Title

Principal Component Analysis of Categorical Data, with Applications to Roll-Call Analysis

Authors
Jan de Leeuw
Jeffrey Lewis

Publication Date
2011-10-25

Principal Component Analysis of Categorical Data, with Applications to Roll-Call Analysis

Jan de Leeuw
Department of Statistics UCLA

Jeffrey Lewis
Department of Political Science UCLA

Choice Models

We have n actors (i=1,...,n) confronted with m (j=1,...,m) choice sets. To summarize 150 years of accumulated wisdom (or at least practice): for unordered alternatives

$$\pi_i(k|K) = \frac{\exp(\phi(x_i, z_k))}{\sum_{\ell \in K} \exp(\phi(x_i, z_\ell))}$$

and for ordered ones use a cdf F and assume

$$\pi_i(k|K) = F(\tau_{k+1} + \phi_i(\theta)) - F(\tau_k + \phi_i(\theta))$$

Deviance

We want to minimize

$$\mathcal{D} = -2 \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k \in K_j} y_{ijk} \log \pi_i(k|K_j)$$

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

$$\sum_{k \in K_j} y_{ijk} = 1$$

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

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 part tricks, part art (like integration).

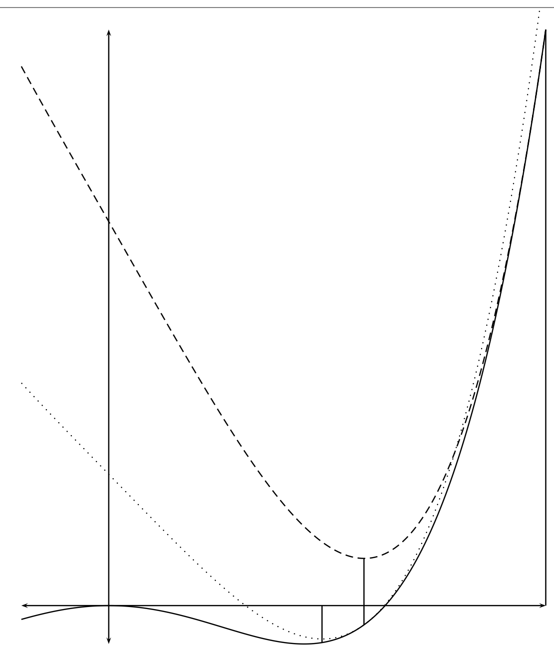
Define the algorithm

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

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

Probit Majorization

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

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

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

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

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

Where does PCA come in ?

$$\mathcal{D} = -2 \sum_{i=1}^n \sum_{j=1}^m \sum_{k \in K_j} y_{ijk} \log \frac{\exp(\tau_{jk} + x_i' z_j)}{\sum_{\ell \in K_j} \exp(\tau_{jk} + x_i' z_j)}$$

$$\mathcal{D} = -2 \sum_{i=1}^n \sum_{j=1}^m \sum_{k \in K_j} y_{ijk} \log\left[\Phi(\tau_{j,k+1} + x_i' z_j) - \Phi(\tau_{jk} + x_i' z_j)\right]$$

And thus ...

Using quadratic majorization on these deviances amounts to minimizing in each iteration a function of the form

$$\sum_{i=1}^n \sum_{j=1}^m (\tilde{y}_{ij} - x_i' z_j)^2$$

and we know how to do this. In fact, two additional observations are very useful here. First, there is no need to actually minimize the majorization function, it suffices to merely decrease it. Second, it is trivial to incorporate missing data, in which i does not choose from j. This last observation makes it possible to analyze rank orders and more general choice structures.