Majorizing logit loss functions for the multidimensional representation of categorical data

Jan de Leeuw
UCLA Statistics
Presentation at UC Riverside, 02/15/05

The Data

The data we study are measurements of n *objects* on each of m *variables*. This corresponds with the spreadsheet format in many software packages and with the data-frame in S/R.

In our setup all data are *categorical*, which means each variable maps the observations into a finite number of categories. Categories can be a finite subset of the reals, a finite ordered set, the set {0,1}, or just an arbitrary finite set.

Observe the finiteness assumption can be made without loss of generality.

Examples

- legislators and votes
- students and multiple choice items
- animals and morphology
- plants and transects
- artifacts and graves
- interviewees and survey questions

Coding

$$G = (G_1 \mid \cdots \mid G_m)$$

The submatrices G(j) are *indicator matrices* (or dummies) for variables. They indicate which category the objects are in. They are binary, add up to one row-wise, and have orthogonal columns. Each indicator matrix codes a partitioning of the objects.

Variable j has k(j) categories, thus G(j) is n x k(j).

There can be missing data (incomplete indicators).

Homogeneity

Objects are represented as points in low-dimensional Euclidean Space.

In *homogeneity* or *clumping* techniques we want the subsets of objects that are in the same category to be "small". Or: we want within-category distances to be relatively small. And we want this for all variables simultaneously.

Many definitions of "small" are possible, we use the size of the Gifi star.

The Gifi System

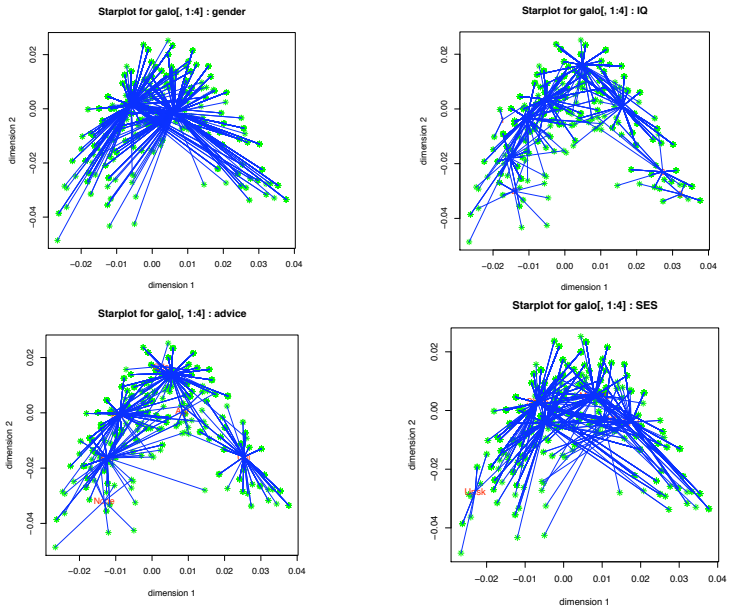
Minimize star-size: the sum of the squared distances of the object points to the centroid of their category.

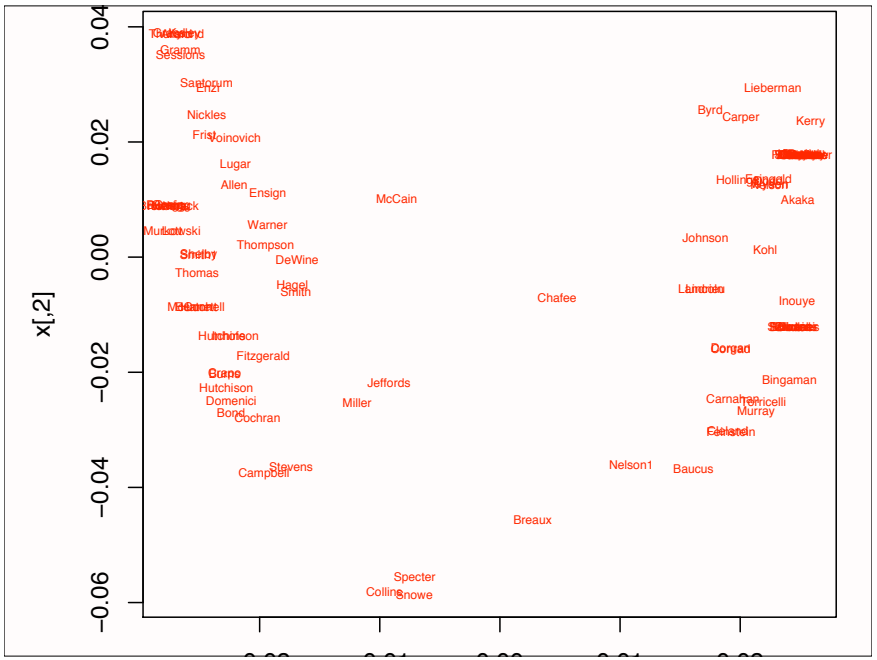
$$\min_{X'X=I} \min_{Y_j} \sum_{j=1}^{m} \mathbf{SSQ}(X - G_j Y_j)$$

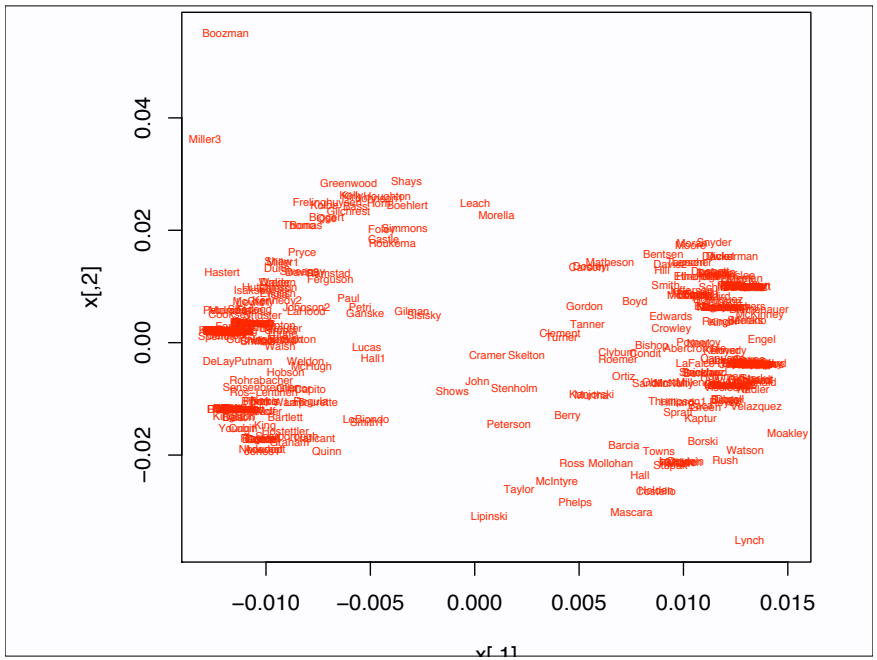
In addition the Gifi System (Gifi, 1990, also Michailides and De Leeuw, Statistical Science, 1998) allows for rank and additivity constraints on the Y(j). This makes it possible to have regression, principal component analysis, canonical analysis, and so on as special cases.

Homogeneity analysis with Gifi Stars is also known as *multiple correspondence analysis*.

- Advantage: computationally simple (SVD and ALS). Matrix calculations. Using sparsity.
- Advantage: all of classical descriptive MVA, extended to mixed level data.
- Disadvantage: a normalization is needed (and rather arbitrary).
- Disadvantage: horseshoes (a least squares effect).
- Disadvantage: not maximum likelihood and no model (if you feel that is important).
- Disadvantage: Statistical stability analysis (standard errors, confidence intervals) is available, but tedious.







Separation

We now use a probability model for the same type of data, summarizing 150 years of accumulated wisdom (or at least practice). The probability that $g_{ij\ell} = 1$ is

$$\pi_{ij\ell}(X,Y) = \frac{\exp(f(x_i, y_{j\ell}))}{\sum_{\nu=1}^{k_j} \exp(f(x_i, y_{j\nu}))}$$

A similar expression can be used for ordered categories, using a cdf such as the normal or the double exponential.

Deviance

We want to minimize

$$\mathcal{D} = -2\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} g_{ij\ell}\log \pi_{ij\ell}(X,Y)$$

where the data are as before (indicators, dummies) and satisfy, for all i and j,

$$\sum_{\ell=1}^{k_j} g_{ij\ell} = 1$$

We call this *separation methods* because we have minimum loss (perfect fit) if f(x,y) is homogeneous and if the solution satisfies the following inequalities

$$g_{ij\ell}\{f(x_i, y_{j\ell}) - f(x_i, y_{j\nu})\} \geq 0 \qquad \forall i, j, \ell, \nu$$

For most geometric models, using inner products or distances, we have homogeneity, and the inequalities have a straightforward geometrical interpretation in terms of separation.

Algorithms

So far, many ad hoc techniques have been proposed to compute maximum likelihood estimates for various specific models. Some work well, some don't.

Our purpose in this presentation is to present a general approach based on *quadratic majorization*. This class of algorithms has the desirable property that it computes maximum likelihood estimates by solving a sequence of least squares problems, which are generally much simpler. It also produces an algorithm which is globally convergent.

Majorization: A digression

The problem we want to solve

$$\min_{\theta \in \Theta} \phi(\theta)$$

Now suppose there is a *majorization function* $\psi(\theta, \xi)$ such that

$$\phi(\theta) \leq \psi(\theta, \xi) \qquad \forall \theta, \xi \in \Theta$$

$$\phi(\theta) = \psi(\theta, \theta) \qquad \forall \theta \in \Theta$$

Finding a suitable majorization function is partly a box of tricks, and partly art (like integration). Quadratic majorization is one of the main tricks in the box.

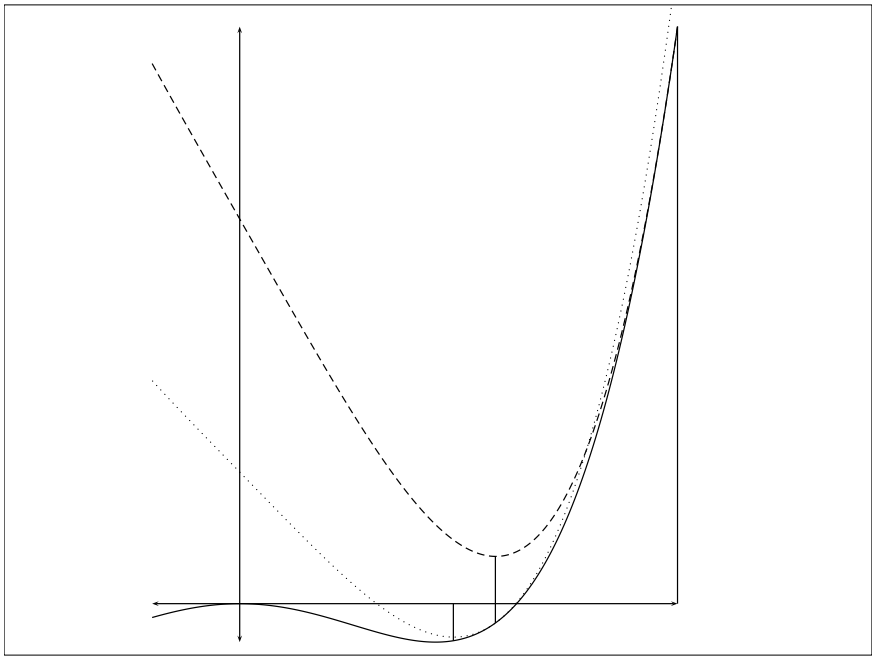
Define the algorithm

$$\theta^{(k+1)} = \operatorname*{argmin}_{\theta} \psi(\theta, \theta^{(k)})$$

Then (sandwich inequality)

$$\phi(\theta^{(k+1)}) \leq \psi(\theta^{(k+1)}, \theta^{(k)}) \leq \psi(\theta^{(k)}, \theta^{(k)}) = \phi(\theta^{(k)})$$

Thus minimizing the majorization function decreases the objective function. Under some additional conditions, this guarantees convergence of the algorithm to a local minimum.



Quadratic Majorization

In this presentation we are interested in the case where we can find a matrix H such that

$$\mathcal{D}^2\phi(\theta) \leq H \qquad \forall\theta \in \Theta$$

Then

$$\phi(\theta) \leq \phi(\xi) + (\theta-\xi)'\mathcal{D}\phi(\xi) + \frac{1}{2}(\theta-\xi)'H(\theta-\xi)$$

which provides a majorization function quadratic in θ. Completing the square gives

$$\phi(\theta) \leq \phi(\xi) - \frac{1}{2}\tilde{\theta}'H\tilde{\theta} + \frac{1}{2}(\theta-\tilde{\theta})'H(\theta-\tilde{\theta})$$

with $\tilde{\theta} = \xi - H^{-1}\mathcal{D}\phi(\xi)$

Logit Majorization

Theorem: Suppose $x \in \mathbb{R}^K$

$$\pi_k(x) = \frac{\exp(x_k)}{\sum_{\ell=1}^K \exp(x_\ell)}$$

$$f(x) = -\sum_{k=1}^K y_k \log \pi_k(x)$$

then

$$0 \leq \mathcal{D}^2 f(x) = \Pi(x) - \pi(x)\pi(x)' \leq \frac{1}{2}I.$$

Probit Majorization

Although we do not use this result in the presentation, we'll throw it in for good measure.

Theorem: Suppose $-\infty \leq \alpha < \beta \leq +\infty$ and

$$f(x) = -\log[\Phi(\beta+x) - \Phi(\alpha+x)]$$

then $\quad 0 < f''(x) < 1 \qquad \forall x$

Where does Gifi come in ?

$$\mathcal{D} = -2\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} g_{ij\ell}\log\frac{\exp(f(x_i,y_{j\ell}))}{\sum_{\nu=1}^{k_j}\exp(f(x_i,y_{j\nu}))}$$

If

$$f(x_i,y_{j\ell}) = \tau_{j\ell} + x_i'y_j$$

we have the deviance for PCA. But other functions (besides the inner product) are possible too. And we can have mixed level data (some logit, some probit, some tobit), and constraints on the Y as in Gifi.

The loss function is of the general form

$$\mathcal{D} = \sum_{i=1}^{n}\sum_{j=1}^{m} g(f(x_i,y_j))$$

where we have an upper bound B for the second derivative of g. Quadratic majorization leads to the minimization of

$$\mathcal{S} = \sum_{i=1}^{n}\sum_{j=1}^{m}[f(x_i,y_j) - h(\tilde{x}_i,\tilde{y}_j)]^2$$

in each iteration, where

$$h(\tilde{x}_i,\tilde{y}_j) = f(\tilde{x}_i,\tilde{y}_j) - \frac{1}{B}g'(f(\tilde{x}_i,\tilde{y}_j))$$

is evaluated at the current solution (with the tilde).

And thus ...

We have replaced logit/probit/tobit maximum likelihood by iterative least squares, and often we know how to solve these LS subproblems (in PCA, use the SVD).

Two additional observations are very useful here.

First, there is no need to actually minimize the majorization function, it suffices to decrease it.

Second, it is easy to incorporate missing data. This last observation makes it possible to analyze rank orders and more general choice structures.