HOMOGENEITY ANALYSIS USING EUCLIDEAN MINIMUM SPANNING TREES

JAN DE LEEUW

1. INTRODUCTION

In *homogeneity analysis* the data are a system of m subsets of a finite set of n *objects*. The purpose of the technique is to locate the objects in low-dimensional Euclidean space in such a way that the points in each of the m subsets are close together. Thus we want the subsets to be *compact* or *homogeneous*, relative to the overall size of the configuration of the n points.

In the usual forms of homogeneity analysis (which are also known as *multiple correspondence analysis* or MCA) the size of a point set is defined as the size of the *star plot* of the set. The star plot is the graph defined by connecting all points in the set with their centroid, and the size is the total squared length of the edges. In order to prevent trivial solutions the configuration X is usually centered and normalized such that $X'X = I$. Minimizing total size of the star plots under the normalization constraints amounts to solving a (spare) singular value decomposition problem.

Recently, there have been a number of proposals to use different measures of point set size [De Leeuw, 2003]. One reason for looking at alternatives is the well known sensitivity of squared distances to outliers. Ordinary homogeneity analysis tends to locate subsets with only a few points way on the outside of the plot, which means that they dominate the solution by determining the scale. A second reason is that ordinary homogeneity analysis tends to create *horseshoes* for many types of data sets, i.e. two dimensional representations in which the second dimension is a quadratic function of the first [Schriever, 1985; Rijckevorsel, 1987; Bekker and Leeuw, 1988]. This is generally seen as wasteful, because we use two dimensions to represent an essentially one dimensional structure.

Date: December 14, 2003.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

The most straightforward modification of MCA is to measure the size of the star plots by using distances instead of squared distances [Heiser, 1987; Deleeuw and Michailides, 2003; Michailides and Deleeuw, 2003]. If we continue to apply the normalization $X'X = I$ then the solutions to this modified problem map all n objects into $p + 1$ points if we compute a representation in p dimensions. This means the solutions are only marginally dependent on the data and do not show sufficient detail to be interesting. From the data analysis point of view they are close to worthless.

In this paper we study another measure of size, which is quite different from the star plot based ones. The size of a point set is defined as the length of the Euclidean minimum spanning tree, and our loss function is the sum of these lengths over all m trees. We continue to use the same normalization as in MCA, i.e. we only consider centered and orthonormal X.

2. ALGORITHM

The problem we study is to minimize

$$\sigma(X) = \sum_{j=1}^{m} \min_{W \in \mathcal{W}_j} \mathbf{tr}\ W_j D(X)$$

over $X'X = I$, where $\mathcal{W}_j$ is the set of adjacency matrices for the spanning trees of the subset j and $D(X)$ is the matrix of Euclidean distances.

One obvious algorithm is to apply block relaxation [De Leeuw, 1994] to

$$\sigma(X, W_1, \cdots, W_m) = \sum_{j=1}^{m} \mathbf{tr}\ W_j D(X).$$

Thus we minimize our loss function by alternating minimization over X with minimization over the W_j.

It is easy to deal with the subproblem of minimimizing over $W_j \in \mathcal{W}_j$ for X fixed at its current value, because that is simply computation of the minimum spanning tree, for which we can use Kruskal's or Prim's algorithm [Graham and Hell, 1985]. The problem of minimizing loss over normalized X for fixed W_j is somewhat more complicated, but we can solve it using majorization De Leeuw [1994]; Heiser [1995]; Lange et al. [2000].

We first regularize the problem to deal with the possibility of zero distances, because in such configurations our loss function is not differentiable. Suppose $\epsilon > 0$ is a fixed small number. For each pair of points i and k we define

$$d_{ik}(X,\epsilon) = \sqrt{(x_i - x_k)'(x_i - x_k) + \epsilon}.$$

Clearly this regularized distance is everywhere positive and continuously differentiable in X.

As in Heiser [1987] we now apply the Arithmetic-Geometric mean inequality to obtain, for two different configurations X and Y,

$$d_{ik}(X,\epsilon) \leq \frac{1}{d_{ik}(Y,\epsilon)}\frac{1}{2}(d_{ik}^2(X,\epsilon) + d_{ik}^2(Y,\epsilon)).$$

We use the representation, which is standard in multidimensional scaling literature,

$$d_{ik}^2(X,\epsilon) = d_{ik}^2(X) + \epsilon = \mathbf{tr}\ X'A_{ik}X + \epsilon,$$

where A_{ik} is defined in terms of the unit vectors e_i and e_k as

$$A_{ik} = (e_i - e_k)(e_i - e_k)'.$$

Using this notation

$$\sigma(X,\epsilon) = \sum_{j=1}^m \mathbf{tr}\ W_j D(X,\epsilon) \leq$$

$$\leq \frac{1}{2}(\mathbf{tr}\ X'B(Y,\epsilon)X + \mathbf{tr}\ Y'B(Y,\epsilon)Y) + \epsilon\phi(Y,\epsilon),$$

where we define the matrix

$$B(Y,\epsilon) = \sum_{i=1}^n \sum_{k=1}^n \frac{\sum_{j=1}^m w_{jik}}{d_{ik}(Y,\epsilon)} A_{ik},$$

and the scalar

$$\phi(Y,\epsilon) = \sum_{i=1}^n \sum_{k=1}^n \frac{\sum_{j=1}^m w_{jik}}{d_{ik}(Y,\epsilon)}.$$

Given an intermediate solution $X^{(\nu)}$, the majorization algorithm finds an update $X^{(\nu+1)}$ by minimizing $\mathbf{tr}\ X'B(X^{(\nu)},\epsilon)X$ over $X'X = I$.

Convergence follows from general results on majorization methods, and in the particular case we can use the fact that $\sigma(X,\epsilon)$ is decreasing in ϵ to show, in addition , that letting $\epsilon \to 0$ gives convergence to the solution of the original non-regularized problem.

In general we do not iterate the majorization algorithm, which finds an optimal X for given W_j, until convergence, but we just take a single step before computing new minimum spanning trees. An implementation of the algorithm, in the R programming language, is given in the Appendix. For the MST calculation it depends on the ape package from the CRAN repository, written by Paradis, Strimmer, Claude, Jobb, Noel, and Bolker.

3. EXAMPLES

4. A MODIFICATION

The algorithm simplifies greatly if we define our minimum spanning trees by using the square of the Euclidean distance. In fact, we need neither regularization nor majorization. We have

$$\sigma(X, W_1, \cdots, W_m) = \sum_{j=1}^m \mathbf{tr}\ W_j D^2(X) = \mathbf{tr}\ X'VX,$$

where

$$V = \sum_{i=1}^n \sum_{k=1}^n \sum_{j=1}^m w_{jik} A_{ik}.$$

Now minimizing over normalized X for fixed W_j means finding eigenvectors corresponding with the smallest non-zero eigenvalues of V. And computing the minimum over the W_j fox fixed X means computing the MST for the squared distances.

Because there are only a finite number of spanning trees, and the algorithm stops if the loss does not decrease, it is clear that in this case we have convergence in a finite number of steps. This sounds much better than it is, because examples show us that the finite convergence gets us to different stationary points from different random starts.

5. DISCUSSION

There are many variations possible. For instance travelling salesman tour. For instance squared distance MST.

APPENDIX A. CODE

```
require(homals)
require(ape)

hommst<-function(g,pow=1,eps=1e-10) {
n<-dim(g)[1]; m<-dim(g)[2]
x<-svd(matrix(rnorm(2*n),n,2))$u
repeat{
        d<-eudist(x,pow,eps)
        c<-matrix(0,n,n); s<-0.0
        for (j in 1:m)
                {
                tmp<-spanner(as.factor(g[,j]),d)
                s<-s+tmp$s; c<-c+tmp$c
                }
        b<-bmat(c,d,pow)
        x<-eigen(b)$vectors[,c(n-1,n-2)]
        print(s)
        for (j in 1:m) spanplot(as.factor(g[,j]),x,eudist(x,pow,eps))
        }
}

spanner<-function(g,d) {
lev<-levels(g); nn<-length(g)
s<-0.0; c<-matrix(0,nn,nn);
for (k in lev) {
        ind<-which(k==g)
        n<-length(ind)
        if (n==1) dd<-mm<-matrix(0,1,1)
        else {
                dd<-d[ind,ind]
                mm<-mst(dd) }
        c[ind,ind]<-mm
        s<-s+sum(dd*mm)
        }
list(c=c,s=s)
}

spanplot<-function(g,x,d) {
plot(x,col="GREEN",pch=8)
lev<-levels(g); nn<-length(g)
rb<-rainbow(length(lev))
for (k in lev) {
        ind<-which(k==g)
        n<-length(ind)
        if (n==1) dd<-mm<-matrix(0,1,1)
        else {
                dd<-d[ind,ind]
                mm<-mst(dd) }
        for (i in 1:n) {
                jnd<-which(1==as.vector(mm[i,]))
```

```
51                  sapply(jnd,function(r) lines(rbind(x[ind[i],],x[ind[r],]),col=rb[
                        which(lev==k)]))
                    }
                }
54  }


57  eudist<-function(x,pow,eps){
    c<-crossprod(t(x)); s<-diag(c)
    dd<-outer(s,s,"+")-2*c
60  if (pow==2) return(dd)
            else return(sqrt(dd+eps))
    }
63
    bmat<-function(c,d,pow){
    if (pow==2) b<-c else b<-c/d
66  r<-diag(rowSums(b))
    return(r-b)
    }
```

REFERENCES

P. Bekker and J. De Leeuw. Relation between variants of nonlinear principal component analysis. In J.L.A. Van Rijckevorsel and J. De Leeuw, editors, *Component and Correspondence Analysis*. Wiley, Chichester, England, 1988.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis*, Berlin, 1994. Springer Verlag.

J. De Leeuw. Homogeneity Analysis of Pavings. URL `http://jackman.stanford.edu/ideal/MeasurementConference/abstracts/homPeig.pdf`. August 2003.

J. Deleeuw and G. Michailides. Weber correspondence analysis: The one-dimensional case. Preprint 343, UCLA Department of Statistics, 2003. URL `http://preprints.stat.ucla.edu/343/twopoints.pdf`.

R.L. Graham and P. Hell. On the History of the Minimum Spanning Tree Problem. *Annals of the History of Computing*, 7:43–57, 1985.

W.J. Heiser. Correspondence analysis with least absolute residuals. *Computational Statistica and Data Analysis*, 5:357–356, 1987.

W.J. Heiser. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J. Krzanowski, editor, *Recent Advanmtages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 1995.

K. Lange, D.R. Hunter, and I. Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9:1–20, 2000.

G. Michailides and J. Deleeuw. Homogeneity analysis using absolute deviations. Preprint 346, UCLA Department of Statistics, 2003. URL `http://preprints.stat.ucla.edu/346/paper3.pdf`.

J.L.A. Van Rijckevorsel. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. PhD thesis, University of Leiden, The Netherlands, 1987. Also published in 1987 by DSWO-Press, Leiden, The Netherlands.

B.F. Schriever. *Order Dependence*. PhD thesis, University of Amsterdam, The Netherlands, 1985. Also published in 1985 by CWI, Amsterdam, The Netherlands.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

E-mail address, Jan de Leeuw: `deleeuw@stat.ucla.edu`

URL, Jan de Leeuw: `http://gifi.stat.ucla.edu`