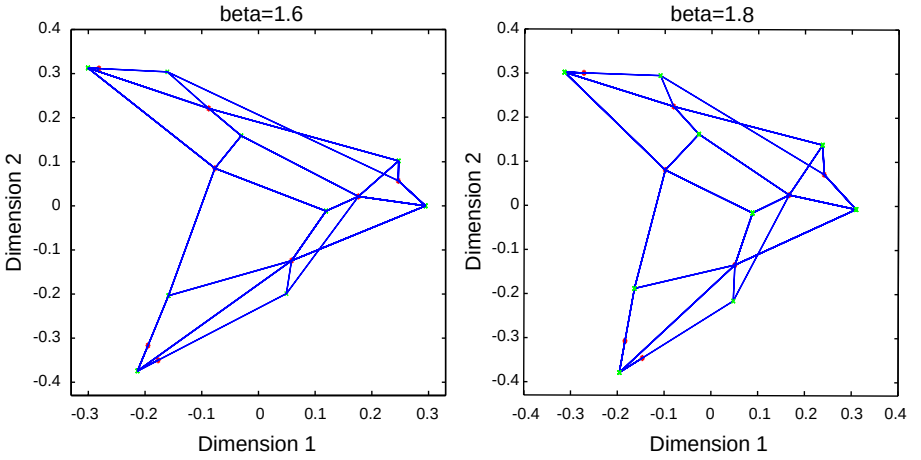

Fig. 6. Left panel: $\beta = 1.6$; Right panel: $\beta = 1.8$.

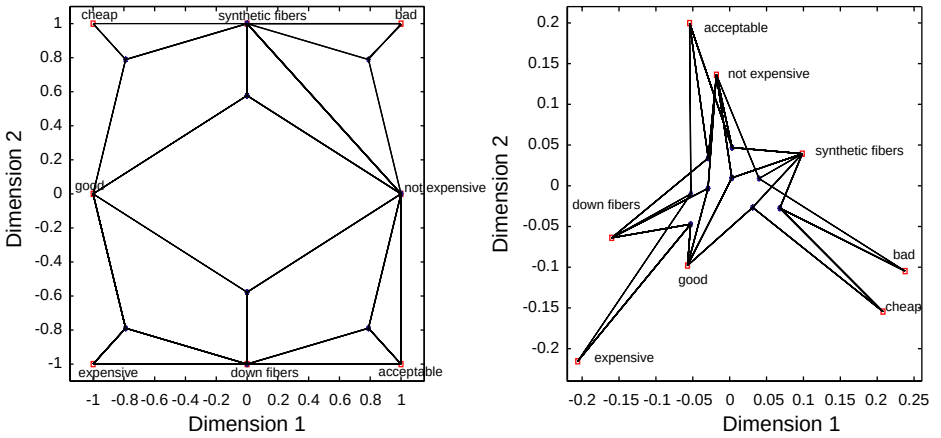

Fig. 7. Left panel: graph layout of the sleeping bag data set under Tutte normalization on the categories; Right panel: graph layout of sleeping bag data set under the (1.1) loss function and with the category points normalized.

different types of normalization constraints must be found that are more suitable to these other loss functions.

An interesting alternative is provided by the Tutte normalization (Tutte, 1963) that requires fixing before hand the locations of a number of points (e.g. the category points of one or even all the variables) and then in the case of a $\mathbf{pull}_1$-like loss function find the locations of the remaining points by calculating their multivariate medians (by solving the corresponding Weber problems). This goes towards the direction of facility location problems (Drezner, 1995), which may provide interesting alternatives

in visualizing categorical multivariate data. The resulting graph layout of the sleeping bags data set with the category points located on a square and the object points corresponding to the multivariate medians (Weber points) of the categories they belong to is shown in Fig. 7 (left panel). It should be noted that the arrangement of the category points Y on the square is such that it gives the minimum $\mathbf{pull}_1$ loss over all (8!) possible arrangements. It is also interesting to note that unlike the homogeneity analysis solution under the $\mathbf{pull}_2$ loss function, but with the category points normalized (i.e. $Y'\mathrm{diag}(G'G)Y = I$) shown in the right panel of Fig. 7, the patterns in the data are such that they give rise to a *planar layout* (edges do not intersect).

Acknowledgements

The authors would like to thank the AE and an anonymous referee for useful suggestions that improved the presentation. The work of George Michailidis was supported in part by NSF under grants IIS-9988095 and DMS-0214171.

References

De Leeuw, J., Michailidis, G., 2000a. Graph layout techniques and multidimensional data analysis. In: Bruss, F.T., Le Cam, L. (Eds.), Game Theory, Optimal Stopping, Probability and Statistics. Papers in honor of Thomas S. Ferguson, IMS Lecture Notes-Monograph Series, Harvard, CA, pp. 219–248.

De Leeuw, J., Michailidis, G., 2000b. Block relaxation algorithms in statistics. J. Comput. Graphical Statist. 9, 26–31.

De Leeuw, J., Michailidis, G., 2003. Weber correspondence analysis: the one dimensional case. Technical Report #343, Department of Statistics, UCLA.

De Leeuw, J., van Rijckevorsel, J.L.A., van der Wouden, H., 1980. Nonlinear principal components analysis with B-splines. Methods Oper. Res. 43, 379–394.

Drezner, Z. (Ed.) 1995. Facility Location. Springer, New York.

Edelman, A., Arias, T.A., Smith, S.T., 1999. The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20, 303–353.

Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, Chichester.

Guttman, L., 1968. A general non-metric technique for fitting the smallest coordinate space for a configuration of points. Psychometrika 33, 469–506.

Heiser, W.J., 1987. Correspondence analysis with least absolute residuals. Comput. Statist. Data Anal. 5, 337–356.

Huber, P., 1981. Robust Statistics. Wiley, New York.

Kiers, H.A.L., 1997. Weighting least squares fitting using ordinary least squares algorithms. Psychometrika 62, 251–266.

Kuhn, H.W., 1967. On a pair of dual nonlinear programs In: Abadie, J., (Ed.), Methods of Nonlinear Programming, 37054, North Holland, Amsterdam.

Lange, K., Hunter, D.R., Yang, I., 2000. Optimization transfer algorithms in statistics (with discussion). J. Comput. Graphical Statist. 9, 1–50.

Michailidis, G., de Leeuw, J., 1998. The Gifi system for descriptive multivariate analysis. Statist. Sci. 13, 307–336.

Michailidis, G., de Leeuw, J., 2000. Homogeneity analysis by alternating least absolute deviations. Technical Report, Department of Statistics, UCLA.

Michailidis, G., de Leeuw, J., 2001. Data visualization through graph drawing. Comput. Statist. 16, 435–450.

Tutte, W.T., 1963. How to draw a graph. Proc. London Math. Soc. 13, 743–767.

Vardi, Y., Zhang, C.H., 2001. A modified Weiszfeld algorithm for the Fermat–Weber location problem. Math. Programming 90, 559–566.

Verboon, P., 1994. A Robust Approach to Nonlinear Multivariate Analysis. DSWO Press, Leiden.

Voss, H., Eckhardt, U., 1980. Linear convergence of generalized Weiszfeld's methods. Computing 25, 243–251.

Weiszfeld, E., 1937. Sur le Point par laquel la Somme des Distances de n Points Donnés Est Minimum. Tohoku Math. J. 43, 355–386.

Witzgall, C.J., 1964. Optimal location of a central facility: mathematical models and concepts. Technical Report, National Bureau of Standards.