Voronoi Analysis of Categorical Data

Jan de Leeuw

May 5, 2026

TBD

Table of contents

Note: This is a working manuscript which will be expanded/updated frequently. All suggestions for improvement are welcome. All Rmd, tex, html, pdf, R, and C files are in the public domain. Attribution will be appreciated, but is not required. The files can be found at https://github.com/deleeuw/voronoi

1 Introduction

1.1 Homogeneity Analysis

Homogeneity analysis, a.k.a. *Multiple Correspondence Analysis*, can be introduced in many different ways. We refer to Guttman (1941), Tenenhaus and Young (1985), Bekker and De Leeuw (1988), Gifi (1990), Greenacre and Blasius (2006), Le Roux and Rouanet (2010), De Leeuw (2023) for a bouquet of discussions and interpretations of the equations and algorithms defining the technique.

For our purposes a geometric interpretation using *star plots* seems the most natural one. We have n objects measured by m categorical variables. Variable j has k_j categories, which are the different values the variable can assumes for the n objects. Suppose the objects are mapped into n points of Euclidean space $\mathbb{R}^p$. For each variable we can construct an additional k_j points in the same space by computing for each category the centroids of all object-points in that category. Thus each variable defines its own plot. We then draw lines from the n object points to "their" category point and we obtain k_j "stars". Homogeneity analysis maps the objects into space in such a way that the stars are as small as possible, where the size of the star is measured as the sum of squares of the distances between each category points and "its" object points (summed over categories and over variables). In order to avoid the trivial solution in which all objects are mapped into the same point homogeneity requires that the *configuration* of object points (an $n \times p$ matrix) is column-centered and orthonormal.

There are a huge number of variations possible on the same theme. Instead of the centroids we could use the Weber points to define our stars. Instead of Euclidean distance we could use any other distance on $\mathbb{R}^p$. Instead of the stars we could use the convex hulls or the minimum spanning trees, or any other way to visualize and quantify the compactness of the point clouds of each category. Any one of these choices defines a technique that could be called homogeneity analysis, but in this paper we mean the classical choice that uses stars, squared Euclidean distance, and orthonormal object points.

The key mathematical object in homogeneity analysis is the *indicator matrix*, which is a classical way to code categorical data. For a variable j with k_j categories it is an $n \times k_j$ binary matrix G_j in which each row i contains a single element equal to one, indicating into which category the object i maps for variable j. The cross product $D_j := G_j'G_j$ is diagonal, with the *marginal frequencies* of the categories on the diagonal. For two different variables the crossproduct $C_{j\ell} = G_j'G_\ell$ is the $k_j \times k_\ell$ *cross table* or *contingency table* of the two variables.

We can also define homogeneity analysis in the traditional non-metric multidimensional scaling

framework, using the loss function (Kruskal (1964a), Kruskal (1964b)). Define

$$\sigma(X, Y_1, \cdots, Y_m, \Delta_1, \cdots, \Delta_m)) := \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{\ell=1}^{k_j}(\delta_{i\ell}^j - d(x_i, y_\ell^j))^2. \tag{1}$$

Homogeneity analysis minimizes this loss over all configurations X with $X'X = I$, over the Y_j, and over all Δ_j that satisfy $\delta_{i\ell}^j = 0$ whenever $g_{i\ell}^j = 1$. The other elements of Δ are only constrained to be non-negative.

If

$$\sigma_\star(X, Y_1, \cdots, Y_m) := \min_{\Delta_1, \cdots, \Delta_m} \sigma(X, Y_1, \cdots, Y_m, \Delta_1, \cdots, \Delta_m), \tag{2}$$

then

$$\sigma_\star(X, Y_1, \cdots, Y_m) = \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{\ell=1}^{k_j} g_{i\ell}^j d(x_i, y_\ell^j)^2. \tag{3}$$

Now define

$$\sigma_{\star\star}(X) := \min_{Y_1, \cdots, Y_m} \sigma_\star(X; Y_1, \cdots, Y_m). \tag{4}$$

The minimum over the Y_j is attained at the centroids $\hat{Y}_j := D_j^{-1}G_j'X$, and if we substitute these optimal $\hat{Y}_j$ in (3) we see that $\sigma_{\star\star}(X)$ is indeed equal to our measure of the total size of all stars. Minimizing $\sigma_{\star\star}(X)$ over $X'X = I$ is homogeneity analysis, and computing the solution turns out to be a straightforward eigenvalue-eigenvector problem. The details are in the publications we referenced at the beginning of this section.

1.2 Voronoi Analysis

Although using loss function (1) makes homogeneity analysis a form of (metric) multidimensional scaling there are some important differences. There are no local minima in homogeneity analysis, because it is a symmetric eigenvalue problem. Also, and most importantly, we do not really expect the minimum loss function value to be zero, or even close to zero. Zero loss would imply all object points coincide with all category points they map into, which is clearly unrealistic. It can only be realized if all object and category points map into a single point, which is prohibited by the constraint $X'X = I$. Thus we get a perfectly determinate and stable solution, but at the cost of imposing unrealistic constraints and the rather arbitrary normalization $X'X = I$.

This makes it interesting to generalize loss function (1) to

$$\sigma(X, Y_1, \cdots, Y_m, \Delta_1, ..., \Delta_m) := \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{\ell=1}^{k_j}(\delta_{i\ell}^j - d(x_i, y_\ell^j))^2, \tag{5}$$

with the constraints that if $g_{i\ell}^j = 1$ then merely $\delta_{i\ell}^j \leq \delta_{i\nu}^j$ for all $\nu = 1, \cdots, k_j$. This is significantly weaker than the homogeneity analysis constraints, which require the $\delta_{i\ell}^j$ to be zero. If object i is in category ℓ of variable j we do not require the object point to coincide with the category point, but we merely require that the object point is at least as close to category point ℓ as to any of the other $k_j - 1$ category points of that variable.

As a consequence, the geometry corresponding with loss function (5) is also different. As in homogeneity analysis there is a plot for each variable j with both n object points and k_j category points. The k_j category points define k_j *voronoi regions* that partition the space. Voronoi region $\mathcal{V}_\ell^j$ is the region of space where the points are at least as close to category point y_ℓ^j as to any of the other category points. And loss (4) is zero (for a variable) if all object points are in the "correct" Voronoi region, i.e. in the Voronoi region of the category point corresponding with the category they are in.

One (very inefficient, but conceptually easy) way of constructing Voronoi regions is to draw the perpendicular bisectors of the (hypothetical) lines connecting all pairs of category points of a variable. The convex polyhedral regions thus formed, some bounded and some unbounded, define the Voronoi regions. The literature on properties and computation on Voronoi regions is large, and we refer to the textbooks of Okabe et al. (2000) and Aurenhammer et al. (2013).

Since the category point itself is obviously in its own Voronoi region and since Voronoi regions are convex it follows that if we make stars by connecting category points with objects points as in homogeneity analysis then we actually require each star to be in its Voronoi region.

Voronoi analysis of categorical data, as defined here, minimizes loss function (5) over X, over the Y_j and over the Δ_j satisfying the ordinal constraints. It is consequently another form of non-metric (ordinal) multidimensional scaling. To be more precise, we have m linked non-metric multidimensional unfolding problems, which are linked because they share the same X. The data are binary, and we use Kruskal's primary approach to ties (De Leeuw (1977b)).

The constraint $X'X = I$ from homogeneity analysis is no longer needed, but some form of normalization is still necessary. We still need to exclude the trivial solution in which all points and regions collapse into a single point and all Δ_j are zero. But, more urgently, we have to deal with the trivial solutions familiar from multidimensional unfolding (Heiser (1981), De Leeuw (1983), Busing (2010)). For example, we can collapse all category points into a single point and put all object points on a sphere with the category point at its center. Then all distances between category and object points are the same and we can choose all dissimilarities equal to that distance value and obtain stress zero. It does not help to require $X'X = I$, because we can easily construct orthogonal configurations with all points on the sphere. Stronger normalization constraints are required.

In this paper we use the constraint

$$\sum_{\ell=1}^{k_j}(\delta_{i\ell}^j - \bar{\delta}_{ij})^2 = c_j, \tag{6}$$

with $\bar{\delta}_{ij}$ the mean of the $\delta_{i\ell}^j$ and with c_j constants. For instance we could choose $c_j = 1$ or $c_j = k_j$ for all j. In addition, of course, the $\delta_{i\ell}^j$ must satisfy the ordinal constraints. Constraint (6) explicitly excludes the possibility that the $\delta_{i\ell}^j$ for given i and j are equal for all ℓ.

2 Algorithm

We use an *Alternating Least Squares* of *ALS* algorithm, similar to many other non-metric muldimensional scaling methods, to minimize loss (5). For context, see Borg and Groenen (2005). We alternate minimization over Δ for given X and Y with minimization over X and Y for given Δ. Using superscript (μ) for the iteration index we can write

$$\Delta^{(\mu)} = \underset{\Delta}{\operatorname{argmin}}\,\sigma(X^{(\mu)}, Y^{(\mu)}, \Delta), \tag{7}$$

$$(X^{(\mu+1)}, Y^{(\mu+1)}) = \underset{X,Y}{\operatorname{argmin}}\,\sigma(X, Y, \Delta^{(\mu)}). \tag{8}$$

The first substep of iteration ν is a form of monotone regression for a very simple partial order, in which one of k_j elements must be smaller than the other elements. The second substep is a metric multidimensional scaling step. It cannot be carried out completely, because it would need an infinite iterative process by itself. Thus the second substep consist of a finite number of *inner iterations* within the second substep of an *outer iteration*. For the inner iterations we use standard smacof majorization steps (De Leeuw (1977a)).

2.1 Monotone Regression

The nm monotone regressions are applied to each row i of each variable j. Forgetting about these indices for a moment we need to minimize

$$\sigma(\delta) := \sum_{\ell=1}^{k}(\delta_\ell - d_\ell)^2$$

$$\sum_{\ell=1}^{k}(\delta_\ell - \bar{\delta})^2 = c$$

3 Majorization

$$A_{ij} := \begin{bmatrix} e_i e_i' & -e_i e_j' \\ -e_j e_i' & e_j e_j' \end{bmatrix}$$

$$Z := \begin{bmatrix} X \\ Y \end{bmatrix}$$

$$d_{ij}(X, Y) = \text{tr } Z' A_{ij} Z$$

$$V := \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} = \begin{bmatrix} mI & -ee' \\ -ee' & nI \end{bmatrix}$$

If we define

$$V_1 := (n+m)^{-1} \begin{bmatrix} \frac{m}{n} E & -E \\ -E & \frac{n}{m} E \end{bmatrix}, \tag{9}$$

$$V_2 := \begin{bmatrix} J_n & 0 \\ 0 & 0 \end{bmatrix}, \tag{10}$$

$$V_3 := \begin{bmatrix} 0 & 0 \\ 0 & J_m \end{bmatrix}, \tag{11}$$

then

$$V = (n+m)V_1 + mV_2 + nV_3$$

Matrices V_1, V_2, and V_3 are symmetric projectors on orthogonal subspaces. It follows that V^+, the Moore-Penrose inverse3 of V, is

$$V^+ = \frac{1}{n+m} V_1 + \frac{1}{m} V_2 + \frac{1}{n} V_3$$

$$B(X, Y) := \begin{bmatrix} B_{11}(X, Y) & B_{12}(X, Y) \\ B_{21}(X, Y) & B_{22}(X, Y) \end{bmatrix}$$

Now $V_1 B(X, Y) = 0$ and thus

$$V^+ B(X, Y) = \frac{1}{m} V_2 B(X, Y) + \frac{1}{n} V_3 B(X, Y)$$

4 The Cone

Consider the cone of all vectors x in $\mathbb{R}^n$ for which $x_i \leq x_1$ for all $i \neq 1$ and $e'x = 0$. The extreme rays of the cone will be the $n-1$ vectors of the form $\alpha e_i + \beta(e - e_i)$ with $i \neq 1$. Thus element i is equal to α and all other elements (including x_1) are equal to beta. Thus we must have $\alpha > \beta$. The sum of the elements is $\alpha + (n-1)\beta$, which must be zero. Thus $\alpha = -(n-1)\beta$, which implies $\beta < 0$ and $\alpha > 0$. The extreme rays are positive multiples of the vectors $e_i - n^{-1}e$ with $i \neq 1$ and the cone is the set of all weighted sums with non-negative weights of these $n-1$ vectors. Here is a small example.

```
a <- -cbind(1, -diag(4))
b <- (diag(5)-(1 / 5))[, -1]
print(a)
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]   -1    1    0    0    0
[2,]   -1    0    1    0    0
[3,]   -1    0    0    1    0
[4,]   -1    0    0    0    1
```

```
print(b)
```

```
     [,1] [,2] [,3] [,4]
[1,] -0.2 -0.2 -0.2 -0.2
[2,]  0.8 -0.2 -0.2 -0.2
[3,] -0.2  0.8 -0.2 -0.2
[4,] -0.2 -0.2  0.8 -0.2
[5,] -0.2 -0.2 -0.2  0.8
```

```
print(a %*% b)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

For later reference the sum of squares of the vectors $e_i - n^{-1}e$ is $(n-1)/n$, so the vectors $\sqrt{n/(n-1)}(e_i - n^{-1}e)$ are the vectors of length 1 on the extreme rays. The average of the extreme rays is

5 Normalized Monotone Regression

Suppose y is a vector with n elements. The problem we study is to minimixe the sum of squares $\sigma(x) := \|x - y\|^2$ over $x \in K$ and $x'Jx = 1$. Here

- K is the cone of vectors with $x_1 \leq x_2 \leq \cdots \leq x_n$,
- J is the centering matrix $J := I - n^{-1}ee'$, where
- e is the vector with all n elements equal to one.

Change variables to $\tilde{x} = Jx$. Then $x'Jx = \tilde{x}'\tilde{x}$ and $x \in K$ if and only if $\tilde{x} \in K$. Also $\tilde{x} = Jx$ if and only if $x = \tilde{x} + \beta e$. So the transformed problem is to minimize $\|\tilde{x} + \beta e - y\|^2$ over β and $\tilde{x} \in K$, with $e'\tilde{x} = 0$ and $\tilde{x}'\tilde{x} = 1$. The minimizer for β is $\overline{y} := n^{-1}e'y$, and thus the problem becomes minimizing $\|\tilde{x} - Jy\|^2$ over $\tilde{x} \in K$, $\tilde{x}'\tilde{x} = 1$, and $e'\tilde{x} = 0$.

Forget about the constraint $e'\tilde{x} = 0$. If M is the monotone regression operator it follows that the solution for $\tilde{x}$ is $M(Jy)/\|M(Jy)\|$, which automatically satisfies $e'\tilde{x} = 0$. Thus the solution of our original problem is

$$\hat{x} = \frac{M(y - \overline{y})}{\|M(y - \overline{y})\|} + \overline{y}.$$

$M(y - \overline{y}) = M(y) - \overline{y}$ and thus $\|M(y - \overline{y})\|^2 = \|M(y)\|^2 + n\overline{y}^2 - 2\overline{y}e'M(y)$ nor $e'M(y) = e'y = n\overline{y}$. Thus $\|M(y - \overline{y})\|^2 = \|M(y)\|^2 - n\overline{M(y)}^2$

$$\hat{x} = \frac{M(y) - \overline{y}e}{\|M(y - \overline{y})\|} + \overline{y}e = \frac{1}{\|M(y - \overline{y})\|}M(y) + \left\{1 - \frac{1}{\|M(y - \overline{y})\|}\right\}\overline{y}e$$

References

Aurenhammer, Franz, Rolf Klein, and Der-Tsai Lee. 2013. *Voronoi Diagrams and Delaunay Triangulations*. World Scientific Publishing Co.

Bekker, Paul, and Jan De Leeuw. 1988. "Relation Between Variants of Nonlinear Principal Component Analysis." Chap. 1 in *Component and Correspondence Analysis*, edited by Jan L. A. Van Rijckevorsel and Jan De Leeuw. Wiley Series in Probability and Mathematcal Statistics. Wiley.

Borg, Ingwer, and Patrick J. F. Groenen. 2005. *Modern Multidimensional Scaling*. Second Edition. Springer.

Busing, Frank M. T. A. 2010. "Advances in Multidimensional Unfolding." PhD thesis, Leiden University. https://scholarlypublications.universiteitleiden.nl/access/item%3A2837800/view.

De Leeuw, Jan. 1977a. "Applications of Convex Analysis to Multidimensional Scaling." In *Recent Developments in Statistics*, edited by J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem. North Holland Publishing Company.

De Leeuw, Jan. 1977b. "Correctness of Kruskal's Algorithms for Monotone Regression with Ties." *Psychometrika* 42: 141–44.

De Leeuw, Jan. 1983. *On Degenerate Multidimensional Unfolding Solutions*. Research Report. Department of Data Theory FSW/RUL. https://jansweb.netlify.app/publication/deleeuw-r-06-a/deleeuw-r-06-a.pdf.

De Leeuw, Jan. 2023. "Deconstructing Multiple Correspondence Analysis." Chap. 22 in *Analysis of Categorical Data from Historical Perspectives. Essays in Honour of Shizuhiko Nishisato.*, edited by Eric J. Beh, Rosaria Lombardo, and Jose G. Clavel. Springer.

Gifi, Albert. 1990. *Nonlinear Multivariate Analysis*. Wiley.

Greenacre, Michael J., and Jorg Blasius. 2006. *Multiple Correspondence Analysis and Related Methods*. Chapman; Hall.

Guttman, Louis. 1941. "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction." In *The Prediction of Personal Adjustment*, edited by Paul Horst. Social Science Research Council.

Heiser, Willem J. 1981. “Unfolding Analysis of Proximity Data.” PhD thesis, Leiden University.

Kruskal, Joseph B. 1964a. “Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis.” *Psychometrika* 29: 1–27.

Kruskal, Joseph B. 1964b. “Nonmetric Multidimensional Scaling: a Numerical Method.” *Psychometrika* 29: 115–29.

Le Roux, Brigitte, and Henry Rouanet. 2010. *Multiple Correspondence Analysis*. Sage.

Okabe, Atsuyuki, Barry Boots, Kokichi Sugihara, and Chiu Sung-Nok. 2000. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Second Edition. Wiley.

Tenenhaus, Michel, and Forest W. Young. 1985. “An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data.” *Psychometrika* 50: 91–119.