SINGLE-PEAKED EXPONENTIAL DISTANCE MODELS FOR BINARY DATA

JAN DE LEEUW

1. Introduction

Suppose $G = \{g_{ij}\}$ is a binary $n \times m$ matrix. There are various ways to create low-dimensional Euclidean representations of this matrix. The first key choice we have to make is to decide on the *conditionality* of the matrix, the second choice is if we want to use *compactness* or *separation* as out leading scaling concept, and the third choice is how to deal with the *zeroes*.

Let us illustrate these choices with a simple example. Suppose n senators vote on m issues. Thus issues correspond with columns, senators with rows. If we decide to treat these data as *column-conditional*, we think of each issue as a variable classifying the senators into two groups: those who vote "aye" and those who vote "nay" on the issue. If we decide to use *compactness*, then we want to compute a geometric representation in which the senators in the "aye" group are close together. If zeroes (in this case "nay's") are treated as *informative* then we also want the senators in the "nay" group to be close together. If zeroes are *non-informative*, then the "nay" group can be dispersed arbitrarily, and only the "aye" group must be *homogeneous*. If we choose to emphasize *separation* instead of *compactness*, then in the informative case we will generally look for a hyperplane separating the "aye" and the "nay" groups, while in the non-informative case we will look for a sphere containing all "aye" senators and no "nay" senators.

Date: May 16, 2006.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

In the case of *row-conditionality*, we reverse the role of rows and columns. Instead of each issue defining a hyperplane separating the two groups of senators, we find a hyperplane for each senator, separating the issues the senator voted “aye” on from those she voted “nay” on. Or each senator defines a sphere containing only the issues she voted “aye” on. Since row-conditionality is simply column-conditionality applied to the transpose, there is no need to treat it separately.

Another example are archeological data, in which we either find or do not find each of *m* types of artifacts in each of *n* graves. The general idea is that objects are only put in the graves in a certain time interval. Thus zeroes are non-informative, because not finding an object can either mean that the grave is too early or the grave is too late. Only graves that contain objects should be close, in this case probably on a one-dimensional scale.

The same distinction can be made between ability items an attitude items. For ability items zeroes are informative, and we want those who fail an item to be close to each other and those who pass an item to be close as well. But for attitude items conservatives may disagree because they think an item is too progressive, and progressives may disagree with the same item because they think it is too concervative. So here zeroes are non-informative, and we only want the people who agree with an item to be close.

The discussion in this introduction, however brief, owes a great deal to the classical psychometric work of Stephenson [1953] and Thurstone [1959]. Our general approach to scaling qualitative data is largely due to Guttman [1941] and Coombs [1964]. The modern computerized approach is also due to Guttman, and is perhaps best described in Lingoes [1968]. There are many obvious connections with correspondence analysis [Guttman, 1946, 1950; Benzécri, 1973, 1980; Greenacre, 1984; Gifi, 1990] which can be thought of as the least squares version of compactness scaling.

2. Loss Function

The basic EDM loss function for binary matrices with a single-peaked representation is

$$\mathcal{D}(X,Y,\xi)=-2\left\{\sum_{(i,j)\in I_1}\log\frac{\beta_{ij}(\xi)\exp(\phi(x_i,y_j))}{1+\beta_{ij}(\xi)\exp(\phi(x_i,y_j))}+\right.\\ \left.+\sum_{(i,j)\in I_0}\log\frac{1}{1+\beta_{ij}(\xi)\exp(\phi(x_i,y_j))}\right\}\tag{1}$$

Here I_1 is the set of index pairs for which $g_{ij}=1$, and I_0 are the index pairs for which $g_{ij}=0$. There may be missing data, for which g_{ij} is neither zero or one.

In this paper we assume that the $\beta_{ij}(\xi)$ are specified using log-linear regression. This means we have

$$\beta_{ij}(\xi)=\exp(\gamma_{ij}(\xi)),$$

with

$$\gamma_{ij}(\xi)=\sum_{s=1}^{p}z_{ijs}\xi_s,$$

and with the z_{ijs} known regressors (which could be dummies or a design matrix). Thus the parameters separate into a *regression part*, which involves ξ, and a *geometry part*, which involves X and Y. One simple additive specification of the regression part has $\gamma_{ij}=\theta_i+\epsilon_j$. Then if $\phi(x_i,y_j)\equiv 0$ we have the Rasch model.

Observe there is no specific type of conditionality built into this specification, the rows and columns are interchangeable. We can introduce conditionality by further specifying the regression part. If $\gamma_i(\xi)$ only depends on the row-index i, for example, we have a form of column-conditionality, in which the regressors have external information about the n objects. And conversely we can have row-conditionality if $\gamma_j(\xi)$ only depends on the column index. Of course switching between these two forms of conditionality is easily done by just transposing the data.

3. Log-odds and Dédoublement

3.1. **Model Equivalences.**

The loss function we studied previously [De Leeuw, 2006a], if specialized to binary data, is

(2)
$$\mathcal{D}(X,Y,\xi)=-2\left\{\sum_{(i,j)\in I_0}\log\frac{\beta_{ij0}(\xi)\exp(\phi(x_i,y_{j0}))}{\beta_{ij0}(\xi)\exp(\phi(x_i,y_{j0}))+\beta_{ij1}(\xi)\exp(\phi(x_i,y_{j1}))}+\right.$$
$$\left.+\sum_{(i,j)\in I_1}\log\frac{\beta_{ij1}(\xi)\exp(\phi(x_i,y_{j1}))}{\beta_{ij0}(\xi)\exp(\phi(x_i,y_{j0}))+\beta_{ij1}(\xi)\exp(\phi(x_i,y_{j1}))}\right\}.$$

From Equation (1) the log-odds are given by

$$\lambda(x_i,y_j,\xi)=\log\beta_{ij}(\xi)+\phi(x_i,y_j), \tag{3a}$$

while from Equation (2) we have

$$\lambda(x_i,y_j,\xi)=\log\frac{\beta_{ij0}(\xi)}{\beta_{ij1}(\xi)}+[\phi(x_i,y_{j0})-\phi(x_i,y_{j1})]. \tag{3b}$$

If ϕ is the inner product, then Equation (3b) can be rewritten as

$$\lambda(x_i,y_j,\xi)=\log\frac{\beta_{ij0}(\xi)}{\beta_{ij1}(\xi)}+\phi(x_i,y_{j0}-y_{j1}). \tag{3c}$$

and the two representations are equivalent if their regression parts are.

If ϕ is squared Euclidean distance, and the regression part includes the additive specification $\gamma_{ij}=\theta_i+\epsilon_j$, then both (3a) and (3b) are just reparametrizations of the inner product specification. The differences between (3a) and (3b) are essential in the case in which ϕ is Euclidean distance, or in the case of squared Euclidean distance without a suitable regression part.

3.2. **Perfect Fit.**

As shown in De Leeuw [2006a] we will have perfect fit if we can find a solution to the system of inequalities $\lambda(x_i,y_j,\xi)>0$. In that case we can use the homogeneity of the specification to show that $\lambda(\kappa x_i,\kappa y_j,\kappa\xi)\rightarrow\infty$ if $\kappa\rightarrow\infty$, and thus $\mathcal{D}(\kappa X,\kappa Y,\kappa\xi)\rightarrow 0$.

The system of strict inequalities can be rewritten as

$$\phi(x_i,y_j)>-\gamma_{ij}(\xi).$$

For the negative distance combination rule this means

$$\|x_i - y_j\| < \gamma_{ij}(\xi).$$

Suppose $\gamma_{ij}(\xi)$ only depends on j. Then the x_i for which $g_{ij} = 1$ must be in a sphere centered at y_j with radius $\gamma_j(\xi)$. Each issue defines a sphere, with only the senators endorsing the issue in the sphere. Conversely, if $\gamma_{ij}(\xi)$ only depends on j, then the y_j for which which $g_{ij} = 1$ must be in a sphere centered at x_i with radius $\gamma_i(\xi)$. Each senator defines a sphere with only the issues she endorses in the sphere.

It may be interesting to require that some of the y_j are the same, i.e. some of the spheres are concentric. This is one possible way to introduce multi-category variables.

4. Algorithm

The algorithm combines two different majorization steps. In the first step we improve the regression part for fixed geometry, in the second part we improve the geometry for fixed regression part. Each of the two substeps can involve one or more inner iterations. The basic majorization result that both substep algorithms are based on is given in the Appendix.

4.1. **Improving Bias Estimates for Fixed Geometry.** Define

$$\begin{aligned}
\eta_{ij} &= \exp(\phi(x_i, y_j)),\\
\gamma_{ij}(\xi) &= \log\beta_{ij}(\xi) = \sum_{s=1}^{p} z_{ijs}\xi_s,\\
\pi_{ij}(\xi) &= \frac{\eta_{ij}\exp(\gamma_{ij}(\xi))}{1 + \eta_{ij}\exp(\gamma_{ij}(\xi))},\\
a_{ij}(\xi) &= \frac{2\pi_{ij}(\xi) - 1}{\gamma_{ij}(\xi)}
\end{aligned}$$

Also define the vectors u and $v(\xi)$ with elements

$$\begin{aligned} u_s &= \sum_{i=1}^{n}\sum_{j=1}^{m} g_{ij}z_{ijs}, \\ v_s(\xi) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij}(\xi)z_{ijs}, \end{aligned}$$

and the matrix $R(\xi)$ with elements

$$r_{st}(\xi) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}(\xi)z_{ijs}z_{ijt}.$$

The majorization algorithm for the logistic regression is simply

$$\xi^{(k+1)} = \xi^{(k)} + R(\xi^{(k)})^{-1}(u - v(\xi^{(k)})).$$

4.2. Improving Geometry Estimates for Fixed Bias.

Now

$$\begin{aligned} \pi(x_i, y_j) &= \frac{\beta_{ij}\exp(\phi(x_i, y_j))}{1 + \beta_{ij}\exp(\phi(x_i, y_j))}, \\ a(x_i, y_j) &= \frac{2\pi_{ij}(\xi) - 1}{\phi(x_i, y_j)}, \\ \tau(x_i, y_j) &= \phi(x_i, y_j) + \frac{g_{ij} - \pi(x_i, y_j)}{a(x_i, y_j)}. \end{aligned}$$

In a majorization step, while currently at $(X^{(k)}, Y^{(k)})$, we minimize

$$\sigma(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a(x_i^{(k)}, y_j^{(k)})(\phi(x_i, y_j) - \tau(x_i^{(k)}, y_j^{(k)}))^2,$$

over X and Y.

Appendix A. Basic Majorization Result

The core of the majorization algorithms used in this paper is the inequality in Lemma A.1, which is a minor generalization of a result by Jaakkola and Jordan [2000] and Groenen et al. [2003]. See also De Leeuw [2006b] for a proof and more discussion.

Lemma A.1. *Define*

$$\omega(x) = \log(1+\alpha\exp(x)),$$
$$\pi(x) = \frac{\alpha\exp(x)}{1+\alpha\exp(x)}.$$

Then

$$\omega(x) \leq \omega(y) + \pi(y)(x-y) + \frac{2\pi(x)-1}{2x}(x-y)^2.$$

Proof. Suppose

$$g(x) = \omega(y) + \pi(y)(x-y) + \frac{1}{2}a(x-y)^2$$

majorizes $\omega(x)$. Clearly $g(y) = \omega(y)$. Suppose there is a $z \neq y$ such that $g(z) = \omega(z)$. Thus

$$a = \frac{\omega(z) - \omega(y) - \pi(y)(z-y)}{\frac{1}{2}(z-y)^2}$$

Since g majorizes ω, we must also have $g'(z) = \omega'(z) = \pi(z)$. Thus $\pi(y) + a(z-y) = \pi(z)$, and

$$a = \frac{\pi(z)-\pi(y)}{z-y}.$$

Thus y and z are both support points of ω if

$$\frac{\pi(z)-\pi(y)}{z-y} = \frac{\omega(z) - \omega(y) - \pi(y)(z-y)}{\frac{1}{2}(z-y)^2}$$

or

$$\frac{\omega(z)-\omega(y)}{z-y} = \frac{1}{2}(\pi(z)+\pi(y))$$

□

References

J.P. Benzécri. *Analyse des Données: Correspondances*, volume 2. Dunod, Paris, 1973.

J.P. Benzécri. *Pratique de l'Analyse des Données: Analyse des Correspondances: Exposé Élémentaire*, volume 1. Dunod, 1980.

C. H. Coombs. *A Theory of Data*. Wiley, 1964.

J. De Leeuw. Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39, 2006a.

J. De Leeuw. Sharp Quadratic Majorization in One Dimension. Preprint Series 464, UCLA Department of Statistics, 2006b.

A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, England, 1990.

M.J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, New York, New York, 1984.

P.J.F. Groenen, P. Giaquinto, and H.L Kiers. Weighted Majorization Algorithms for Weighted Least Squares Decomposition Models. Technical Report EI 2003-09, Econometric Institute, Erasmus University, Rotterdam, Netherlands, 2003.

L. Guttman. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, editor, *The Prediction of Personal Adjustment*, pages 321–348. Social Science Research Council, New York, 1941.

L. Guttman. The Principal Components of Scale Analysis. In S.A. Stouffer and Others, editors, *Measurement and Prediction*. Princeton University Press, Princeton, 1950.

L. Guttman. An Approach for Quantifying Paired Comparisons and Rank Order. *Annals of Mathematical Statistics*, 17:144–163, 1946.

T.S. Jaakkola and M. I. Jordan. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, 10:25–37, 2000.

J.C. Lingoes. The Multivariate Analysis of Qualitative Data. *Multivariate Behavioral Research*, 3:61–94, 1968.

W. Stephenson. *The Study of Behavior*. University of Chicago Press, 1953.

L.L. Thurstone. *The Measurement of Values*. University of Chicago Press, 1959.

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`