HOMOGENEITY ANALYSIS USING ABSOLUTE DEVIATIONS

GEORGE MICHAILIDIS AND JAN DE LEEUW

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 drawing is computed by an alternating least squares algorithm. Heiser (1987) looked at homogeneity analysis under a weighted least absolute deviations criterion. In this paper, we take a closer 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 $(p \geq 2)$ 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.

1. INTRODUCTION

Homogeneity Analysis (also known as Multiple Correspondence Analysis (MCA)) is a well-known technique to make graphical representations of categorical multivariate data [7]. It can also be presented as a technique to produce informative layouts of *bipartite* graphs [14, 2].

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 $W = \{w_{ij}|,\ i = 1, ..., N,\ j = 1, ..., K\}$. It can be easily shown that the 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

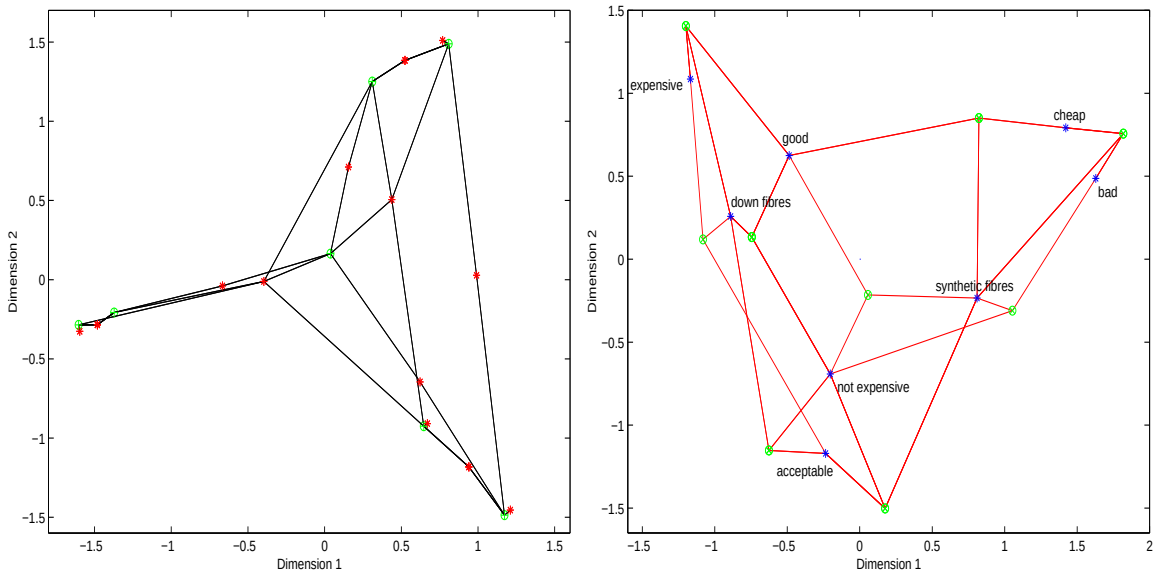

FIGURE 1. Left panel: Guttman-Bell graph drawing, with $*$ denoting the object points and $\diamond$ the category points. Right panel: Sleeping bags graph drawing. Both examples illustrate quite clearly the centroid principle.

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 [7, 14].

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 [7], 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 [7]. 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 Figure 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 [14], as can be seen on the left part of the picture for the Guttman-Bell drawing and (ii) the presence of *horseshoes* [1].

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 [15], 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 [10]. The case (1.2) was discussed earlier in [9] 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 [9, page 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 Figure 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 emerges for the object points; i.e. all of 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 [16]. Analogously, the 3-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 [4]. 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.

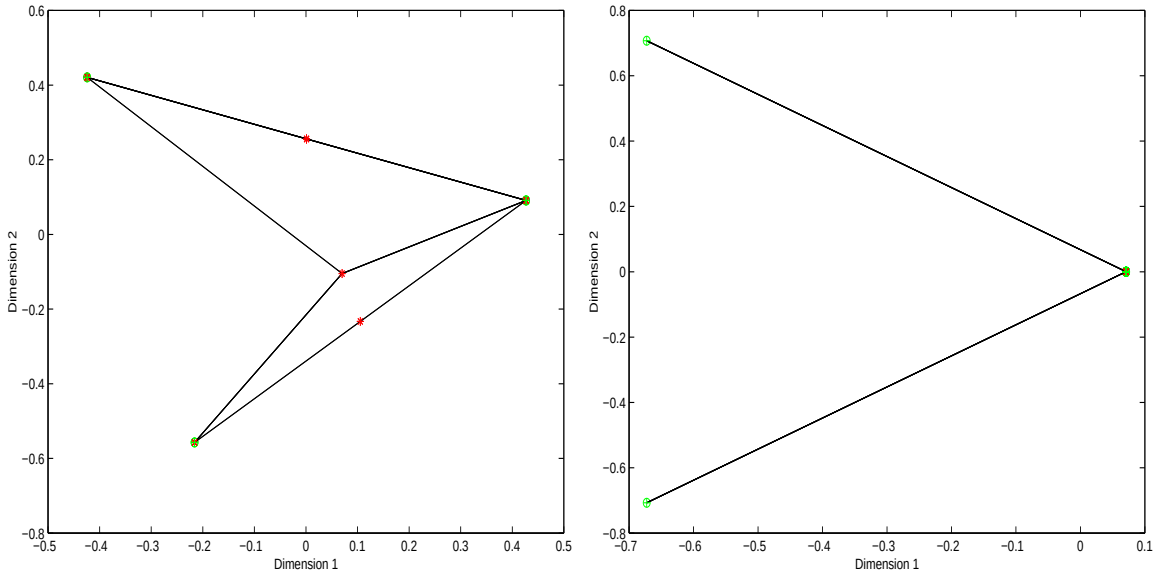

FIGURE 2. Left panel: Guttman-Bell graph drawing under loss function (1.2); Right panel: sleeping bags graph drawing.

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 $\mathcal{G}$.

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

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

throughout, where ϵ is small, and we minimize

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

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

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 i^{th} (j^{th}) 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

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

In matrix notation we can write

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

with

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

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

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

and

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

Observe that $A_\epsilon(X,Y)$ and $B_\epsilon(X,Y)$ 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

$$\mathbf{pull}(\epsilon)\overset{\Delta}{=}\min_{X'X=I}\min_Y\mathbf{pull}_{(1,\epsilon)}(X,Y).$$

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

Proposition 2.1. 1. *The objective function* $\mathbf{pull}(\epsilon)$ *is increasing in the parameter* ϵ.
2. $\lim_{\epsilon\to 0}\mathbf{pull}(\epsilon)=\mathbf{pull}(0)$.

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

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

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 $|\mathbf{pull}(\epsilon)-\mathbf{pull}(0)|=\sqrt{\epsilon}$ and the result follows. □

Experience has shown that for values of $\epsilon<10^{-5}$ 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 [3].

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 $k = 0, 1, \ldots$

$$\begin{aligned} Y^{(k)} &= \text{argmin}_Y \big(\mathbf{pull}(X^{(k)}, Y)\big), \\ X^{(k+1)} &= \text{argmin}_{X'X=I} \big(\mathbf{pull}(X, Y^{(k)})\big). \end{aligned}$$

The first subproblem, updating Y, due to the Euclidean distance function used, amounts to solving K separate Weber problems [19]. To find the coordinates in $\mathbf{R}^p$ of category point y_j we minimize

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

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 [12]. The classical algorithm is the one by Weiszfeld [22], which is a linearly convergent majorization method [19, 21]. 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. [6] could then be used.

A second approach can be based on the concept of majorization [13, 3]. By the Arithmetic Mean/Geometric Mean inequality we have that

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

and thus

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

This implies

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

where

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

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

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

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

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

with

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

and with Λ a diagonal matrix containing the p smallest eigenvalues of the matrix $D_\epsilon(\tilde{X},\tilde{Y})$. It is worth noting that the smallest eigenvalue is 0, since both the rows and the columns of $D_\epsilon(\tilde{X},\tilde{Y})$ 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 $D_\epsilon(X,Y)$, while X is the corresponding set of eigenvectors. The matrix Y contains the weighted centroid $B_\epsilon^{-1}(X,Y)C_\epsilon'(X,Y)X$, 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 $D_\epsilon(X,Y)$, 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

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

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

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

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

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

and the third term as where $V_\epsilon(\tilde{X},\tilde{Y}) \stackrel{\Delta}{=} A_\epsilon(\tilde{X},\tilde{Y})Y - C_\epsilon(\tilde{X},\tilde{Y})\tilde{X}$. Then, the second term above can be written (using the definition of $\overline{X}$) as

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

and the third term as

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

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

$$\begin{aligned}\mathbf{pull}(X,Y|\tilde{X},\tilde{Y}) \leq \mathbf{pull}(\tilde{X},Y|\tilde{X},\tilde{Y}) + \\ \alpha(\tilde{X},\tilde{Y})\mathbf{tr}\,(X-\overline{X})'(X-\overline{X}) - \frac{1}{\alpha(\tilde{X},\tilde{Y})}\mathbf{tr}\{V_\epsilon'(\tilde{X},\tilde{Y})V_\epsilon(\tilde{X},\tilde{Y})\}.\end{aligned}$$

Minimizing this second majorization over X is the same as minimizing $\mathbf{tr}\ (X-\overline{X})'(X-\overline{X})$, which is a so-called *orthogonal procrustes* problem, whose solution is classical. If $\overline{X} = K\Gamma L'$ is the singular value decomposition of $\overline{X}$, then the solution is $\hat{X} = KL'$. This is the algorithm proposed by [9], compare also [11].

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 [8] and in [7]) consists of 7 objects and 5 variables with a total of $K = 17$ categories. In figure 3 the homogeneity analysis solution under the $\mathbf{pull}-2$ loss function superimposed the corresponding solution under $\mathbf{pull}_1$ are given. The 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 the following table 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.

3.2. **Sleeping Bags.** This data set is taken from [2] 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 figure 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 [14]). In the following table the decomposition of the total loss for the optimal solution is given, along with the correspondence between the 8 category points and the 3

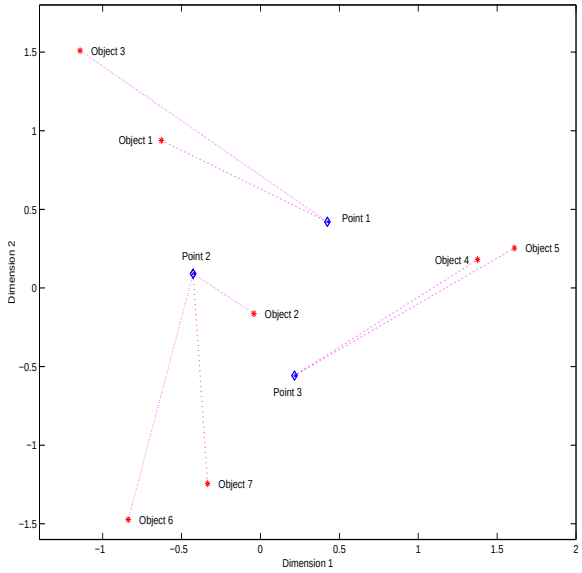

FIGURE 3. The homogeneity analysis solution for the Guttman-Bell data set under (1.1) (red * points) and the solution under (1.2) (blue diamond-shape points) and the correspondence between the two solutions

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

TABLE 3.1. Decomposition of the total loss for the Guttman-Bell data set.

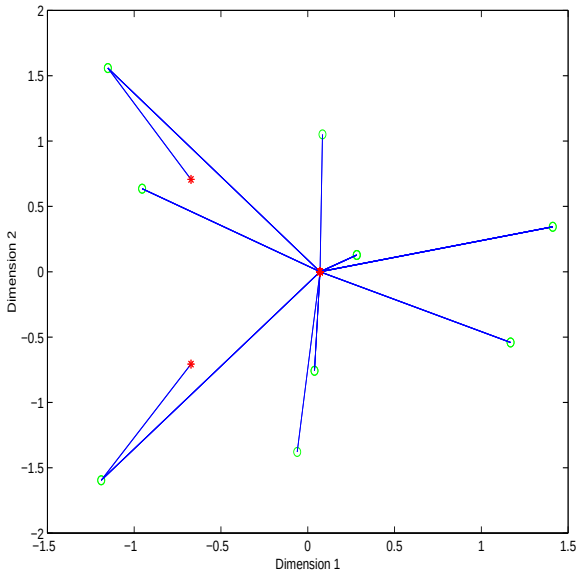

FIGURE 4. The homogeneity analysis solution for the sleeping bags data set under (1.1) (red * points) and the corresponding absolute deviations solution (blue diamond-shape points) and the correspondence between the two solutions

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

TABLE 3.2. Decomposition of the total loss for the sleeping bags data set.

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.

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 (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 the following Table the number of times the various

Data Set	A1	A2	A3
Guttman-Bell	72	88	84
Sleeping Bags	86	96	94

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 [16]) 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*

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

Proof. See [12], 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)**.** *If $w_k \geq \sum_{i \neq k} w_i$ then $\hat{y} = x_k$.*

Proof.

$$\| \sum_{i \neq k} w_i \frac{x_i - x_k}{d(x_i, x_k)} \| \leq \sum_{i \neq k} w_i \| \frac{x_i - x_k}{d(x_i, x_k)} \| = \sum_{i \neq k} w_i.$$

□

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 $\mathbf{pull}$ 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 $k_j \geq 3,\ j = 1, ..., J$, 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 $i = 2, ..., p+1$ and given that points $x_2, ..., x_{p+1}$ have L nonzero w_{ij}'s, the result follows. □

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

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

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 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

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

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 Figures 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.

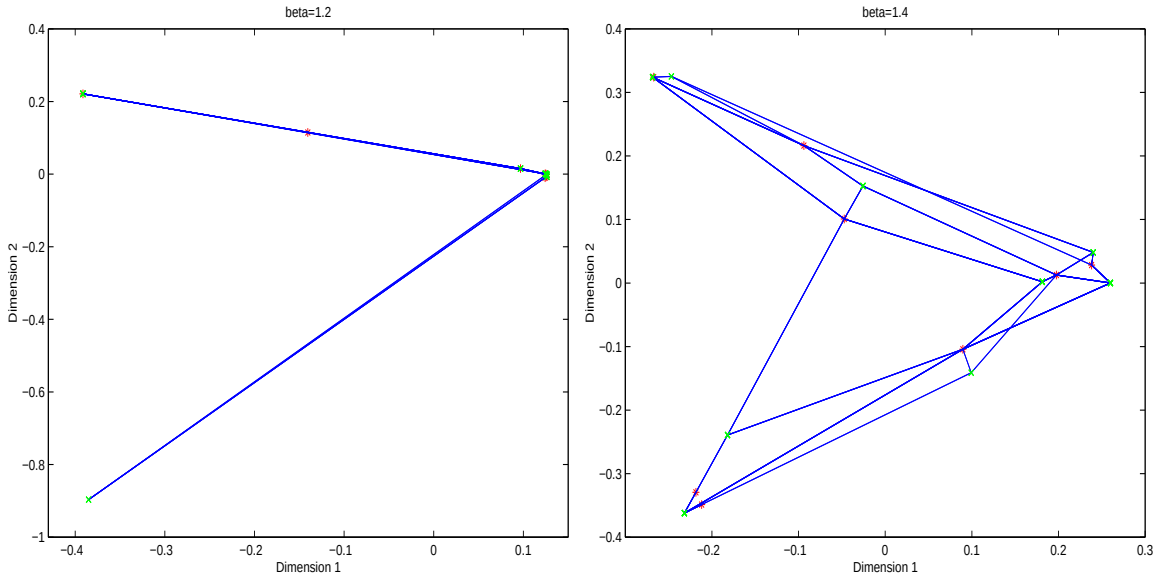

FIGURE 5. Left Panel: $\beta = 1.2$; Right panel: $\beta = 1.4$

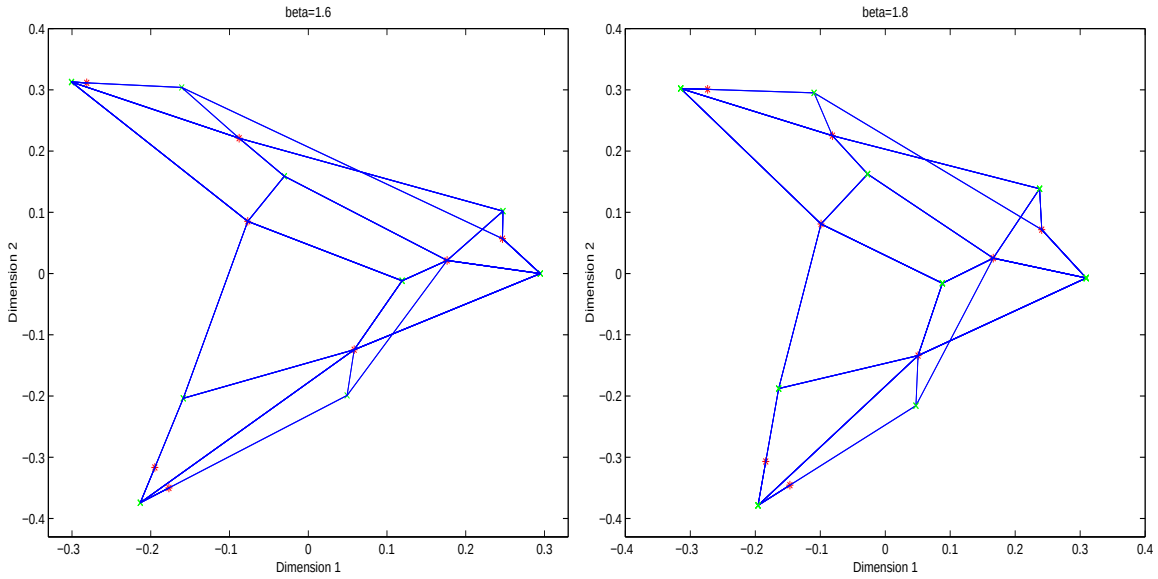

FIGURE 6. Left Panel: $\beta = 1.6$; Right panel: $\beta = 1.8$

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 [20] 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.

For loss functions that attempt to robustify the distances involved, the normalization constraint becomes highly problematic. The message of our investigations is that 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 [18] 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 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 [5], 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 Figure 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. $Y'\text{diag}(G'G)Y = I$) shown in the right panel of Figure 7, the patterns in the data are such that they give rise to a *planar layout* (edges do not intersect).

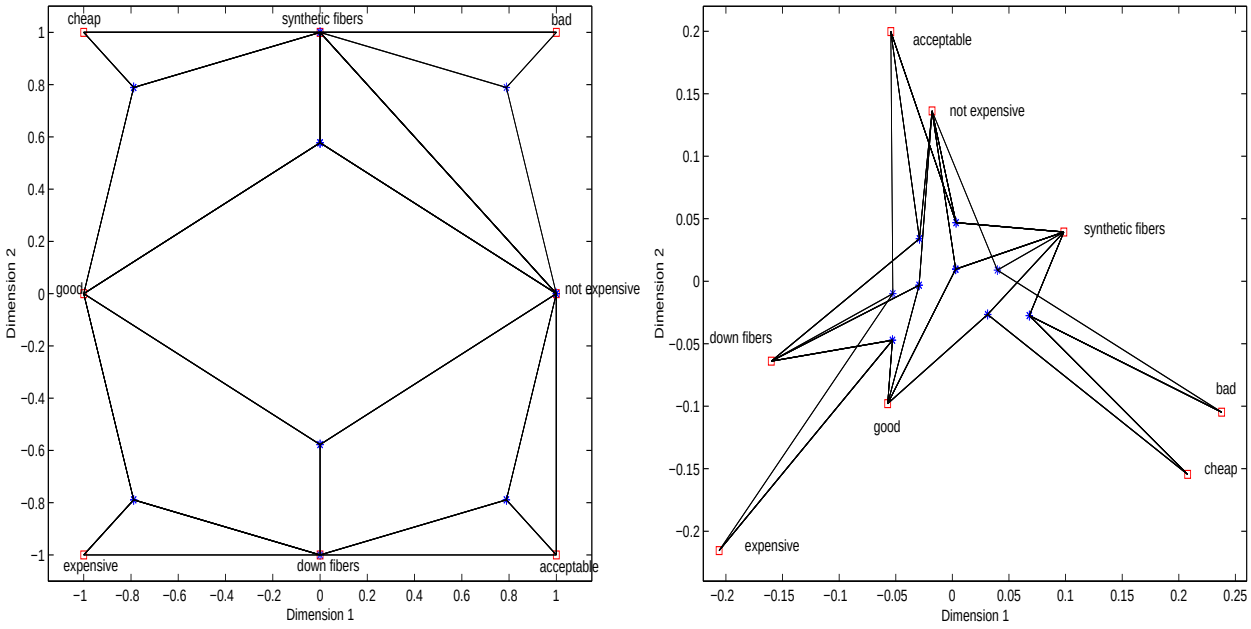

FIGURE 7. Left panel: graph layout of the sleeping bag data set under Tutte normalization on the categories; Right panel: graph layout of sleeping bag data set under the (1.1) loss function and with the category points normalized.

REFERENCES

[1] De Leeuw, J., van Rijckevorsel, J.L.A. & van der Wouden, H. (1980), Nonlinear Principal Components Analysis with B-splines, *Methods of Operations Research*, **43**, 379-394

[2] De Leeuw, J. & Michailidis, G. (2000), Graph Layout Techniques and Multidimensional Data Analysis, *Game Theory, Optimal Stopping, Probability and Statistics. Papers in honor of Thomas S. Ferguson*, F.T. Bruss and L. Le Cam (eds), IMS Lecture Notes-Monograph Series, 219-248
[3] De Leeuw, J. & Michailidis, G. (2000), Block Relaxation Algorithms in Statistics, *Journal of Computational and Graphical Statistics*, **9**, 26-31
[4] De Leeuw, J. & Michailidis, G. (2003), Weber Correspondence Analysis: The One Dimensional Case, Technical Report # 343, Department of Statistics, UCLA
[5] Drezner, Z. (ed) (1995), *Facility Location*, Springer, New York
[6] Edelman, A., Arias, T.A. & Smith, S.T. (1999), The Geometry of Algorithms with Orthogonality Constraints, *SIAM Journal of Matrix Analysis and Applications*, **20**, 303-353
[7] Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Chichester: Wiley
[8] Guttman, L. (1968), A General Nonmetric Technique for Fitting the Smallest Coordinate Space for a Configuration of Points, *Psychometrika*, **33**, 469-506
[9] Heiser, W.J. (1987), Correspondence Analysis with Least Absolute Residuals, *Computational Statistics and Data Analysis*, **5**, 337-356
[10] Huber, P. (1981), *Robust Statistics*, New York: Wiley
[11] Kiers, H.A.L. (1997), Weighting Least Squares Fitting Using Ordinary Least Squares Algorithms, *Psychometrika*, **62**, 251-266
[12] Kuhn, H.W. (1967), On a Pair of Dual Nonlinear Programs, *Methods of Nonlinear Programming*, J. Abadie (ed), 37054, Amsterdam: North Holland
[13] Lange, K., Hunter, D.R. and Yang, I. (2000), Optimization Transfer Algorithms in Statistics (with discussion), *Journal of Computational and Graphical Statistics*, **9**, 1-50
[14] Michailidis, G. & de Leeuw, J. (1998), The Gifi System for Descriptive Multivariate Analysis, *Statistical Science*, **13**, 307-336
[15] Michailidis, G. & de Leeuw, J. (2001), Data Visualization through Graph Drawing, *Computational Statistics*, **16**, 435-450
[16] Michailidis, G. & de Leeuw, J. (2000), Homogeneity Analysis by Alternating Least Absolute Deviations, Technical Report, Department of Statistics, UCLA
[17] Rockefellar, R. T. (1970), *Convex Analysis*, Princeton Univ. Press, Princeton, New Jersey.
[18] Tutte, W.T. (1963), How to Draw a Graph, *Proceedings of the London Mathematical Society*, **13**, 743-767
[19] Vardi, Y. & Zhang, C.H. (2001), A Modified Weiszfeld Algorithm for the Fermat-Weber Location Problem, *Mathematical Programming*, **90**, 559-566
[20] Verboon, P. (1994), *A Robust Approach to Nonlinear Multivariate Analysis*, DSWO Press, Leiden
[21] Voss, H. and Eckhardt, U. (1980), Linear Convergence of Generalized Weiszfeld's Methods, *Computing*, **25**, 243-251
[22] Weiszfeld, E. (1937), Sur le Point par laquel la Somme des Distances de n Points Donnés Est Minimum, *Tohoku Mathematics Journal*, **43**, 355-386
[23] Witzgall, C.J. (1964), Optimal Location of a Central Facility: Mathematical Models and Concepts, Technical Report, National Bureau of Standards

DEPARTMENT OF STATISTICS, UNIVERSITY OF MICHIGAN

E-mail address: gmichail@umich.edu

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES

E-mail address: deleeuw@stat.ucla.edu