WEBER CORRESPONDENCE ANALYSIS: THE ONE-DIMENSIONAL CASE

JAN DE LEEUW AND GEORGE MICHAILIDIS

1. INTRODUCTION

Correspondence Analysis (CA) can be interpreted as a technique for drawing weighted bipartite graphs [Michailides and de Leeuw, 2001]. The adjacency matrix of a bipartite graph takes the form

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

and therefore we can restrict our attention to the upper diagonal block W, that contains the connections between the two sets of vertices (points) of say, size n and m, respectively. Therefore, the $n \times m$ matrix $W = \{w_{ij}\}$, with $w_{ij} \geq 0$ for all $i = 1, \cdots n$ and $j = 1, \cdots, m$, becomes the focus of the analysis.

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 [Gifi, 1990] 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

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

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

Date: March 24, 2003.

In *Weber Correspondence Analysis* or WCA we use Euclidean distances, instead of their squares, and hence minimize

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

using the same normalization conditions on X. The WCA problem was first discussed by Heiser [1987], and we discuss it in general in Michailides and de Leeuw [2003].

A more specialized version can be studied in which W is the adjacency matrix of an unweighted graph and therefore binary (i.e. $w_{ij}\in\{0,1\}$). We call it the *binary case*. An even more specialized case corresponds to W being the super-indicator matrix of a multiple correspondence analysis problem [Gifi, 1990]. In this case W is also binary, but it can be partitioned column-wise into a number of submatrices, where each submatrix is an indicator matrix[1]. Thus each submatrix codes a partitioning of the n row-elements into a number of subsets (equal to its number of columns). If there are M submatrices, then all rows of W sum to M. We call this the *indicator case*.

2. Problem Formulation

In this paper we discuss the special case $p=1$, in which both sets of points are mapped into the real line. 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 W is a non-negative matrix, representing the data. The problem $\mathcal{P}$ we want to solve in this paper is to minimize the loss function

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

[1] An indicator matrix is a binary matrix with orthogonal columns whose rows sum to one.

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

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

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 S with elements ± 1 such that $s_{ij} = \mathbf{sign}(x_i - y_j)$ 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

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

over x and y and $S \in \mathcal{S}$, 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 examine this new problem $\mathcal{P}(S)$, for a fixed S.

Lemma 1. *Problem $\mathcal{P}(S)$ 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 $\{x | x'x = 1\}$.*

Proof. In the first place,

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

which is linear in x and y.

Second, the set of all x and y such that

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

and

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

is a polyhedral convex cone, say K. Because $s_{ij}(x_i - y_j) = 0$ and $\sum_{i=1}^{n} x_i = 0$ has the unique solution $x = y = 0$, the cone K is pointed. □

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

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

i.e. 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.

We combine next the linearity and the structure of the cone in order to characterize the optimality of the fixed $\mathcal{P}(S)$ problem. We prove a lemma that is a little bit more general than actually needed in the present setting.

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 $j = 1, \cdots, m$ 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.*

Proof. Each feasible point can be written in the form $x = \eta(\hat{x})^{-1}\hat{x}$, where $\hat{x}$ is a convex combination, with coefficients α_j, of the e_j. Because

$$w'\hat{x} = \sum_{j-1}^{m} \alpha_j w' e_j \geq \min_{j=1}^{m} w' e_j,$$

and

$$\eta(\hat{x}) \leq \sum_{j=1}^{m} \alpha_j \eta(e_j) = 1,$$

we have

$$w'\tilde{x} = \frac{w'\hat{x}}{\eta(\hat{x})} \geq \min_{j=1}^{m} w' e_j.$$

□

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. Nevertheless the result still holds.

Theorem 3. *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.*

Proof. For each fixed S the solution of $\mathcal{P}(S)$ for x and y has the structure stated in the Theorem. Therefore, the solution for x and y has also this structure for the optimal S. □

Due to the normalization constraint, one of the x values must be negative and is denoted by $x_\_$ and the other one must be positive and denoted by x_+. Some simple calculations show that if the negative value $x_\_$ occurs $n_\_$ times and the positive value x_+ occurs n_+ times, then

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

with $n = n_+ + n_-$. 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 G and an $m \times 2$ indicator matrix H, with x_- and x_+ in the two element vector $\overline{x}$. We summarize what we have shown so far in a corollary.

Corollary 4. *The optimal value for problem $\mathcal{P}(S)$ is given by*

$$\sigma(\star,\star) \stackrel{\Delta}{=} \min_x \min_y \sigma(x,y) = \min_G \min_H \sigma(G\overline{x}, H\overline{x})$$

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

4. Eliminating H

Suppose we know G, i.e. we know which x_i get the negative value and which get the positive value, which in turn implies that we can calculate these values. Now the problem becomes to find the optimal H for this given G.

Theorem 5. *For given G, the optimal value of the loss function σ is given by*

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

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

Proof. That problem can obviously be solved for each y_j, i.e. for each row of H, separately. For a fixed j we must minimize

$$\sum_{i=1}^n w_{ij}|x_i - y| = u_j|x_- - y_j| + v_j|x_+ - y|,$$

For y between x_- and x_+ we have

$$u_j(y - x_-) + v_j(x_+ - y) = (u_j - v_j)y - (u_j x_- - v_j x_+).$$

If $u_j > v_j$ then the function increases on the interval, and thus it attains its minimum at x_-, with value $v_j(x_+ - x_-)$. If $u_j < v_j$ then the optimal y_j is x_+, and the optimum value is $u_j(x_+ - x_-)$. If $u_j = v_j$ then the optimal value can be anything in the interval, because the function has slope zero. □

Corollary 6. *The optimal value for problem $\mathcal{P}(S)$ is given by*

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

Proof. Observe that by using $\min(a,b) = \frac{1}{2}(a + b - |a - b|)$ the result in Theorem 5 can also be written as

$$\sigma(G,\star) = \sqrt{\frac{n}{n_- n_+}} \sum_{j=1}^{m} \left\{ \frac{1}{2} w_{\bullet j} - |\frac{1}{2} w_{\bullet j} - v_j| \right\},$$

where we replace an index by a bullet if we sum over the index. If $\mathcal{G}_s$ is the set of indicator matrices for $s : n - s$ allocations, then minimizing our loss function means solving the problem in the corollary, where g_i are the elements of the first column of G, which sum to s for G in $\mathcal{G}_s$. □

We examine next some specific optimal allocations.

5. Specific Optimum Allocations

5.1. **Optimum** $1 : n - 1$ **allocation.** If $x_i = x_-$ and all other elements of x are x_+, then

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

This implies

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

and we have equality if $w_{ij} \leq \frac{1}{2} w_{\bullet j}$ for all j, which will be the most common case. In that case, we select i to correspond to the smallest row sum.

But if only element i in column j of 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 W is any permutation matrix, then the minimum loss is zero.

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

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

5.2. **Optimum** $2 : n-2$ **allocation.** Clearly

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

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

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

Since

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

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

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

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.

6. Real Examples

We apply next WCA to some real data sets. One obvious algorithm to find the optimum solution is to solve the problem for each $s = 1, \cdots, n-1$ 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 that we have to solve is, due to Corollary 6, the maximization of

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

i.e. 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.

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

The Guttman-Bell data [Gifi, 1990] describe seven types of sociological groups in terms of five variables. This is a really small example, and enumeration gives the results in Table 1. Columns of the table are the Tukey five numbers 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 Michailides and de Leeuw [2001]. Three 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.

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

Allocation	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

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 data sets with a fairly large number of observations, the above pattern would be most likely to occur, unless something very special is going on (for instance, if the indicator matrices are permutation matrices). 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).

7. Additional Orthogonal Dimensions

In Michailides and de Leeuw [2003] we discuss 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 it is present 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 data analysis technique, unless it is applied to small examples.

Suppose we have found a solution $(\overrightarrow{x}, \overrightarrow{y})$ to problem $\mathcal{P}$, and we are interested in obtaining a second solution that minimizes the same loss function under the same normalization constraints as before, but in addition we impose the condition that $\overrightarrow{x}'x = 0$. Thus the new solution for x must be orthogonal to the previous one. We call this problem $\mathcal{P}_1$.

Theorem 7. *Suppose (I_-, I_+) is the optimal assignment found in problem $\mathcal{P}$. Then we solve problem $\mathcal{P}_1$ by computing the optimal assignment for the subsets I_- and I_+, and keep the best of these two.*

Proof. We follow the same reasoning as before, i.e. we first solve problem P_1 for a fixed sign matrix S, which still means finding the minimum of a linear function over a now generally smaller pointed convex cone $K_1 \subseteq K$. For all the two-valued edges e_s of our previous cone we compute $\alpha_s \triangleq \overrightarrow{x}'e_s$. If $\alpha_s = 0$ we keep the edge in our list of edges, and it addition we add $\alpha_s e_t - \alpha_t e_s$ for all pairs (s,t) such that $\alpha_s > 0$ and $\alpha_t < 0$. The edges of the smaller cone K_1 are now in our new list [Uzawa, 1958]. Observe that the edges now have at most three different values.

To compute the values of these edges for a given assignment, we have the two possibilities shown in the table below. In the first case we must have

n_{11}	x_{-}	a
n_{12}	x_{+}	b
n_{13}	x_{+}	c

n_{11}	x_{-}	a
n_{21}	x_{-}	b
n_{31}	x_{+}	c

$n_{11} = n_{-}$ and $n_{12} + n_{13} = n_{+}$. Also

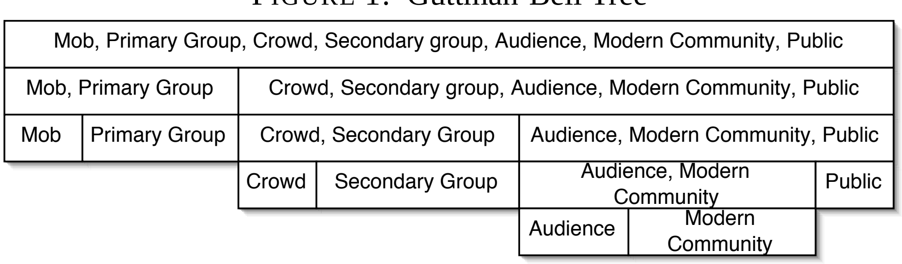
$$
\begin{aligned}
n_{11}ax_{-} + n_{12}bx_{+} + n_{13}cx_{+} &= 0, \\
n_{11}a + n_{12}b + n_{13}c &= 0,
\end{aligned}
$$

which implies $a = 0$. But this means that we find the optimal assignment, and the corresponding values of x and y by eliminating the $n_{11} = n_{-}$ rows for which $x_i = x_{-}$ and solving the original optimization problem for the remaining n_{+} rows. This will give us an optimal split for those rows, and an orthogonal dimension.

The second possibility in the table is solved in the same way, except that we split the set of indices corresponding with x_{+}. The required optimum under orthogonality constraints is then the best solution of the two. The optimal two-dimensional solution can be plotted as three points in the plane. □

Clearly we can continue this procedure 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. We illustrate this for the Guttman-Bell data, which is small enough to show some structure.

FIGURE 1. Guttman-Bell Tree



8. Discussion and Concluding Remarks

In this paper the problem of CA under Euclidean distances on the line is examined. It is shown that similarly to multidimensional scaling (MDS), it corresponds to a combinatorial optimization structure. However, unlike MDS, the optimal solution is most often data independent and thus of limited practical use.

For example, as shown in two real data sets, the optimal assignment is given by the $1 : n - 1$ allocation for problem $\mathcal{P}$. In this case, when one is interested in looking at additional dimensions, there is no need to split the first set, but proceed directly to solve problem $\mathcal{P}$ again for the second set of points. Experience has shown that only one element would be split off at each subsequent stage, until the set becomes small enough. However, an equally serious problem is that all $1 : n - 1$ allocations give the same optimum value for the objective function; hence, the optimal allocation is not uniquely defined and we have to resort to building a forest of n binary trees. Given the combinatorial nature of the problem, these difficulties severely limit the usefulness of WCA as a data analytic technique.

The message of the paper is that the normalization constraint $X'X = I$ in the context of CA is suitable for squared Euclidean distances, since it leads to an eigenvalue problem. The main difference with MDS where Euclidean distances work is that in MDS the objective is to approximate as well as possible graph-theoretic distances by Euclidean distances, while in CA the emphasis is on the connections between the points as represented by W. Therefore, the above normalization constraint must be abandoned if one is interested in using Euclidean distances in the adjacency model implied by CA.

References

J. de Leeuw and W.J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. In J.C. Lingoes, editor, *Geometric Representations of Relational Data*, chapter 32. Mathesis Press, Ann Arbor, Michigan, 1977.

A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, New York, 1990.

A.J. Goldman and A.W. Tucker. Polyhedral convex cones. In H.W. Kuhn and A.W. Tucker, editors, *Linear Inequalities and Related Systems*, chapter 2. Princeton University Press, Princeton, New Jersey, 1956.

W. J. Heiser. Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis*, 5:337–356, 1987.

G. Michailides and J. de Leeuw. Data visualization through graph drawing. *Computational Statistics*, 16:435–450, 2001.

G. Michailides and J. de Leeuw. Homogeneity analysis using absolute deviations. 2003.

H. Uzawa. A theorem on convex polyhedral cones. In K.J. Arrow, L. Hurwicz, and H. Uzawa, editors, *Studies in Linear and Non-linear Programming*, chapter 2. Stanford University Press, Stanford, California, 1958.