UCLA

Department of Statistics Papers

Title

Distance-Based Transformations of Biplots

Permalink

<https://escholarship.org/uc/item/0bd8s7fk>

Author Jan de Leeuw

Publication Date 2011-10-25

DISTANCE-BASED TRANSFORMATIONS OF BIPLOTS

JAN DE LEEUW

1. INTRODUCTION

In principal component analysis and related techniques we approximate (in the least squares sense) an $n \times m$ matrix *F* by an $n \times m$ matrix *G* which satisfies rank $(G) \leq p$, where $p < \min(n, m)$. Or, equivalenty, we want to find an $n \times p$ matrix *X* and an $m \times p$ matrix *Y* such that $G = XY'$ approximates *F* as closely as possible. The rows of *X* and *Y* are then often used in graphical displays. In particular, *biplots* [\[Gower and Hand, 1996\]](#page-5-0) represent *X* and *Y* jointly as $n + m$ points in Euclidean p space.

If formulated in this way, there is an important form of indeterminacy in this approximation problem. If R of order p is nonsinsular, then we can define $\tilde{X} = XR$ and $\tilde{Y} = YR^{-T}$ and we have $\tilde{X}\tilde{Y}' = XY'$, where A^{-T} is the transpose of the inverse (or the inverse of the transpose). Thus \tilde{X} and \tilde{Y} give exactly the same approximation, but plotting them may give quite different results, depending on *R*. To give a simple example, we can choose *R* scalar, and make \tilde{X} arbitrarily small and \tilde{Y} arbitrarily big. In particular for biplots, which are often interpreted in terms of distances between the points, the indeterminacy is a nuisance and can lead to unattractive representations.

In this note we choose R in such a way that the distances, more specifically the squared Euclidean distances, between selected rows of \tilde{X} and \tilde{Y} are small. This takes care of both the relative scaling of the two clouds of points, as well as rotating them to some form of conformance.

Date: April 14, 2006.

²⁰⁰⁰ *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis.

2 JAN DE LEEUW

2. PROBLEM FORMULATION

The squared distance between rows *i* and *j* of the $n + m$ matrix

$$
Z = \begin{bmatrix} XR\\YR^{-T}\end{bmatrix}
$$

can be written as

$$
d_{ij}^2(R) = (e_i - e_j)'C(e_i - e_j) = \text{tr } CA_{ij}.
$$

Here the e_i are unit vectors (columns of the identity matrix) and we define

$$
C = \begin{bmatrix} XSX' & XY' \\ YX' & YS^{-1}Y' \end{bmatrix},
$$

as well as $S = RR'$ and $A_{ij} = (e_i - e_j)(e_i - e_j)'$

Thus summing over a selected subset I of squared distances leads to a loss function of the form

$$
\lambda(S) = \sum_{(i,j)\in\mathcal{I}} d_{ij}^2(S) = \mathbf{tr}\, S X' A_{11} X + \mathbf{tr}\, S^{-1} Y' A_{22} Y
$$

where A_{11} and A_{22} are the two principal submatrices of

$$
A = \sum_{(i,j)\in\mathcal{I}} A_{ij}.
$$

If we minimize the sum of squares of all *nm* distances between the *n* points in *X* and the *m* points in *Y*, for example, we find $A_{11} = mI$ and $A_{22} = nI$. If $n = m$ and we want to minimize the sum of the *n* squared distances between the corresponding points x_i and y_i then $A_{11} = A_{22} = I$.

3. PROBLEM SOLUTION

Let us minimize $\lambda(S) = \text{tr } SP + \text{tr } S^{-1}Q$, where both *P* and *Q* are positive definite. If *P* and/or *Q* are singular, the more general results of [De](#page-5-1) [Leeuw](#page-5-1) [\[1982\]](#page-5-1) must be used, but in most applications we have in mind nonsingularity is guaranteed.

The stationary equations for the problem of minimizing $\lambda(S)$ are

(1)
$$
P = S^{-1}QS^{-1},
$$

which we have to solve for a positive definite *S* . We can use the symmetric square root to rewrite Equation [\(1\)](#page-2-0) as

(2)
$$
I = P^{-\frac{1}{2}} S^{-1} P^{-\frac{1}{2}} \left[P^{\frac{1}{2}} Q P^{\frac{1}{2}} \right] P^{-\frac{1}{2}} S^{-1} P^{-\frac{1}{2}},
$$

from which

(3)
$$
P^{-\frac{1}{2}}S^{-1}P^{-\frac{1}{2}}=\left[P^{\frac{1}{2}}QP^{\frac{1}{2}}\right]^{-\frac{1}{2}},
$$

and thus

(4)
$$
S^{-1} = P^{\frac{1}{2}} \left[P^{\frac{1}{2}} Q P^{\frac{1}{2}} \right]^{-\frac{1}{2}} P^{\frac{1}{2}},
$$

and

(5)
$$
S = P^{-\frac{1}{2}} \left[P^{\frac{1}{2}} Q P^{\frac{1}{2}} \right]^{\frac{1}{2}} P^{-\frac{1}{2}}.
$$

If we want to minimize the sum of squares of all distances between the points in *X* and those in *Y* we have seen that $A_{11} = mI$ and $A_{22} = nI$. In many forms of principal component analysis *X* is chosen such that $X'X = I$, and thus $P = mI$. In that case, from [\(5\)](#page-3-0),

$$
S = \sqrt{\frac{n}{m}} (Y'Y)^{\frac{1}{2}}.
$$

If $Y = L\Lambda L'$ is an eigen-decomposition of *Y*, we can choose

$$
R = \left[\frac{n}{m}\right]^{\frac{1}{4}} L\Lambda^{\frac{1}{4}},
$$

$$
R^{-T} = \left[\frac{m}{n}\right]^{\frac{1}{4}} L\Lambda^{-\frac{1}{4}}.
$$

4. EXAMPLE

To illustrate the problem, consider the following output from the scalAssoc() program [\[De Leeuw, 2006\]](#page-5-2). These are 20 votes of 100 US senators. Each vote is presented by a plus ("aye") point and a minus ("nay") point, and the technique jointly scales senators and votes in such a way that senators are closest to the vote points they endorse. Or, equivalently, senators voting "aye" must be separated by a straight line from senators voting "nay". In Figure 1 all senators are clumped around the origin, and this makes it impossible to read and interpret the plot.

4 JAN DE LEEUW

Column Objects (Red) Row Objects (Green) senate

$V19_+$ 50 50
0
0 V15_− $V14_+$ dimension 2 dimension 2 V11__ $V₅$ V20_− VZ_V V3_− V4_+ V4_− V12_+ V8_+ V8_− V9_+V9_− V10_+ V10_− N42D−Y V49⊟∀ V18_V \circ V17∨1 ^{V18}V16. – William Vil6_{V18_+}V1v+7_V₁ M 103 $+$ V1_− Hutchinson BrownbackRockefeller MurkowskiMcConnellThompson Lieberman Thurmond HutchisonBingaman Fitzgerald Santorum CarnahanStabenow WellstoneVoinovichCampbellSarbanesDomeniciSessions FeinsteinGrassley SchumerLandrieuKennedyEdwardsCochranFeingoldCantwellJohnsonBunning ThomasNelson1TorricelliRoberts StevensDeWineaschl GrahamJeffordsHollingsMikulskiBennettSpecterMcCainCorzineGramm ClelandBaucusWarnerNicklesDorgan ConradhafeeBreauxSmith1LincolnDayton EnsignShelbyNelsonWydenCarperClintonSnoweMurrayCollinsInouyeHelmsDurbinHak Gregg InhofeCrapoAkakaBurns HatchLeahy HagelAllardLugarSmithBidenoxerCraig MillerBond Dd ReedKerryLevinAllenBayhByrd Reid FristEnziKohlLttKylV1_+ V2_+ ™130_∸ V6_+ 12 V3_+ V7_+ $V20'$ 수 V5_− V11_+ V14_− V15_+ −50 V19_− −60 −40 −20 0 20 40 60

Now let us apply the scaling outlines in this paper. Figure 2 gives the results,

dimension 1

which are clearly much more satisfactory.

Column Objects (Red) Row Objects (Green) senate

REFERENCES

- J. De Leeuw. Majorization Methods for Distance Association Models. Technical report, UCLA Department of Statistics, 2006.
- J. De Leeuw. Generalized Eigenvalue Problems with Positive Semidefinite Matrices. *Psychometrika*, 47:87–94, 1982.
- J.C. Gower and D.J. Hand. *Biplots*. Number 54 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1996.

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