Available online at www.sciencedirect.com
SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 21 – 39

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Principal component analysis of binary data by iterated singular value decomposition

Jan de Leeuw*

Department of Statistics, University of California, Los Angeles, 8130 Math Sciences Blvd., Los Angeles, CA 90095-1554, USA

Available online 21 August 2004

Abstract

The maximum-likelihood estimates of a principal component analysis on the logit or probit scale are computed using majorization algorithms that iterate a sequence of weighted or unweighted singular value decompositions. The relation with similar methods in item response theory, roll call analysis, and binary choice analysis is discussed. The technique is applied to 2001 US House roll call data.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Multivariate analysis; Factor analysis; Binary data; Item response models; Applications to social sciences

1. Introduction

Suppose $P=\{p_{ij}\}$ is an $n \times m$ binary data matrix, i.e. a matrix with elements equal to zero or one (representing yes/no, true/false, present/absent, agree/disagree). For the moment we suppose that P is complete, the case in which some elements are missing is discussed in a later section.

There are many examples of such binary data in the sciences. We give a small selection in Table 1, many more could be added.

Many different statistical techniques have been developed to analyze data of this kind. One important class is latent structure analysis (LSA), which includes latent class analysis, latent trait analysis and various forms of factor analysis for binary data. Alternatively,

* Tel.: +1-310-825-9550; fax: +1-310-206-5658.
E-mail address: deleeuw@stat.ucla.edu (Jan de Leeuw).

doi:10.1016/j.csda.2004.07.010

Table 1
Binary data

Discipline	Rows	Columns
Political science	Legislators	Roll calls
Education	Students	Test items
Systematic zoology	Species	Characteristics
Ecology	Plants	Transects
Archeology	Artefacts	Graves
Sociology	Interviewees	Questions

by recoding the data as a 2^m table, log-linear decompositions and other approximations of the multivariate binary distribution become available. There are also various forms of cluster analysis which can be applied to binary data, usually by first computing some sort of similarity measure between rows and/or columns. And finally there are variations of principal component analysis (PCA) specifically designed for binary data, such as multiple correspondence analysis (MCA).

In this paper, we combine ideas of LSA, more particularly item response theory and factor analysis of binary data, with PCA and MCA. This combination produces techniques with results that can be interpreted both in probabilistic and in geometric terms. Moreover, we propose algorithms that scale well, in the sense that they can be fitted efficiently to large matrices.

Our algorithm is closely related to the logistic majorization algorithm proposed by Groenen et al. (2003). We improve on their somewhat heuristic derivation, propose an alternative uniform logistic majorization, and a uniform probit majorization.

2. Problem

The basic problem we solve in this paper is geometric. We want to represent the rows of the data matrix as points and the columns as hyperplanes in low-dimensional Euclidean space $\mathbb{R}^r$, i.e. we want to make a drawing of our binary matrix. Rows i are represented as points a_i and the hyperplanes corresponding with columns j are parametricized as vectors of slopes b_j and as scalar intercepts c_j. The parameter r is the dimensionality of the solution. It is usually chosen to be equal to two, but drawings in different dimensionalities are also possible.

The drawing should be constructed in such a way that points a_i for which $p_{ij}=1$ should be on one side of hyperplane (b_j, c_j) and the points for which $p_{ij}=0$ should be on the other side. Or, equivalently, if we define the point sets $\mathscr{A}_{j1} = \{a_i \mid p_{ij} = 1\}$ and $\mathscr{A}_{j0} = \{a_i \mid p_{ij} = 0\}$, the convex hulls of $\mathscr{A}_{j1}$ and $\mathscr{A}_{j0}$ should be disjoint. Of course we want these disjoint convex hulls for all columns j simultaneously, and this is what makes the representation restrictive. Depending on the context, such a representation, if possible, is known as an inner product representation, a vector representation, or a compensatory representation. In the multidimensional scaling literature the algebraic version of the compensatory or vector model is usually attributed to Tucker (1960), although Coombs (1964) reviews some earlier work by his students and co-workers. The vector representation is most often applied to preference rank orders, but also quite often to binary choices and paired comparisons.

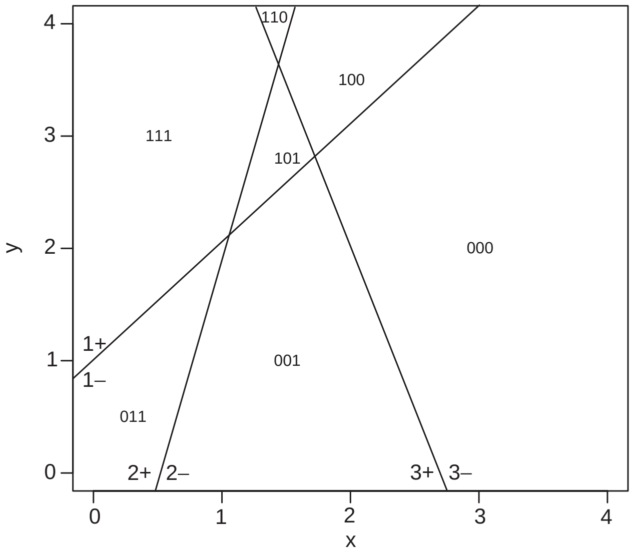


Fig. 1. Three variables.

The geometry of the compensatory representation was studied in a remarkable monograph by Coombs and Kao (1955, especially Chapter 5), and the results are summarized in Coombs (1964, Chapter 12). They show that m hyperplanes "in general position" in $\mathbb{R}^r$ will partition space into

$$\tau(m,r) = \sum_{k=0}^{r} \binom{m}{k}$$

disjoint convex regions. Each region corresponds with a profile, i.e. a vector of m zeroes and ones. Of the $\tau(m,r)$ regions there are $2m$ "open" regions, which extend to infinity. Since in most actual data analysis situations $n < \tau(m,r)$ we do not find all possible profiles (for a given set of hyperplanes) in our data, and since profiles correspond with regions the position of each row point (again, for a given set of hyperplanes) is only partially determined by the data. A small example, with three variables, is given in Fig. 1. We see $\binom{3}{0}+\binom{3}{1}+\binom{3}{2}=7$ regions, of which 6 are open. For these three variables, we can obtain perfect fit for any data matrix which does not contain the 010 profile.

The geometry of the compensatory model is different from that of MCA. In the general form of MCA recently discussed by De Leeuw (2003a) various measures of the size of a cloud of points in $\mathbb{R}^r$ are considered. Each column j of the data matrix defines a subset of the rows, those i for which $p_{ij}=1$, corresponding to a cloud of points a_i in the drawing. Call them the hit-points. The drawing is then constructed in such a way that the average over columns of the sizes of these clouds of hit-points is minimized. If cloud size is defined as squared distance to the centroid of the cloud this leads to multiple correspondence analysis (Greenacre, 1984; Gifi, 1990), a technique which is computationally relatively simple, because it requires computation of just one single SVD. More complicated measures of cloud size lead to more complicated optimization problems.

Thus in MCA we are satisfied if the clouds are small, not necessarily if they are linearly separated. Aiming for a small cloud suggests that the hit-points are in a small sphere, and the miss-points are outside that sphere. There is a variation of the technique which uses what the French MCA school calls dédoublement. In that case we let the miss-points also define a cloud, and we try to make both clouds small. This suggests two small disjoint spheres, and thus linear separation, which means that in the case of excellent fit MCA and the technique in this paper will be close. The difference between MCA with and without dédoublement can be made more precise by using classical theorems of Guttman and Mosteller (De Leeuw, 2003a).

In algebraic terms, we want to find a solution to the system of strict inequalities

$$\langle a_i, b_j \rangle > c_j \quad \forall p_{ij} = 1, \tag{1a}$$

$$\langle a_i, b_j \rangle < c_j \quad \forall p_{ij} = 0, \tag{1b}$$

where $\langle a_i, b_j \rangle$ is the inner product of a_i and b_j. Coombs and Kao (1955) give a simple pencil-and-paper procedure to solve these inequalities in the case $r = 2$, assuming that such a solution actually exists.

In general the system of inequalities defined by the data will not have a perfect solution. We have to find an approximate solution which is as good as possible, in the sense that it minimizes some loss function. Some simplifications are useful before we discuss the loss function we use. By defining $q_{ij} = 2p_{ij} - 1$, i.e. $q_{ij} = -1$ if $p_{ij} = 0$ and $q_{ij} = +1$ if $p_{ij} = 1$, we can write the system in the more compact form $q_{ij}(\langle a_i, b_j \rangle - c_j) > 0$. By appending c_j to b_j and appending -1 to a_i we can even write $q_{ij}(\langle a_i, b_j \rangle) > 0$, where a_i and b_j now have $r + 1$ elements. We collect them in an $n \times (r + 1)$ matrix of row scores A and in an $m \times (r + 1)$ matrix of column scores B. Now we can use the Hadamard product to rewrite our inequalities as $Q\, o AB' > 0$.

We fit a predicted matrix $\Pi = \{\pi_{ij}\}$ to the observed binary data matrix $P = \{p_{ij}\}$. The predicted matrix Π is a function of A and B. More specifically, we define $\pi_{ij}(A, B) = F(\langle a_i, b_j \rangle)$, where F is some cumulative distribution function such as the normal or logistic. Our computational problem is to minimize the distance between P and $\Pi(A, B)$ over (A, B), where distance is measured by the loss function

$$\mathscr{D}(A, B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} [p_{ij} \log \pi_{ij}(A, B) + (1 - p_{ij}) \log(1 - \pi_{ij}(A, B))]. \tag{2a}$$

The distance interpretation comes from the fact that $\mathscr{D}(A, B) \geqslant 0$ with equality if and only if $p_{ij} = \pi_{ij}(A, B)$ for all i and j. Remember that we require that the last column of A has all elements equal to -1. If we drop this restriction, we still fit separating hyperplanes, but they are now located in $\mathbb{R}^{r+1}$ and must pass through the origin.

By assuming symmetry of F, i.e. $F(-x) = 1 - F(x)$ for all x, and by using the fact that the p_{ij} are binary, we can also write

$$\mathscr{D}(A, B) = -\sum_{i=1}^{n} \sum_{j=1}^{m} \log F(q_{ij} \langle a_i, b_j \rangle). \tag{2b}$$

The usual way to motivate loss function (2) is to assume that the p_{ij} are outcomes of independent Bernoulli trials with probability of success $\pi_{ij} = F(\langle a_i, b_j\rangle)$. Then $\mathscr{D}$ is the negative log-likelihood, and minimizing $\mathscr{D}$ produces maximum-likelihood estimates. We do not emphasize this interpretation, since we do not think it is a realistic representation of actual data generating processes in any of the situations we are familiar with. Nevertheless, if you are so inclined, you can think of our computed points and lines as maximum-likelihood estimates. In any case, our loss function is a suitable way to measure distance between the observed and expected frequencies, or, alternatively, a measure how well the system of inequalities (1a) is satisfied.

It is easy to see that if the system (1a) has a solution, then minimizing $\mathscr{D}$ will find it, and the minimum of $\mathscr{D}$ in that case will be zero. Conversely, we can only make $\mathscr{D}$ converge to zero by letting (A, B) converge to a solution of (1a). In fact what minimizing $\mathscr{D}$ is trying to achieve is

$$\langle a_i, b_j\rangle \to \infty \quad \forall p_{ij} = 1, \tag{3a}$$

$$\langle a_i, b_j\rangle \to -\infty \quad \forall p_{ij} = 0, \tag{3b}$$

although it will generally not succeed in its goal. It can only do it perfectly if the system (1a) is solvable. Using the analysis of Coombs and Kao (1955) we can make these statements more precise. Of the $\frac{1}{2}(m^2 + m + 2)$ regions defined by m hyperplanes in two-space, there are $2m$ unbounded regions. All row points a in such an unbounded region will have the same profile, and moving them to infinity along the direction of recession of the region will make $\mathscr{D}$ smaller.

Define a matrix Λ with elements $\lambda_{ij} = F^{-1}(\pi_{ij})$. Then, we can write the basic relationship we have to fit as $\Lambda \approx AB'$. This expression shows that we are dealing with a rank r approximation problem of our data matrix on the F^{-1} scale, a problem that is solved by principal component analysis (PCA) or, equivalently, singular value decomposition (SVD), in the linear case in which Λ is observed directly.

3. Algorithm

We develop a majorization algorithm for minimizing (2), based on bounding the second derivative of the likelihood function. General information about majorization algorithms is in De Leeuw (1994), Heiser (1995), Lange et al. (2000). Majorization by bounding the second derivative is discussed in detail in Böhning and Lindsay (1988). In the computer science literature majorization is known as the variational method, or variational bounding (Jaakkola and Jordan, 2000).

So far we have only assumed that the cdf F is symmetric. We now make an additional assumption. Define the hazard function

$$h(x) \triangleq -\frac{\mathrm{d}\ \log\ F(x)}{\mathrm{d}x}$$

and assume there exists a function $w \geqslant 0$ such that

$$-\log\ F(x) \leqslant -\log\ F(y) + h(y)(x-y) + \tfrac{1}{2}w(y)(x-y)^2 \tag{4}$$

for all x and y. We say that w provides a quadratic majorization of log F. If we can choose $w(y)$ to be a constant, say w, then w provides a uniform quadratic majorization. In Appendix A, we show that for uniform majorization of the logistic cdf

$$w(y) \equiv \tfrac{1}{4}$$

for non-uniform majorization of the logistic cdf

$$w(y) = \frac{1-2F(y)}{2y}$$

and for uniform majorization of the normal cdf

$$w(y) \equiv 1.$$

The majorization of $-\log\ F$ can now be used to find a quadratic majorization of the negative log-likelihood.

Theorem 1. *Suppose A and $\tilde{A}$ are $n \times p$ matrices of row scores and B and $\tilde{B}$ are $m \times p$ matrices of column scores. Define the $n \times m$ matrices $\tilde{w}_{ij} \triangleq w(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)$, $\tilde{h}_{ij} \triangleq h(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)$, and $\tilde{z}_{ij} \triangleq \langle \tilde{a}_i, \tilde{b}_j\rangle - q_{ij}\frac{\tilde{h}_{ij}}{\tilde{w}_{ij}}$. Then*

$$\mathscr{D}(A, B) \leqslant \mathscr{D}(\tilde{A}, \tilde{B}) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\tilde{h}_{ij}^2}{\tilde{w}_{ij}} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\tilde{w}_{ij}(\langle a_i, b_j\rangle - \tilde{z}_{ij})^2.$$

Proof. By completing the square we can also write (4) as

$$-\log\ F(x) \leqslant -\log\ F(y) + \frac{1}{2}w(y)\left(x - \left(y - \frac{h(y)}{w(y)}\right)\right)^2 - \frac{1}{2}\frac{h^2(y)}{w(y)}. \tag{6}$$

Now substitute $q_{ij}\langle a_i, b_j\rangle$ for x and $q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle$ for y. Then

$$\begin{aligned} -\log\ F(q_{ij}\langle a_i, b_j\rangle) \leqslant & -\log\ F(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle) - \frac{1}{2}\frac{h^2(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)}{w(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)} \\ & + \frac{1}{2}w(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)\big(\langle a_i, b_j\rangle \\ & - \left(\langle \tilde{a}_i, \tilde{b}_j\rangle - q_{ij}\frac{h(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)}{w(q_{ij}\langle \tilde{a}_i, \tilde{b}_j\rangle)}\right)\Bigg)^2. \end{aligned} \tag{7}$$

Sum over i and j to obtain the required result. □

The algorithm works as follows. Start with some $A^{(0)}$ and $B^{(0)}$. Suppose $A^{(k)}$ and $B^{(k)}$ are the current best solution. We update them to find a better solution in two steps, similar to the E-step and the M-step in the EM-algorithm.

Algorithm 1 (*Majorization*).
***Step k*(1)** *Compute the matrices* $W^{(k)}$, $H^{(k)}$ *and* $Z^{(k)}$ *with elements*

$$\begin{aligned} w_{ij}^{(k)} &= w(q_{ij}\langle a_i^{(k)}, b_j^{(k)}\rangle), \\ h_{ij}^{(k)} &= h(q_{ij}\langle a_i^{(k)}, b_j^{(k)}\rangle), \\ z_{ij}^{(k)} &= \langle a_i^{(k)}, b_j^{(k)}\rangle - q_{ij}\frac{h_{ij}^{(k)}}{w_{ij}^{(k)}}. \end{aligned}$$

***Step k*(2)** *Solve the least squares matrix approximation problem*

$$\min_{A,B} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(k)}(z_{ij}^{(k)} - \langle a_i, b_j\rangle)^2$$

by using the (*weighted*) *SVD.*

Theorem 2. *The Majorization Algorithm* 1 *produces a decreasing sequence* $\mathscr{D}(A^{(k)}, B^{(k)})$ *of loss function values, and all accumulation points of the sequence* $(A^{(k)}, B^{(k)})$ *of iterates are stationary points*.

Proof. This follows by applying general results on majorization algorithms to Theorem 1. □

4. Implementation details

In the logistic case, we can choose if we want to use uniform on non-uniform majorization. If we use uniform majorization, then the substeps of the algorithm are unweighted SVDs. In an interpreted matrix language such as R or Matlab the SVD is usually implemented in object code in a very efficient manner, and consequently using uniform majorization seems a sensible choice. If we use non-uniform majorization we either have to write interpreted code for the substeps, or we have to apply majorization a second time to approximate the weighted SVD by an unweighted one (Kiers, 1997; Groenen et al., 2003). This last strategy basically means we are using uniform majorization in a roundabout way.

In our implementations of the algorithm so far, in the R language, we have always chosen uniform approximation. This is mainly to avoid using interpreted code as much as possible. Versions of the algorithm written in a compiled language such as C or FORTRAN could very well benefit from the more precise non-uniform majorization. But even in the compiled case customized iterative methods will probably have a hard time beating highly optimized SVD library routines.

In our R implementation, the initial estimate for A and B is simply taken as zero. This start is obviously not a good one, and we may obtain some improvement by using MCA instead. But we have to remember that the SVD majorization algorithm converges very fast in the initial steps, and then slows down to its linear (or even sublinear) rate, so the improvements of a very good start will presumably be not very large.

Also observe that starting with A and B equal to zero means that the first iteration computes the singular value decomposition of a matrix z_{ij} with element $p_{ij} - \frac{1}{2}$. This will be often close to the MCA solution, because MCA also computes a singular value decomposition of a matrix of weighted and centered p_{ij}.

If there are missing data then the matrix approximation problem becomes

$$\min_{X,Y} \sum \{(z_{ij}^{(k)} - \langle a_i, b_j \rangle)^2 \mid (i,j) \in N\},$$

where N is the subset of non-missing index pairs. We now use the classical least-squares augmentation trick, used in non-balanced ANOVA by Yates and Wilkinson and in least squares factor analysis by Thomson and Harman. See De Leeuw (1994), De Leeuw and Michailidis (1999) for references and for further discussion of augmentation.

We define inner iterations in each iteration of our majorization algorithm to impute the missing data. The inner iterations start with $a_i^{(k,0)} = a_i^{(k)}$ and $b_j^{(k,0)} = b_j^{(k)}$.

$$\tilde{z}^{(k,\ell)} = \begin{cases} z_{ij}^{(k)} & \text{if } (i,j) \in N, \\ \langle a_i^{(k,\ell)}, b_j^{(k,\ell)} \rangle & \text{if } (i,j) \notin N. \end{cases}$$

We then do an SVD to find $A^{(k,\ell+1)}$ and $B^{(k,\ell+1)}$, and continue the inner iterations. Actually, in our R implementation we only perform a single inner iteration, which basically means that we always perform a singular value decomposition on $\tilde{Z}^{(k,0)}$ which is just our previous $Z^{(k)}$ with missing elements imputed by setting them to the corresponding elements of the product of $A^{(k)}$ and $B^{(k)}$.

It may not always be a good idea to do a complete SVD after computing a new Z or $\tilde{Z}$, even if we use an SVD algorithm that only computes p singular vectors. This will depend, again, on the precise nature of the computing environment, the uniformity of the majorization, and the restrictions on the parameters.

We could use an iterative SVD method such as the simultaneous iteration method proposed by Daugavet (1968), and perhaps more familiar as the NILES/NIPALS method (Wold, 1966a, b). Only perform one or a small number of innermost iterations before updating $\tilde{H}$. This may ultimately lead to fewer computations.

Each NIPALS iteration

$$\begin{aligned} X &\leftarrow \tilde{H}Y(Y'Y)^{-1}, \\ Y &\leftarrow \tilde{H}'X(X'X)^{-1}, \end{aligned}$$

basically requires two matrix multiplications, so even for big matrices it is quite inexpensive. To identify along the way, the iterations are typically implemented as

$$\begin{aligned} X &\leftarrow \mathbf{orth}(\tilde{H}Y), \\ Y &\leftarrow \tilde{H}'X, \end{aligned}$$

where **orth** is an orthogonalization method such as Gram–Schmidt or QR. This makes the method identical to the Bauer–Rütishauser simultaneous iteration method, used in a similar context by Gifi (1990, pp. 98–99).

If an iterative SVD method is implemented, then we have to distinguish the outer iterations of the majorization algorithm, the inner iterations of the augmentation method to impute missing values, and the innermost iterations to compute or improve the SVD. The number of inner and innermost iterations will influence the amount of computation in an outer iteration and the convergence speed of the algorithm.

There is another straightforward block relaxation algorithm that can be used to complement (or replace) our majorization technique. If we fix *A* at its current value, then optimizing over b_j is a straightforward logit or probit regression problem. Thus one cycle consists of solving $n + m$ small logit or probit regression problems, which are convex minimization problems with a straightforward globally convergent Newton–Raphson implementation. This is the criss-cross regression algorithm used by researchers in generalized bilinear modeling (de Falguerolles and Francis, 1992; Gabriel, 1998; Van Eeuwijk, 1995). It has also been suggested by Poole (2001) for roll call analysis using the normal *F*.

In computer science, the work of Collins et al. (2002) extends PCA to the exponential family, using the fact that finding the optimal *A* for fixed *B* and finding the optimal *B* for fixed *A* are generalized linear model regression problems. Their development thus extends easily to Poisson and gamma versions of PCA. Schein et al. (2003) combine the ideas in Tipping (1999) and Collins et al. (2002) in the logistic case. Thus they used fixed row scores a_i and the non-uniform majorization of the logistic cdf. The algorithm then uses block relaxation to optimize over blocks *A* and *B*.

From the theoretical point of view, this alternative algorithm uses more and smaller blocks and will probably be slower and more likely to end up in a non-global minimum. On the other hand no majorization is involved and the Newton iterations will tend to be very precise and fast. It may be that this block regression algorithm can be used to refine the result of the majorization algorithm, but more research is necessary to compare the two.

5. Application areas

In this section, we try to give an overview of the various contexts in which the technique we discuss in this paper has occurred previously. As we shall see, the history is complicated, because different areas of statistics and data analysis are not necessarily aware of each other's work.

The multivariate compensatory model has been used for a long time in item response theory (IRT). The logistic version, for example, has been studied by Reckase (1997).

The main distinction we want to make in this section, however, is between fixed score and random score models, basically the same distinction as is made in factor analysis (Anderson and Rubin, 1956). In fixed score models (which we have discussed so far) each row gets a parameter vector a_i, and these "incidental" parameters are estimated along with the column parameters b_j. In random score models rows are sampled from some distribution *G*, and the row parameters can be integrated out of the likelihood. Random score models are better behaved from a statistical point of view, but seems less appropriate in areas such as roll call analysis in which row parameters (legislator ideal points) are really more interesting than column parameters (roll calls).

In Lawley (1944) a random score normal ogive model for the factor analysis of binary matrices was first proposed. Some other key references are the books by Lord and Novick

(1968, Chapter 16), Bartholomew (1987, Chapters 5 and 6) and the articles by Takane and De Leeuw (1987), Bock et al. (1988), McDonald (1997).

We can extend our majorization methods to optimize likelihood for random score models. More work remains to be done on implementation in this case, but as Jaakkola and Jordan (2000) observe it allows us to replace MCMC (Meng and Schilling, 1996; Beguin and Glas, 2001) and other delicate approximations (Bock et al., 1988; Bock and Schilling, 1997) by monotonically converging majorization algorithms.

There have been interesting recent related developments in multidimensional roll call analysis. Let us first outline the basic way of thinking in the field (Clinton et al., 2003). We work in $\mathbb{R}^r$. Each legislator has an ideal point a_i in this space and each roll call has both a yes-point u_j and a no-point v_j.

The utilities for legislator i to vote "yes" or "no" on roll call j have both a fixed and a random component. The fixed component is quadratic, that is squared distance. Thus

$$\begin{aligned}\underline{\xi}_{ij}^{1} &= -\tfrac{1}{2}\|a_i - u_j\|^2 + \underline{\varepsilon}_{ij}^{1},\\ \underline{\xi}_{ij}^{0} &= -\tfrac{1}{2}\|a_i - v_j\|^2 + \underline{\varepsilon}_{ij}^{0}.\end{aligned}$$

This means that the legislator will vote "yes" if $\underline{\xi}_{ij}^{1} > \underline{\xi}_{ij}^{0}$, i.e. if

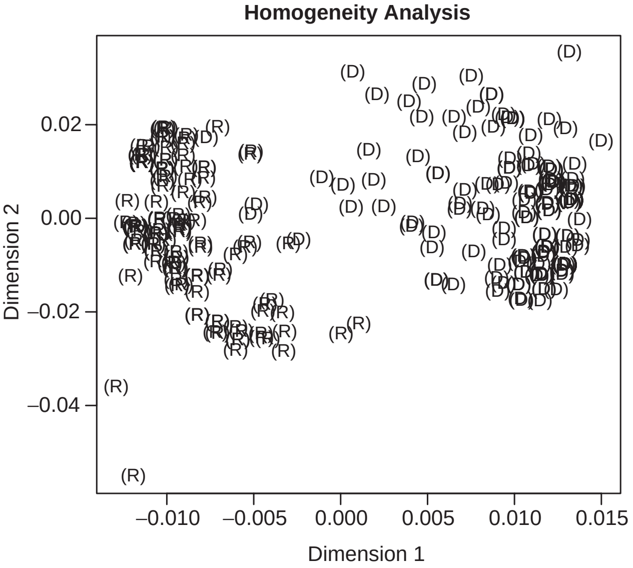
$$\langle a_i, u_j - v_j\rangle - c_j > -\left(\underline{\varepsilon}_{ij}^{1} - \underline{\varepsilon}_{ij}^{0}\right),$$

where $c_j = \frac{1}{2}(\|u_j\|^2 - \|v_j\|^2)$. If F is the cumulative probability distribution of $-(\underline{\varepsilon}_{ij}^{1} - \underline{\varepsilon}_{ij}^{0})$, then π_{ij}, the probability that legislator i will vote "yes" on roll call j is simply $F\{\langle a_i, u_j - v_j\rangle - c_j\}$, which is precisely the binary PCA representation we have studied so far. An identification analysis, using the second derivatives of the likelihood function, has been published by Rivers (2003). There are also random legislator versions of this model (Bailey, 2001), and, similar to IRT, the field has been caught in the maelstrom of MCMC (Jackman, 2000, 2001).

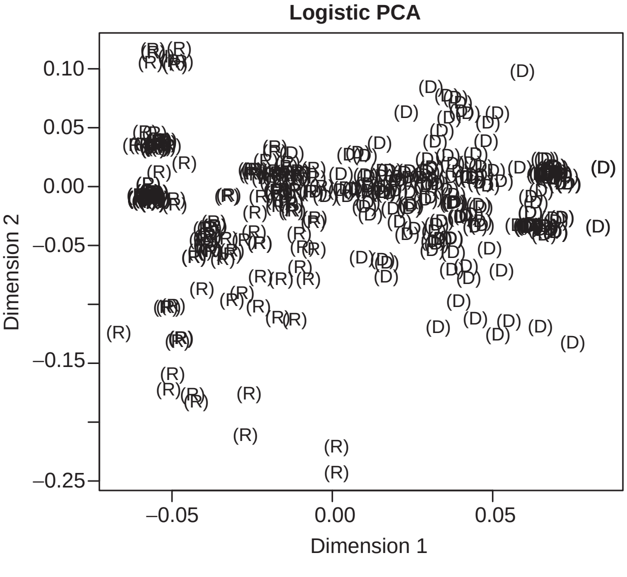
Both IRT and roll call analysis are special cases of binary choice problems. There are many probabilistic version of the compensatory model for preferential choice, which are reviewed very ably in Böckenholt and Gaul (1986) or Takane (1987). We can use the utility formulation of the roll call models to derive the paired comparison model in which individual i prefers j to ℓ, which we also write as $j >_i \ell$, if $\mathbf{prob}(j >_i \ell) = F(\langle a_i, b_j - b_\ell\rangle)$. Applying majorization to this representation allows us to iteratively apply the constrained PCA techniques of Takane and Hunter (2001), where we constrain the scores of pair (j, ℓ) to be of the form $b_j - b_\ell$.

6. Examples

We use the data for the 2001 House of Representatives, with votes on 20 issues selected by Americans for Democratic Action (Ada, 2002). Descriptions of the roll calls are given in Appendix B. We use the logit function for our binary PCA.



The algorithm needs 186 iterations to attain three decimals precision in loss function convergence. The proportion of correct classifications is 0.9387, up just a tiny bit from 0.9322 after the first iteration. The number of fitted probabilities which are indistinguishable from either zero or one is 36%.



In the first of the two plots we give the results of an MCA. The second plot gives the binary PCA. Legislators in the plots are labeled by party. In the MCA solution we have (marginally) fewer correct classifications. Parties are separated more clearly and we see some of the conservative democrats in the republican hull. The logistic PCA solution shows a little bit more detail within party, and less separation between the two party clumps. If we plot the 20 issues as lines in the plot, we find that almost all pass through the origin and separate the republicans from the democrats. It seems that voting in the house is so polarized that not much additional stable structure can be found.

This is one of the cases where MCA and logistic or probit binary PCA are close. In other examples, for instance in IRT, the differences will generally be more pronounced. Observe that binary PCA has the advantage that separating hyperplanes are also computed, while in MCA the issues are merely represented as centroids of the groups of individuals voting “aye” and “nay”. Binary PCA fits a system of inequalities and can be discussed in terms of the number of violations, MCA computes a convenient geometric representation with small clumps and it has not clear concept of violations at all.

A similar analysis of the 2001 Senate shows the same dominant dichotomy, with perhaps a bit more detailed structure.

7. Discussion and conclusions

The logistic (and probit) forms of principal component analysis discussed in this paper, and the corresponding majorization algorithms, seem to work well. Convergence is fast, and MCA provides a very good starting point. In the examples we have analyzed they do not show great improvements over simple homogeneity analysis, but this may very well be due to the extreme polarization in US politics. In IRT, for example, it is well known that homogeneity analysis (which is just a PCA of point correlations) gives results which can be quite different from those of simple logistic models. It will be a useful topic for further research to compare the two classes of techniques on different types of examples.

There are at least two additional interesting developments in the general approach we have outlined here. Both are currently being tested. In the first we extend the logistic majorization method to general choice and ranking models with multiple alternatives. This gives, for instance, a logistic version of multiple correspondence analysis for general sets of indicator matrices. In the second development logit and probit majorizations are used, in combination with the EM algorithm, to derive algorithms for random score versions of the various choice models. Although these new developments are largely untested, they seem to indicate that using quadratic majorization for choice models allows us to construct a very rich class of techniques. Compared with similar least squares based techniques, such as the ones in Gifi (1990), they have the advantage of maximizing a Bernoulli likelihood and not requiring arbitrary normalizations. They have the disadvantage of sometimes leading to partially degenerate solutions, in which points are moved to infinity to improve the fit.

Acknowledgements

Comments by George Michailides, Coen Bernaards, Doug Rivers, Yoshio Takane, Patrick Groenen, Jeff Lewis, Ulf Böckenholt and two anonymous reviewers on previous versions of this paper have been very helpful.

Appendix A. Logit and probit bounds

A.1. The logit case

We first derive a result familiar from Böhning and Lindsay (1988), Lange et al. (2000). Define

$$f(x) = -p\log\Psi(x) - (1-p)\log(1-\Psi(x)) = -\log\Psi(qx),$$

where $q = 2p - 1$ and

$$\Psi(x) = \frac{1}{1+\exp(-x)}.$$

Theorem 3. *f is strictly convex on* $(0, 1)$ *and has a uniformly bounded second derivative satisfying* $0 < f''(x) < \frac{1}{4}$.

Proof. Simple calculation gives

$$\begin{aligned} f'(x) &= q(\Psi(qx) - 1), \\ f''(x) &= \Psi(x)(1 - \Psi(x)). \end{aligned}$$

Clearly

$$0 < f''(x) < \tfrac{1}{4}$$

for all $0 < x < 1$, which is all we need.

This result can be improved by using a non-uniform bound due, independently, to Jaakkola and Jordan (2000) and Groenen et al. (2003). The following result shows the non-uniform bound is, in fact, sharper than the uniform one. □

Theorem 4.

$$\begin{aligned} \log\Psi(x) &\geqslant \log\Psi(y) + (1-\Psi(y))(x-y) + \frac{1-2\Psi(y)}{4y}(x-y)^2 \\ &\geqslant \log\Psi(y) + (1-\Psi(y))(x-y) - \tfrac{1}{8}(x-y)^2. \end{aligned}$$

Proof. Let

$$f(x) = \log\Psi(\sqrt{x}) - \frac{\sqrt{x}}{2}.$$

Then f is convex on $x \geqslant 0$ and

$$f'(x) = \frac{1-2\Psi(\sqrt{x})}{4\sqrt{x}}.$$

Convexity implies $f(x) \geqslant f(y) + f'(y)(x-y)$, i.e.

$$\log\, \Psi(\sqrt{x}) - \frac{\sqrt{x}}{2} \geqslant \log\, \Psi(\sqrt{y}) - \frac{\sqrt{y}}{2} + \frac{1-2\Psi(\sqrt{y})}{4\sqrt{y}}(x-y).$$

Collecting terms and changing variables gives

$$\log\, \Psi(x) \geqslant \log\, \Psi(y) + \frac{x-y}{2} + \frac{1-2\Psi(y)}{4y}(x^2-y^2),$$

which is, after some additional manipulation, the first inequality in the theorem.

The second inequality follows by a simple application of the mean value theorem.

$$\frac{1-2\Psi(y)}{4y} = \frac{\Psi(-y)-\Psi(y)}{4y} = \frac{-2\Psi(\xi)(1-\Psi(\xi))y}{4y}$$
$$= -\tfrac{1}{2}\Psi(\xi)(1-\Psi(\xi)) \geqslant -\tfrac{1}{8}.$$

In Groenen et al. (2003) the authors look for a quadratic majorization of the form

$$\log\, \Psi(x) \geqslant \log\, \Psi(y) + (1-\Psi(y))(x-y) + a(y)(x-y)^2,$$

which has the additional property that the derivatives of the functions on both sides of the inequality sign are not only equal at y but also at $-y$. They call this logistic majorization. For this we must have

$$a(y) = \frac{\Psi(-y)-\Psi(y)}{4y} = \frac{1-2\Psi(y)}{4y}$$

and this leads to Jaakkola–Jordan majorization, except for the fact that Groenen et al. (2003) still need an elaborate proof to show that this choice of $a(y)$ does indeed give a majorization function.

The difference between uniform and non-uniform majorization is clear from Fig. A.1, where we plot $\delta(-y, y)$ as a function of y. Uniform majorization corresponds with the horizontal line at -0.125. □

A.2. The probit case

For uniform majorization in the probit case we use a result previously given by Böhning (1999). Our proof is different. Define

$$f(x) = -p\, \log\, \Phi(x) - (1-p)\log(1-\Phi(x)) = -\log\, \Phi(qx),$$

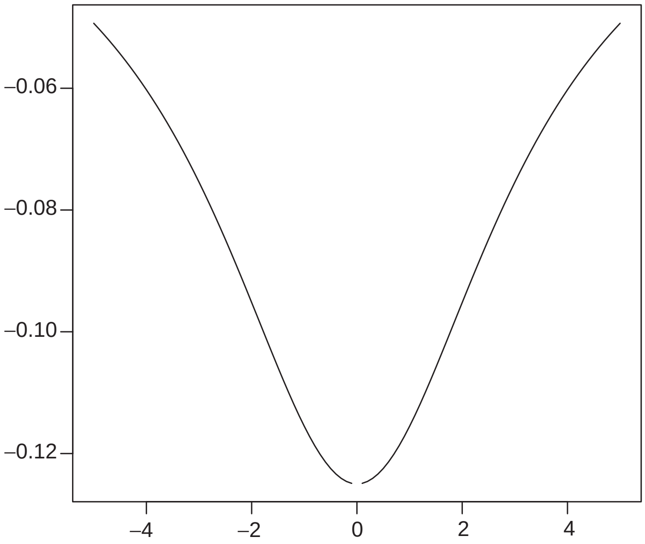


Fig. A.1.

where $q = 2p - 1$ and

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)\, \mathrm{d}z$$

with

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}.$$

Theorem 5. *The function f is strictly convex on* $(0, 1)$ *and has a uniformly bounded second derivative satisfying* $0 < f''(x) < 1$.

Proof. By simple computation

$$f'(x) = -qm(qx)$$

and

$$f''(x) = qxm(qx) + m^2(qx),$$

where

$$m(x) = \frac{\phi(x)}{\Phi(x)}$$

is the Inverse Mills' Ratio.

We now use a trick from Sampford (1953). Consider a standard normal random variable, truncated on the right (from above) at qx. Its variance is $1 - qxm(qx) - m^2(qx)$, see Johnson et al. (1994, Section 10.1), and because variance is positive, we see that $f''(x) < 1$. On the other hand, the variance, which is a conditional variance, must be less than that of the standard normal, which implies $f''(x) > 0$.

A.3. Sharpest quadratic majorization

The sharpest quadratic majorization (De Leeuw, 2003b) is obtained by defining

$$\delta(x, y) = \frac{f(x) - f(y) - f'(y)(x - y)}{(x - y)^2}$$

and then choosing the coefficient of the quadratic term in the majorization

$$f(x) \geqslant f(y) + f'(y)(x - y) + a(y)(x - y)^2$$

as

$$a(y) = \inf_{x \neq y} \delta(x, y).$$

In the logit case we know, from Theorem 4 that

$$\delta(x, y) \geqslant \frac{1 - 2\Psi(y)}{4y}$$

and, from the proof, we have equality if and only if $x^2 = y^2$. Thus

$$\inf_{x \neq y} \delta(x, y) = \delta(-y, y) = \frac{1 - 2\Psi(y)}{4y}$$

and Jaakkola–Jordan majorization is sharp.

In the probit case (we conjecture that) $\delta(x, y)$ is increasing in x, and

$$\inf_{x \neq y} \delta(x, y) = \lim_{x \to -\infty} \delta(x, y) = -\tfrac{1}{2}$$

and uniform majorization is sharp.

Appendix B. ADA roll call descriptions

Code	Description	ADA
HR 333	Bankruptcy overhaul	Yes
SJ Res 6	Ergonomics rule disapproval	No
HR 3	Income tax reduction	No
HR 6	Marriage tax reduction	Yes
HR 8	Estate tax relief	Yes
HR 503	Fetal protection	No
HR 1	School vouchers	No
HR 1836	Tax cut reconciliation bill	No
HR 2356	Campaign finance reform	No
HJ Res 36	Flag desecration	No
HR 7	Faith-based initiative	Yes
HJ Res 50	China normalized trade relations	Yes
HR 4	ANWR drilling ban	Yes
HR 2563	Patients' rights/HMO liability	No
R 2563	Patients' bill of rights	No
HR 2944	Domestic partner benefits	No
HR 2586	US military personnel overseas/abortions	Yes
HR 2975	Anti-terrorism authority	No
HR 3090	Economic stimulus	No
HR 3000	Trade promotion authority/fast track	No

The votes selected cover a full spectrum of domestic, foreign, economic, military, environmental and social issues. We tried to select votes which display sharp liberal/conservative contrasts. In many instances we have chosen procedural votes: amendments, motions to table, or votes on rules for debate. Often these votes reveal true attitudes frequently obscured in the final votes.

References

Ada, 2002. 2001 Voting record: shattered promise of liberal progress. ADA Today 57 (1), 1–17.

Anderson, T.W., Rubin, H., 1956. Statistical inference in factor analysis. In: Neyman, J. (Ed.), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 5, University of California Press, pp. 111–150.

Bailey, M., 2001. Ideal point estimation with a small number of votes: a random-effects approach. Political Anal. 9 (3), 192–210.

Bartholomew, D.J., 1987. Latent Variable Models and Factor Analysis. Griffin, London.

Beguin, A.A., Glas, C.W., 2001. MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika 66, 541–562.

Bock, R.D., Schilling, S., 1997. High-dimensional full information factor analysis. In: Berkane, M. (Ed.), Latent Cariable Modeling and Applications to Causality. Springer, Berlin.

Bock, R.D., Gibbons, R., Muraki, E., 1988. Full-information factor analysis. Appl. Psychol. Meas. 12, 261–280.

Böckenholt, I., Gaul, W., 1986. Analysis of choice behaviour via probabilistic ideal point and vector models. Appl. Stochastic Models Data Anal. 2, 209–226.

Böhning, D., 1999. The lower bound method in probit regression. Comput. Statist. Data Anal. 30, 13–17.

Böhning, D., Lindsay, B.G., 1988. Monotonicity of quadratic-approximation algorithms. Ann. Inst. Statist. Math. 40 (4), 641–663.

Clinton, J., Jackman, S., Rivers, D., 2003. The statistical analysis of roll call data. http://jackman.stanford.edu/papers/masterideal.pdf.

Collins, M., Dasgupta, S., Shapire, R.E., 2002. A generalization of principal component analysis to the exponential family. Adv. Neural Inform. Process. Systems 14.

Coombs, C.H., 1964. A Theory of Data. Wiley, New York.

Coombs, C.H., Kao, R.C., 1955. Nonmetric factor analysis. Engineering Research Bulletin 38, Engineering Research Institute, University of Michigan, Ann Arbor.

Daugavet, V.A., 1968. Variant of the stepped exponential method of finding some of the first characteristics values of a symmetric matrix. USSR Comput. Math. Phys. 8 (1), 212–223.

De Leeuw, J., 1994. Block relaxation methods in statistics. In: Bock, H.H., Lenski, W., Richter, M.M. (Eds.), Information Systems and Data Analysis. Springer, Berlin.

De Leeuw, J., 2003a. Homogeneity analysis of pavings. http://jackman.stanford.edu/ideal/MeasurementConference/abstracts/homPeig.pdf.

De Leeuw, J., 2003b. Quadratic majorization. http://gifi.stat.ucla.edu/pub/quadmaj.pdf.

De Leeuw, J., Michailidis, G., 1999. Block relaxation algorithms in statistics. http://gifi.stat.ucla.edu/pub/block.pdf.

de Falguerolles, A., Francis, B., 1992. Algorithmic approaches for fitting bilinear models. In: Dodge, Y., Whittaker, J. (Eds.), COMPSTAT 1992. Physika, Heidelberg, Germany.

Gabriel, K.R., 1998. Generalized bilinear regression. Biometrika 85, 689–700.

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

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

Groenen, P.J.F., Giaquinto, P., Kiers, H.L., 2003. Weighted majorization algorithms for weighted least squares decomposition models. Technical Report EI 2003-09, Econometric Institute, Erasmus University, Rotterdam, Netherlands.

Heiser, W.J., 1995. Convergent computing by iterative majorization: theory and applications in multidimensional data analysis. In: Krzanowski, W.J. (Ed.), Recent Advantages in Descriptive Multivariate Analysis. Clarendon Press, Oxford.

Jaakkola, T.S., Jordan, M.I., 2000. Bayesian parameter estimation via variational methods Bayesian parameter estimation via variational methods. Statist. Comput. 10, 25–37.

Jackman, S., 2000. Estimation and inference are missing data problems: unifying social science statistics via Bayesian simulation. Polit. Anal. 8 (4), 307–332.

Jackman, S., 2001. Multidimensional analysis of roll call data via Bayesian simulation: identification, estimation, inference, and model checking. Polit. Anal. 9 (3), 227–241.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, 2nd Edition, vol. I. Wiley, New York.

Kiers, H.A.L., 1997. Weighted least squares fitting using iterative ordinary least squares algorithms. Psychometrika 62, 251–266.

Lange, K., Hunter, D.R., Yang, I., 2000. Optimization transfer using surrogate objective functions. J. Comput. Graph. Statist. 9, 1–20.

Lawley, D.N., 1944. The factorial analysis of multiple item tests. Proc. Roy. Soc. Edinburgh 62-A, 74–82.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

McDonald, R.P., 1997. Normal ogive multidimensional model. In: Van Der Linden, W.J., Hambleton, R.K. (Eds.), Handbook of Item Response Theory. Springer, Berlin.

Meng, X.-L., Schilling, S., 1996. Fitting full-information item factor analysis models and an empirical investigation of bridge sampling. J. Amer. Statist. Assoc. 91, 1254–1267.

Poole, K.T., 2001. The geometry of multidimensional quadratic utility in models of parliamentary roll call voting. Polit. Anal. 9 (3), 211–226.

Reckase, M.D., 1997. A linear logistic multidimensional model. In: Van Der Linden, W.J., Hambleton, R.K. (Eds.), Handbook of Item Response Theory. Springer, Berlin.

Rivers, D., 2003. Identification of multidimensional spatial voting models. http://jackman.stanford.edu/ideal/MeasurementConference/abstracts/river03.pdf.

Sampford, M.R., 1953. Some inequalities on Mill's ratio and related functions. Ann. Math. Statist. 24, 130–132.

Schein, A.I., Saul, L.K., Ungar, L.H., 2003. A generalized linear model for principal component analysis of binary data. In: Bishop, C.M., Frey, B.J., (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. http://research.microsoft.com/conferences/aistats2003/proceedings/papers.htm.

Takane, Y., 1987. Analysis of covariance structures and probabilistic binary choice data. Commun. Cognition 20, 45–62.

Takane, Y., De Leeuw, J., 1987. On the relationship between item response theory and factor analysis of discreticized variables. Psychometrika 52, 393–408.

Takane, Y., Hunter, M., 2001. Constrained principal components analysis: a comprehensive theory. Appl. Algebra Eng. Commun. Comput. 12, 391–419.

Tipping, M., 1999. Probabilistic visualization of high-dimensional binary data. Adv. Neural Inform. Process. Systems 11, 592–598.

Tucker, L.R., 1960. Intra-individual and inter-individual multidimensionality. In: Gulliksen, H., Messick, S. (Eds.), Psychological Scaling: Theory and Applications. Wiley, New York.

Van Eeuwijk, F.A., 1995. Multiplicative interaction in generalized linear models. Biometrics 51, 1017–1032.

Wold, H., 1966a. Estimation of principal components and related models by iterative least squares. In: Krishnaiah, P.R. (Ed.), Multivariate Analysis. Academic Press, New York.

Wold, H., 1966b. Nonlinear estimation by iterative least squares procedures. In: David, F.N. (Ed.), Research Papers in Statistics. Festschrift for J. Neyman. Wiley, New York.