Weber Correspondence Analysis: The One-Dimensional Case

Jan DE LEEUW and George MICHAILIDIS

1. INTRODUCTION

Correspondence analysis or CA can be interpreted as a technique for drawing weighted bipartite graphs (Michailidis and de Leeuw 2001). In the adjacency matrix of the bipartite graph we can restrict our attention to the upper diagonal block, which shows the adjacencies of the two sets of, say, n and m points having all the connections. Suppose $\mathbf{W}$, with $w_{ij} \geq 0$ for all $i = 1, \ldots n$ and $j = 1, \ldots, m$, is this upper diagonal submatrix.

The general idea is to make a drawing of the graph in which a large weight w_{ij} corresponds with a small Euclidean distance d_{ij}. In the usual versions of CA we actually use squared Euclidean distances, and we draw the graph by finding n points x_i and m points y_j in $\mathbb{R}^p$ such that

$$\sigma(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} d_{ij}^2(\mathbf{X}, \mathbf{Y}),$$

is minimized. We normalize the solution by requiring that $\mathbf{u}'\mathbf{X} = 0$, where $\mathbf{u}$ has all elements equal to $+1$, and $\mathbf{X}'\mathbf{X} = \mathbf{I}$.

In *Weber correspondence analysis*, or WCA, we use Euclidean distance, not its square, and minimize

$$\sigma(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} d_{ij}(\mathbf{X}, \mathbf{Y}),$$

using the same normalization conditions on $\mathbf{X}$. The WCA problem was first discussed by Heiser (1987), and we discuss it in general in Michailidis and de Leeuw (2003).

A more specialized version can be studied in which $\mathbf{W}$ is the adjacency matrix of an unweighted graph (i.e., $\mathbf{W}$ is binary). We call this the *binary case*. And even more specialized

Jan de Leeuw is Professor and Chair, Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095 (E-mail: deleeuw@stat.ucla.edu). George Michailidis is Associate Professor, Department of Statistics, The University of Michigan, Ann Arbor, MI 48109.

Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 946–953
DOI: 10.1198/106186004X15404

is the case in which $\mathbf{W}$ is the indicator super matrix of a multiple correspondence analysis problem. In that case $\mathbf{W}$ is also binary, but it can be partitioned columnwise into a number of submatrices, where each submatrix is an indicator matrix. (An indicator matrix is a binary matrix with orthogonal columns whose rows sum to one.) Thus, each submatrix codes a partitioning of the n row-elements into a number of subsets (equal to its number of columns) rows. If there are $\mathbf{M}$ submatrices, then all rows of $\mathbf{W}$ sum to $\mathbf{M}$. We call this the *indicator case*.

2. PROBLEM

In this article we discuss the special case $p = 1$, in which both sets of points are mapped into the real line. (Proofs for all results in this article are in the *JCGS* on-line repository at http://www.amstat.org/publications/jcgs/ftp.html.) As is the case in multidimensional scaling (de Leeuw and Heiser 1977) the WCA problem in one dimension turns out to be equivalent to a combinatorial optimization problem, more specifically a nonlinear zero-one optimization problem.

Let us reformulate the one-dimensional WCA problem explicitly. Suppose $\mathbf{W}$ is a nonnegative matrix, representing the data. The problem $\mathcal{P}$ we want to solve in this article is to minimize the loss function

$$\sigma(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} |x_i - y_j|$$

over $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, under the constraints that

$$\begin{aligned} \sum_{i=1}^{n} x_i &= 0, \\ \sum_{i=1}^{n} x_i^2 &= 1. \end{aligned}$$

3. TWO-POINT RESULT

Consider the $n!$ strict orderings of the x_i, and locate each y_j in one of $n - 1$ "interior" cells bordered by the ordered x_i. If y_j and y_ℓ are both placed between the same adjacent $x_{(i)} < x_{(i+1)}$, then the ordering of y_j and y_ℓ does not matter. This gives $n!(n-1)^m$ possible orderings. Each such choice corresponds with an $n \times m$ matrix $\mathbf{S}$ with elements ± 1 such that $s_{ij} = \text{sign}(x_i - y_j)$ and thus $s_{ij}(x_i - y_j) = |x_i - y_j|$. Let $\mathcal{S}$ be the set of such matrices.

Our original problem $\mathcal{P}$ is equivalent to minimizing

$$\sigma(x, y, \mathbf{S}) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} s_{ij} (x_i - y_j)$$

over x and y and $\mathbf{S} \in \mathcal{S}$, under the constraints

$$\sum_{i=1}^{n} x_i = 0,$$
$$\sum_{i=1}^{n} x_i^2 = 1,$$

and

$$s_{ij}(x_i - y_j) \geq 0.$$

Let us look at this problem, say $\mathcal{P}(S)$, for a fixed $\mathbf{S}$.

Lemma 1. *Problem $\mathcal{P}(\mathbf{S})$ is the minimization of a linear function of (x, y) over $K \cup \Gamma$, where K is a pointed convex polyhedral cone and Γ is the sphere $\{x|x'x = 1\}$.*

A pointed cone is the convex hull of its one-dimensional faces (extreme half-rays) (Goldman and Tucker 1956). These extreme half rays are of the form

$$x_{(1)} = \cdots = x_{(i)} < x_{(i+1)} = \cdots = x_{(n)},$$

that is, x only has two different elements (one negative, one positive, with weighted sum equal to zero). The y_j are equal to either the upper of the lower bound of the interval in which the ordering places them. In any case, vectors on the extreme half-rays have precisely two different x elements, and each y element is equal to one of the two x elements.

Now combine linearity and structure of the cone to characterize optimality of the fixed $\mathbf{S}$ problem. We prove a lemma that is a little bit more general than actually needed.

Lemma 2. *Suppose K is a pointed polyhedral convex cone and η is a pseudo-norm. Γ is the sphere of all x such that $\eta(x) = 1$. Suppose e_j, with $j = 1, \ldots, m$ are the extreme half-rays of K, normalized such that $\eta(e_j) = 1$. Consider the problem of minimizing $w'x$ over x such that $x \in K \cap \Gamma$. Then the minimum is attained at one of the e_j.*

Observe that η is a pseudo-norm, and thus it is possible that $\eta(e_j) = 0$ for some of the extreme half-rays, which then obviously cannot be normalized. But the proof still works. From this lemma we get our main result.

Theorem 1. *The solution to problem $\mathcal{P}$ has precisely two different x elements, and each y element is equal to one of the two x elements.*

Some simple calculations show that if value the negative value x_- occurs n_- times and the negative value x_+ occurs n_+ times, then

$$x_- = -\sqrt{\frac{n_+}{nn_-}},$$
$$x_+ = +\sqrt{\frac{n_-}{nn_+}}.$$

We call such a distribution of values an $n_- : n_+$ *allocation*. Observe that

$$x_+ - x_- = \sqrt{\frac{n}{n_- n_+}},$$

which is largest for $1 : n - 1$ allocations and smallest for $n/2 : n/2$ allocations.

The original problem becomes to determine n_- and n_+ and to assign the x_i and the y_j to the two different values. We can formalize this using an $n \times 2$ indicator matrix $\mathbf{G}$ and an $m \times 2$ indicator matrix $\mathbf{H}$, with x_- and x_+ in the two-element vector $\overline{\mathbf{x}}$. We summarize what we have shown so far in a corollary.

Corollary 1.

$$\sigma(\star,\star) \stackrel{\Delta}{=} \min_{x} \min_{y} \sigma(x,y) = \min_{\mathbf{G}} \min_{\mathbf{H}} \sigma(\mathbf{G}\overline{\mathbf{x}}, \mathbf{H}\overline{\mathbf{x}})$$

Here minimization is over normalized x. The indicator matrix $\mathbf{G}$ must be nondegenerate, in the sense that its columns sums are positive. These column sums are n_- and n_+, which means that $\overline{\mathbf{x}}$ is a function of $\mathbf{G}$. $\mathbf{H}$ can be any indicator matrix.

Suppose we know $\mathbf{G}$, that is, we know which x_i get the negative value and which get the positive value, and we know what these values are. Now find the optimal $\mathbf{H}$ for this given $\mathbf{G}$.

Theorem 2.

$$\sigma(\mathbf{G},\star) \stackrel{\Delta}{=} \min_{\mathbf{H}} \sigma(\mathbf{G},\mathbf{H}) = \sqrt{\frac{n}{n_- n_+}} \sum_{j=1}^{m} \min(u_j, v_j),$$

where u_j is the sum of the w_{ij} over the i for which $x_i = x_-$, while v_j is the sum of the remaining w_{ij}.

As a consequence we have the following combinatorial representation of the optimum value.

Corollary 2.

$$\sigma(\star,\star) = \min_{s=1}^{n-1} \sqrt{\frac{n}{s(n-s)}} \left\{ \frac{1}{2} w_{\bullet\bullet} - \min_{\mathbf{G}\in\mathcal{G}_s} \sum_{j=1}^{m} \left| \frac{1}{2} w_{\bullet j} - \sum_{i=1}^{n} w_{ij} g_i \right| \right\}.$$

4. SPECIFIC OPTIMUM ALLOCATIONS

4.1 OPTIMUM 1: $n - 1$ ALLOCATION

If $x_i = x_-$ and all other elements of x are x_+, then

$$\sigma(\mathbf{G},\star) = \sqrt{\frac{n}{n-1}} \sum_{j=1}^{m} \min(w_{ij}, w_{\bullet j} - w_{ij}).$$

This implies

$$\sigma(\mathbf{G},\star) \leq \sqrt{\frac{n}{n-1}} w_{i\bullet},$$

and we have equality if $w_{ij} \leq \frac{1}{2}w_{\bullet j}$ for all j, which will be the usual case. In that case we select i to correspond to the smallest row sum. But if only element i in column j of $\mathbf{W}$ is positive, and the other elements in the column are zero, then $\min(w_{ij}, w_{\bullet j} - w_{ij}) = 0$ and w_{ij} does not enter into the summation. For example, if $\mathbf{W}$ is any permutation matrix, then minimum loss is zero.

In the indicator case all row sums are equal to the number of indicator matrices $\mathbf{M}$, which means that all $1 : n - 1$ allocations have the same value

$$\mathbf{M}\sqrt{\frac{n}{n-1}}.$$

4.2 Optimum 2: $n - 2$ Allocation

Clearly

$$\sigma(\mathbf{G}, \star) = \sqrt{\frac{n}{2(n-2)}} \sum_{j=1}^{m} \min[w_{ij} + w_{kj}, w_{\bullet j} - (w_{ij} + w_{kj})]$$
$$\leq \sqrt{\frac{n}{2(n-2)}}(w_{i\bullet} + w_{k\bullet}).$$

If $(w_{ij} + w_{kj}) \leq \frac{1}{2}w_{\bullet j}$, then we have equality.

In the indicator case the upper bound is equal to

$$2\mathbf{M}\sqrt{\frac{n}{2(n-2)}}.$$

Since

$$2\mathbf{M}\sqrt{\frac{n}{2(n-2)}} > \mathbf{M}\sqrt{\frac{n}{n-1}},$$

which means that, in the "usual" case, $n : n-1$ allocation is better than $2 : n-2$ allocation.

If we do have $(w_{ij} + w_{kj}) > \frac{1}{2}w_{\bullet j}$ for some j then this is no longer true. In fact, the example

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

has optimum $1 : 3$ allocation equal to $2\sqrt{\frac{4}{3}} \approx 2.3094$ and optimum $2 : 2$ allocation equal to 2.

It is clear that in this case the optimal split depends on properties of the data, although the two-point representation is not very revealing from a data analysis point of view.

Table 1. Guttman-Bell

Allocation	*Min*	*Q1*	*Median*	*Q3*	*Max*
1:6	4.3205	4.3205	4.3205	5.4006	5.4006
2:5	2.5100	6.4841	6.6933	7.5299	8.3666
3:4	3.8188	5.3463	7.6376	8.9742	10.6927

5. REAL EXAMPLES

One obvious algorithm is to solve the problem for each $s = 1, \dots, n-1$ separately. In fact, by symmetry, we only have to consider the values less than or equal to $\frac{n}{2}$. For each s we evaluate the objective function for all $\binom{n}{s}$ possible binary vectors g summing to s. For fixed s the problem, by Corollary 2, we have to solve is the maximization of

$$\sum_{j=1}^{m} \left| \frac{1}{2} w_{\bullet j} - \sum_{i=1}^{n} w_{ij} g_i \right| ,$$

that is, the maximization of a convex piecewise linear function of g. We apply this enumerative algorithm to some examples. Observe that for n objects, we must compute the loss function for $2^{(}n-1) - 1$ subsets.

The Guttman-Bell data describe seven types of sociological groups in term of five variables. This is a really small example, and enumeration gives the results in Table 1. Columns of the table are the Tukey fivenums for the various allocations. We see that the optimal allocation is $2:5$ in this case.

Our next example, the sleeping bags data, was analyzed previously by ordinary MCA and other graph drawing techniques in Michailidis and de Leeuw (2001). Four variables are used to describe 21 sleeping bags. Table 2 shows that the $1:20$ allocation is optimal, and that all $1:20$ allocations have precisely the same value.

In the final example, 24 cars are described in terms of four safety features. After looking at more than 8 million subsets we find Table 3, in which again the $1:23$ allocation is the easy winner, and all $1:23$ allocations have the same value. It seems reasonable to assume that for examples that are large enough, this will always be what happens, unless something very special is going on (for instance, if the indicator matrices are permutation matrices).

Table 2. Sleeping Bags

Allocation	*Min*	*Q1*	*Median*	*Q3*	*Max*
1:20	3.0741	3.0741	3.0741	3.0741	3.0741
2:19	4.4604	4.4604	4.4604	4.4604	4.4604
3:18	3.7417	5.6125	5.6125	5.6125	5.6125
4:17	3.8900	6.6686	6.6686	6.6686	6.6686
5:16	4.0988	7.6852	7.6852	7.6852	7.6852
6:15	4.3474	8.2118	8.6948	8.6948	8.6948
7:14	5.0920	9.2582	9.7211	9.7211	9.7211
8:13	5.3923	9.4365	10.3352	10.7846	10.7846
9:12	5.7325	9.7011	10.5830	11.4649	11.9059
10:11	5.6801	10.0494	10.923	11.7971	13.1079

Table 3. Cars

	Min	*Q1*	*Median*	*Q3*	*Max*
1:23	4.0860	4.0860	4.0860	4.0860	4.0860
2:22	5.1698	5.9084	5.9084	5.9084	5.9084
3:21	5.5549	7.4066	7.4066	7.4066	7.4066
4:20	6.5727	8.2158	8.7636	8.7636	8.7636
5:19	7.0367	9.5499	10.0525	10.0525	10.0525
6:18	7.0711	10.3709	10.8423	11.3137	11.3137
7:17	7.1854	11.6763	12.1254	12.1254	12.5745
8:16	7.3612	12.1244	12.9904	13.4234	13.8564
9:15	7.1678	13.0707	13.4924	14.3357	15.1789
10:14	7.0387	13.6633	14.4914	14.9054	16.5616
11:13	6.9644	13.9289	14.7482	15.5676	17.6159
12:12	7.3485	13.8804	14.6969	15.5134	17.5547

This implies that WCA is almost always useless as a data analysis technique, because the solution will almost always be independent of the data (unless we have small or special examples).

6. ADDITIONAL ORTHOGONAL DIMENSIONS

Michailidis and de Leeuw (2003) discussed WCA for $p > 1$. The reduction to a zero-one programming problem, and the possibility to find the optimum solution by enumeration, no longer applies there. In fact, we cannot prove the $(p + 1)$-point property, which is the obvious generalization of the two-point property, although we have found it in all examples we have analyzed. In this section we will briefly discuss an alternative way to compute higher dimensional solutions, which does turn out to give us the $(p + 1)$-property. For the same reasons as for $p = 1$, the solution is not really of interest as a statistical technique, unless it is applied to small examples.

Suppose we have found a solution $(\overrightarrow{x}, \overrightarrow{y})$, and we want to find another solution minimizing the same loss function, under the same conditions, and in addition we impose the condition that $\overrightarrow{x}' x = 0$. Thus the new solution for x must be orthogonal to the previous one. Call this problem P_1.

Theorem 3. *Suppose (I_-, I_+) is the optimal assignment found in problem P. Then we solve problem P_1 by computing the optimal assignment for the subsets I_- and I_+, and keep the best of these two.*

Clearly we can continue the procedure in the theorem by using the optimal split in P_1 and continuing to split these subsets, thus finding an orthogonal third dimension. By continuing this procedure we build up a binary tree, and this binary tree can be used as a WCA representation of the data. Figure 1 illustrates this for the Guttman-Bell data, which is small enough to show some structure.

Of course, if P has a $1 : n - 1$ optimum assignment, then there is no need to split the first set, and we have seen that this will often happen. Thus, we will often split off one element at each stage, until the set becomes small enough. But, more seriously, we have also seen that usually all $1 : n - 1$ splits will have the same value, so the optimum is not

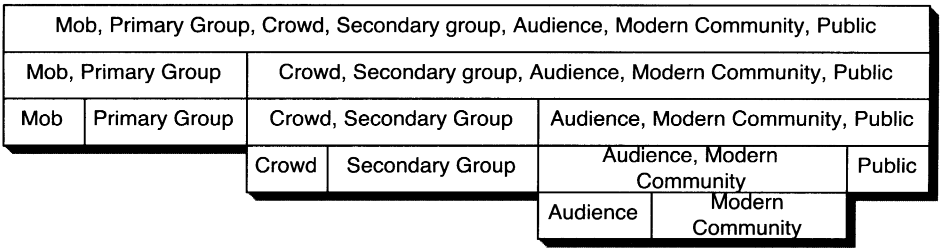

Figure 1. *Guttman-Bell Tree*

defined uniquely, and we must continue by building n trees, and so on. This will make tree building basically impossible, and again we find a very limited practical usefulness.

ACKNOWLEDGMENTS

The work of the second author is supported in part by NSF grant IIS-9988095, who also thanks the Department of Statistics and Actuarial Science of the University of Waterloo for its hospitality.

[Received March 2003. Revised January 2004.]

REFERENCES

de Leeuw, J., and Heiser, W. J. (1977), "Convergence of Correction Matrix Algorithms for Multidimensional Scaling," in *Geometric Representations of Relational Data*, ed. J. C. Lingoes, Ann Arbor: Mathesis Press, chap. 32.

Goldman, A. J., and Tucker, A. W. (1956), "Polyhedral Convex Cones," in *Linear Inequalities and Related Systems*, eds. H. W. Kuhn and A. W. Tucker, Princeton, NJ: Princeton University Press, chap. 2.

Heiser, W. J. (1987), "Correspondence Analysis with Least Absolute Residuals," *Computational Statistics and Data Analysis*, 5, 337–356.

Michailidis, G., and de Leeuw, J. (2001), "Data Visualization Through Graph Drawing," *Computational Statistics*, 16, 435–450.

——— (2003), "Homogeneity Analysis Using Absolute Deviations," Technical Report #346, Department of Statistics, UCLA.