Multidimensional Voronoi Scaling of Categorical Data

Jan de Leeuw - University of California Los Angeles

Started April 13 2024, Version of April 15, 2024

Abstract

smacofVO

Contents

Note: This is a working paper which will be expanded/updated frequently. All suggestions for improvement are welcome.

1 Introduction

1.1 Simultaneous Nonmetric Unfolding

In non-metric unfolding

$$\sigma(X,Y)=\sum_{i=1}^n\min_{\delta_i\in\Delta_i}\sum_{j=1}^m w_{ij}(\delta_{ij}-d(x_i,y_j))^2$$

If we have data where the same individuals give preference judgments over multiple domains, or over the same domain at multiple occasions, or over the same domain with different experimental conditions, then we can use the loss function

$$\sigma(X,Y_1,\cdots,Y_s)=\sum_{r=1}^s\sum_{i=1}^n\min_{\delta_i^r\in\Delta_i^r}\sum_{j=1}^{m_r} w_{ij}^r(\delta_{ij}^r-d(x_i,y_j^r))^2$$

Note that for each occasion there are different transformations Δ_r and different matrices of column scores Y_j, but there is only a single matrix of row scores X.

z## Homogeneity Analysis

Homogeneity Analysis, also known as Multiple Correspondence Analysis, is a basic multivariate data analysis technique. The Gifi System (Gifi (1990), Michailidis and De Leeuw (1998), De Leeuw and Mair (2009)) presents the familiar linear multivariate analysis techniques (regression, analysis of variance, canonical analysis, discriminant analysis, principal component analysis) as special cases of homogeneity analysis.

We give a somewhat non-standard introduction to homogeneity analysis here, to highlight the similarities with unfolding and the voronoi technique we will present further on in this paper.

The data are a number of indicator matrices $G_1,\cdots,G_s$. Indicator matrices are binary matrices, with rows that add up to one. They are used to code categorical variables. Rows corresponds with objects (or individuals), column with the categories (or levels) of a variable. An element g_{ij} is one in row if object i is in category j, and all other elements in row i are zero.

Homogeneity analysis makes a joint maps in p dimensions of individuals and categories (both represented as points) in such a way that category points are close to the points for the individuals in the category. And, vice versa, individuals are close to the category points that they score in. If there is only one variable then it is trivial to make such a homogeneous map. We just make sure the individual points coincide with their category points. But there are $s>1$ indicator matrices, corresponding with s categorical variables, and the solution is a compromise trying to achieve homogeneity as well as possible for all variables simultaneously.

$$\sigma(X,Y_1,\cdots,Y_s)=\sum_{r=1}^s\sum_{i=1}^n w_i^r d^2(x_i,y_i^r)=\sum_{r=1}^s \text{tr}\ (X-G_rY_r)'W_r(X-G_rY_r)$$

$$w_i^r:=\sum_{j=1}^{m^r} g_{ij}^r w_{ij}^r$$

$$y_i^r := \sum_{j=1}^{m^r} g_{ij}^r y_j^r$$

Constraint is that δ_{ij}^r is zero if i is in category j of variable r. There are no constraints on the other delta's in row i. of variable r. Thus we want an object to coincide with all s categories it is in.

star plot

$$Y_r = (G_r' W_r G_r)^{-1} G_r' W_r X$$

$$\min_Y \sigma(X, Y_1, \cdots, Y_s) = \text{tr } X' \left\{ \sum_{r=1}^{s} \{W_r - W_r G_r (G_r' W_r G_r)^{-1} G_r' W_r\} \right\} X$$

$X' W_\star X = I$

2 Voronoi Loss Function

2.1 The Unconstrained Case

Now consider the closely related problem in which we do not require, as in homogeneity analysis, that

$$\delta_i^r := \sum_{j=1}^{m_r} g_{ij}^r \delta_{ij}^r = 0$$

but we merely require that δ_i^r is less than or equal to all other δ_{ij}^r in row i. Formally

$$\delta_i^r \leq \delta_{ij}^r \qquad \forall j = 1, \cdots, m_r$$

or

$$g_{il}^j = 1 \Rightarrow \delta_{il}^j \leq \delta_{iv}^j \qquad \forall v \neq \ell$$

$$\sigma(X, Y_1, \cdots, Y_m) = \sum_{j=1}^{m} \left\{ \sum_{i=1}^{n} \sum_{\ell=1}^{k_j} w_{il}^j (\delta_{i\ell}^j - d(x_i, y_\ell^j))^2 \right\}$$

$$\sum_{j=1}^{m} \left\{ \sum_{i=1}^{n} \sum_{\ell=1}^{k_j} w_{il}^j d^2(x_i, y_\ell^j) \right\} = 1$$

$$Z_j = \begin{matrix} {\scriptstyle n} \\ {\scriptstyle k_j} \end{matrix} \begin{matrix} {\scriptstyle p} \\ \left[\begin{array}{c} X \\ Y_j \end{array} \right] \end{matrix}$$

$$d^2(x_i, y_\ell^j) = \text{tr } Z_j' A_{i\ell}^j Z_j$$

$$\sum_{i=1}^{n} \sum_{\ell=1}^{k_j} w_{il}^j d^2(x_i, y_\ell^j) = \text{tr } Z_j' V_j Z_j$$

$$V_j = \sum_{i=1}^{n} \sum_{\ell=1}^{k_j} w_{il}^j A_{i\ell}^j$$

$$V_j = \begin{matrix} {\scriptstyle n} \\ {\scriptstyle k_j} \end{matrix} \begin{matrix} {\scriptstyle n} \quad {\scriptstyle k_j} \\ \left[\begin{array}{cc} \ddots & \square \\ \square & \ddots \end{array} \right] \end{matrix}$$

$$\sum_{i=1}^{n} \sum_{\ell=1}^{k_j} w_{il}^j \delta_{i\ell}^j d(x_i, y_\ell^j) = \text{tr } Z_j' B_j(Z_j) Z_j \geq \text{tr } Z_j' B_j(\tilde{Z}_j) \tilde{Z}_j = \text{tr } Z_j' V_j G(\tilde{Z}_j)$$

$$\sigma(Z_1, \cdots, Z_m) \leq \sum_{j=1}^{m} \left\{ \eta^2(\Delta_j) + \text{tr } (Z_j - G(\tilde{Z}_j))' V_j (Z_j - G(\tilde{Z}_j)) - \text{tr } G(\tilde{Z}_j)' V_j G(\tilde{Z}_j) \right\}$$

$$\sum_{j=1}^{m} Z_j' V_j Z_j = X' \{ \sum_{j=1}^{m} H_j' V_j H_j \} X.$$

$$\text{tr } (X - \tilde{X}_j)' V_{11}^j (X - \tilde{X}_j) + 2\text{tr } (X - \tilde{X}_j)' V_{12}^j (Y_j - \tilde{Y}_j) + \text{tr } (Y_j - \tilde{Y}_j)' V_{22}^j (Y_j - \tilde{Y}_j)$$

The Unweighted Case

If unweighted then $V_{11}^j = k_j I$ and $V_{22}^j = nI$. V_{12}^j is $n \times k_j$ with all elements equal to -1. The rank of V_j is $n + k_j - 1$, and the only vectors in the null-space are proportional to the vector with all elements equal to one ().

$$V_j^+ = (V_j + ee)^{-1} - \frac{1}{n + k_j} ee'$$

But

$$V_j + ee' = (V_{11}^j + ee') \oplus (V_{22}^j + ee') = (k_j I + ee') \oplus (nI + ee')$$

$$(V_j + ee')^{-1} = (k_j I + ee')^{-1} \oplus (nI + ee')^{-1}$$

Lemma: If A is a symmetric matrix of order n of the form $\alpha I + ee'$ then $(\alpha I + ee')^{-1} =$

Proof:

Heiser and De Leeuw (1979)

Sherman-Morrison

$$(nI + ee')^{-1} = \frac{1}{n} \left\{ I - \frac{1}{k_j + n} ee' \right\}$$

$$(k_j I + ee')^{-1} = \frac{1}{k_j} \left\{ I - \frac{1}{k_j + n} ee' \right\}$$

Schur complement

$$V_{11} - V_{12} V_{22}^{-1} V_{21} = kI - e_n e_k' \frac{1}{n} e_k e_n' = k\{I - \frac{1}{n} e_n e_n'\} = kJ_n$$

2.2 The Constrained Case

$$Z_j = \begin{matrix} n \\ k_j \end{matrix} \begin{bmatrix} \overset{p}{I} \\ D_j^{-1} G_j' \end{bmatrix} X = H_j X$$

References

De Leeuw, J., and P. Mair. 2009. “Homogeneity Analysis in R: the Package homals.” *Journal of Statistical Software* 31 (4): 1–21. https://www.jstatsoft.org/v31/i04/.

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

Heiser, W. J., and J. De Leeuw. 1979. “Metric Multidimensional Unfolding.” *Methoden En Data Nieuwsbrief SWS/VVS* 4: 26–50.

Michailidis, G., and J. De Leeuw. 1998. “The Gifi System for Descriptive Multivariate Analysis.” *Statistical Science* 13: 307–36.