CHAPTER 28

PRINCIPAL COMPONENT ANALYSIS OF SENATE VOTING PATTERNS

Jan de Leeuw

There are various techniques available for the principal component analysis (PCA) of binary matrices. We illustrate some of them in this chapter by analyzing votes on 20 issues in the 2001 US Senate, selected by Americans for Democratic Action (Americans for Democratic Action, 2002). It must be emphasized that the techniques we discuss are general, because they apply to many different types of binary matrices. They can be used equally effectively, for example, to analyze data from choice experiments or from tests and exams.

The PCA techniques we use are chosen from two different classes (De Leeuw, 2006). First, there is *homogeneity analysis* (Gifi, 1990), also known as *multiple correspondence analysis* (Greenacre & Blasius, 2006). In this technique we make a joint two-dimensional plot of both senators and issues. In order to do this, we first have to choose a *dimensionality.* In this chapter we make all our plots in two-dimensional scape. Each issue is represented by two points, an "aye" point and a "nay" point. In homogeneity analysis the objects (in this example the senators) are stan-

Real Data Analysis, pp. 405–409

dardized, in the sense that coordinates on both dimensions are centered, have unit sum of squares, and are uncorrelated with one another. Moreover the "aye" point for each issue is the centroid (average) of all senator points voting "aye" on the issue, and the same is true for all "nay" points.

If we have a representation in the plane, then for each issue we can make a *star plot* (Michailidis & De Leeuw, 1998). This is a joint plot of senator and issue points, which contains 100 line segments, one for each senator. Each senator that votes "aye" is connected to the "aye" point for the issue, while the senators that vote "nay" are connected to the "nay" point. Since the "aye" point and the "nay" points are centroids, this will create two stars for each issue, in which senators are connected with lines to the centroid of the group of senators that voted the same as they did on the issue. Homogeneity analysis moves the senators around iteratively to find the solution in which the squared length of all the lines, over all the issues, is as small as possible. Thus we are aiming for a solution in which each issue divides the cloud of 100 senator points into two clumps, which have a small within-clump variance and a large between-clump variance. We leave it to the reader to translate this same rational into the context of preference, choice, or testing data.

If we apply homogeneity analysis to the senate data, we find the solution (for senators) in Figure 28.1. In a more complete analysis we would make the star plots for the 20 issues and discuss the solution in terms of these plots. Here we merely look at the grouping of senators, with republicans on the left, democrats on the right, and moderates like McCain and Chafee in the middle.

As an alternative to the "clumping" aimed for by homogeneity ananlysis, we can try to place the senators in the plane in such a way that the "aye" groups and "nay" groups can be separated by straight lines, with the "aye's" on one side and the "nay's" on the other. Or, equivalently, that the convex hulls of the two groups are disjoint. Or that the "aye" and the "nay" groups are in complementary half-spaces. This is a *separation technique,* because groups can be very large and still be separable by a straight line.

There are two techniques that have been proposed to quantify and optimize goodness-of-fit in these types of separation models. The first is Nonlinear PCA, a special case of the nonmetric multidimensional scaling approach of Kruskal (1964a, 1964b). It uses a least squares loss function. We find directions in the space, one corresponding with each issue, such that the one-dimensional projections of the senators on the issue direction have all "nay's" to the left of all "aye's." This means that there is a perpendicular to the direction that separates the two groups.

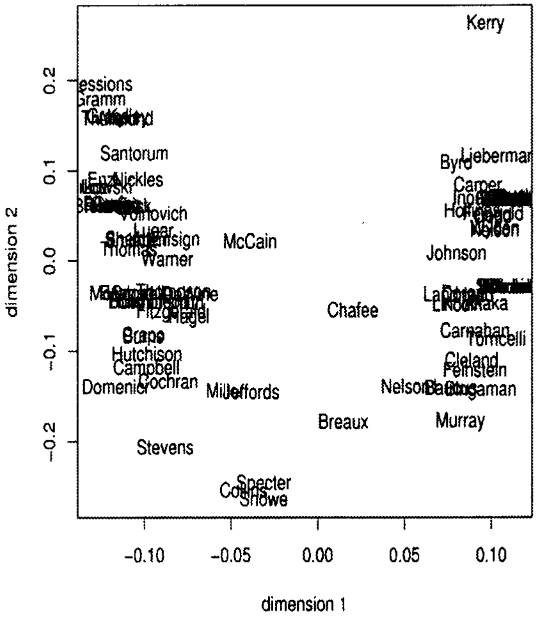

Figure 28.1. Senate homogeneity analysis solution.

The algorithm amounts to iteratively alternating PCA and transforming the binary data by monotone regression with Kruskal's *primary approach* to ties. Code in R is available from the author. Homogeneity can be shown to be Nonlinear PCA with the primary approach to ties, which is much more restrictive. If we apply the algorithm, starting from the homogeneity analysis solution, we need 209 alternating least squares iterations to find the solution in Figure 28.2. It has zero stress, which means that we can find directions for all issues weakly separating the "aye's" from the "nay's." Again, in a more complete analysis, we would look at the issue directions. Here we merely observe that imposing weaker restrictions on the representation comes at a price, because the solution, perfect as it may be in terms of the loss function, shows less detail and is more difficult to interpret. Because the algorithm only aims for weak separation, many senators are actually placed on the separating lines.

The second approach to separation uses a logistic likelihood function, and computes the maximum likelihood solution. This solution has long been popular in item response theory (Reckase, 1997) and in political science (Clinton Jackman, & Rivers, 2004). We use the majorization algorithm discussed in De Leeuw (in press), and after 975 iterations we find the solution in Figure 28.3. Again, code in R is available from the author.

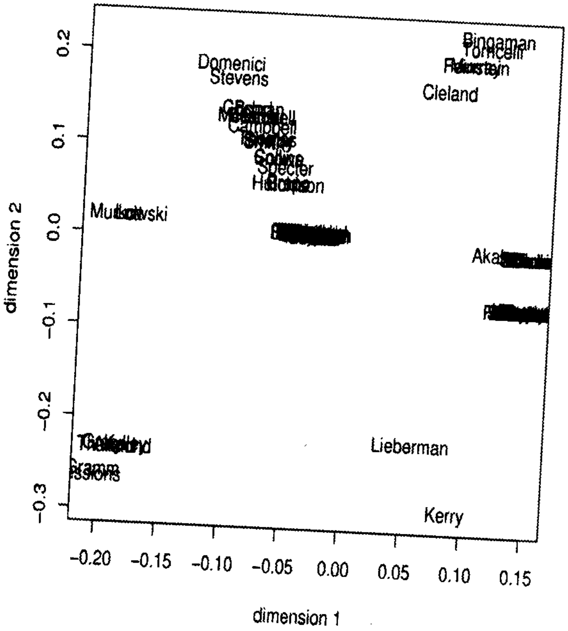

Figure 28.2. Senate nonmetric PCA solution.

For more detail we would have to look at the separating lines again, but at first sight it seems to solution is even harder to handle than the Nonlinear PCA solution.

What we find from these analyses is, in the first place a grouping and classification of the senators in the U.S. Senate. But in the second place we learn that the powerful nonmetric and likeliihood iterative methods for binary data are not necessarily an improvement from the substantive and interpretative point of view.

REFERENCES

Americans for Democratic Action. (2002) Voting record: Shattered promise of liberal progress. *ADA Today, 57*(1), 1–17.

Clinton, J., Jackman, S., & Rivers, D. (2004). The Statistical Analysis of Roll Call Data. *American Political Science Review, 98*, 355-370.

De Leeuw, J. (2006). Nonlinear principal component analysis and related techniques. In M. Greenacre & J. Blasius, (Ed.), *Multiple correspondence analysis and related methods*. New York: Chapman and Hall.

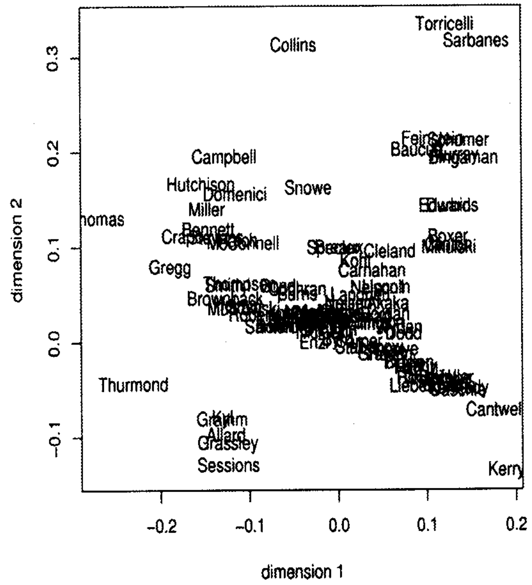

Figure 28.3. Senate logistic PCA solution.

De Leeuw, J. (in press). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis.*

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Greenacre, M., & Blasius, J. (Eds.). (2006). *Multiple correspondence analysis and related methods*. New York: Chapman and Hall.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1-27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29,* 115–129.

Michailidis, G., & De Leeuw, J. (1998). The Gifi system for descriptive multivariate analysis. *Statistical Science, 13,* 307–336.

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. Van Der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 271-286). New York: Springer.