EM algorithms for Thurstonian and Bradley-Terry-type random permutation models

Villő Csiszár

Department of Probability and Statistics
Eötvös Loránd University, Budapest

Prague Stochastics 2010

Outline

Outline

Thurstonian, or order statistics models

- There are n items labelled $1, 2, \ldots, n$.
- Judges are asked to order the items according to preference.
- An ordering is a permutation π, where $\pi(1)$ is the label of the most preferred, $\pi(n)$ is the label of the least preferred item.

The orderings are supposed to be random and independent for all judges, with

$$p(\pi) = P(X_{\pi(1)} < X_{\pi(2)} < \cdots < X_{\pi(n)}),$$

where $(X_1, \ldots, X_n)$ is an n-dimensional continuous random vector.

Special cases

- Suppose the X_i are independent.
- Moreover, suppose the distributions of the X_i differ only in a location parameter, i.e. the ith distribution function is

$$F_i(x) = F(x - \mu_i).$$

- Moreover, suppose the baseline distribution is Gumbel,

$$F(x) = 1 - \exp(-\exp(x)).$$

This is called Plackett-Luce model.

Studied by e.g. Thurstone (1927), Daniels (1950), Luce (1959), Yellott (1977), Henery (1983), Stern (1990).

General independent model

We choose to study the model where X_i are independent, but otherwise arbitrary.

Problem

How to estimate the distribution functions F_i from the data?

Observe that a strictly monotone increasing transformation of the X_i's defines the same permutation model, so the problem is not well-defined.
Instead: Suppose each X_i is concentrated on a finite number of points, w.l.g. let the support of X_i be

$$T_i = \{s_{ij} = j + i/n : j = 1, \ldots, J\}.$$

We need to estimate $p(i,j) = P(X_i = s_{ij})$.

EM algorithm

We can use the EM method. Let the full observation be $\{X_{i,r} : 1 \leq i \leq n, 1 \leq r \leq m\}$. The E-step calculates

$$Q(p, p^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \log p(i,j) \sum_{r=1}^{m} P(X_{i,r} = s_{ij} | \pi_r, p^{(t)}),$$

which is maximized by the M-step as

$$p^{(t+1)}(i,j) = \frac{1}{m} \sum_{r=1}^{m} P(X_{i,r} = s_{ij} | \pi_r, p^{(t)}).$$

The conditional probabilities appearing in the formula can be calculated recursively in polynomial time.

Properties, Example

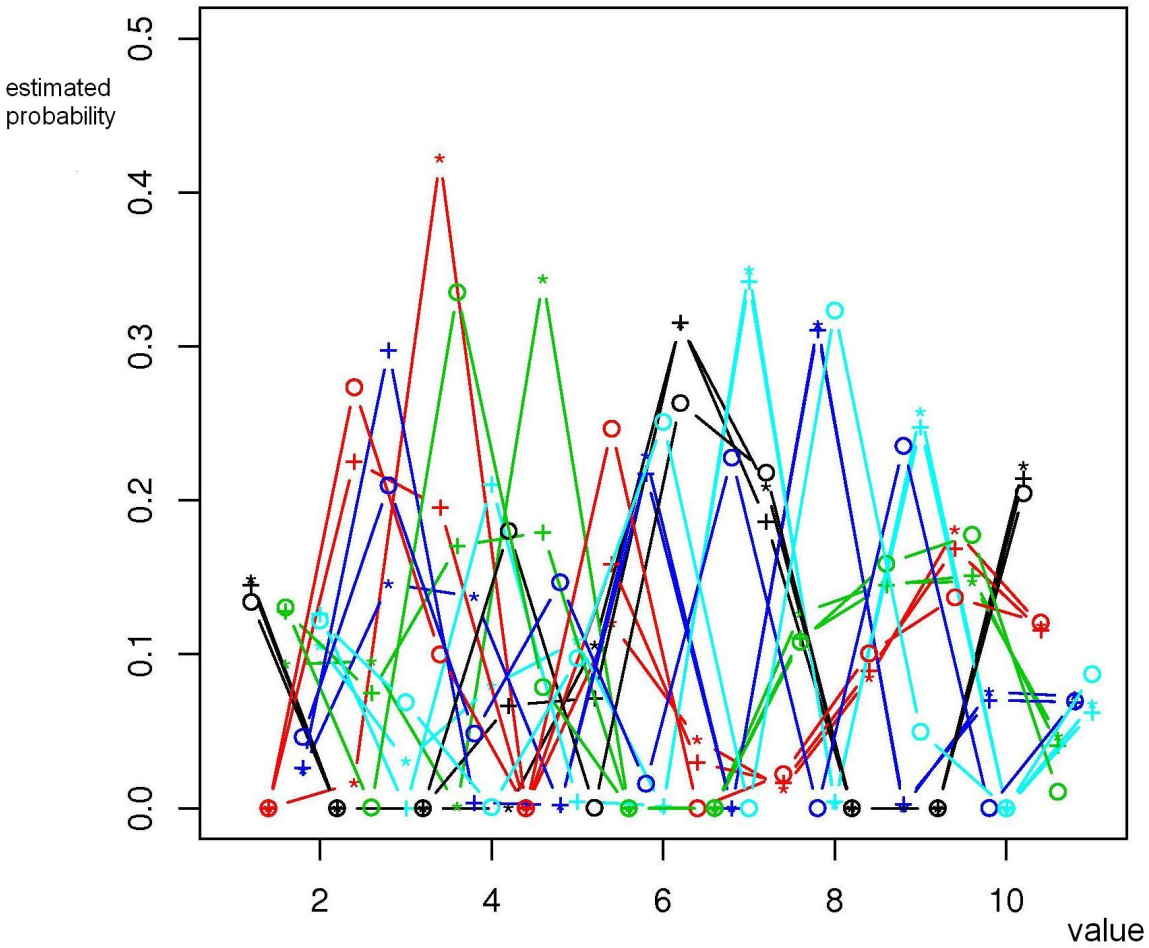
The ML estimate exists, but may not be unique (e.g. if J is too large). Local maxima may abound.

APA data

The Americal Psychological Association chose a president in 1980. There were 5 candidates, and 5738 voters gave a full ordering.

The figure shows the result of two runs of the algorithm (from different starting values), with $J = 10$.

Results for APA data



Model formulation

The Bradley-Terry model for paired comparisons postulates that in the comparison of items i and j, item i is preferred with probability

$$\frac{\lambda_i}{\lambda_i + \lambda_j},$$

where λ_i are positive parameters for each item.
Its extension to orderings of more items is just the Plackett-Luce model, where the probability of ordering $\pi = (\pi(1), \ldots, \pi(|I|)$ of a subset $I \subset \{1, \ldots, n\}$ is

$$p(\pi) = \prod_{k=1}^{|I|} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{|I|} \lambda_{\pi(j)}}.$$

Introduced by Zermelo (1929), Bradley and Terry (1952), Plackett (1975), Silverberg (1980, 1984).

Generalizations

- Home field advantage, Agresti (1990): If team i is at home, then

$$P(i \text{ beats } j) = \frac{\vartheta\lambda_i}{\vartheta\lambda_i + \lambda_j},$$

  where $\vartheta > 0$.
- Ties, Rao and Kupper (1967):

$$P(i \text{ beats } j) = \frac{\lambda_i}{\lambda_i + \vartheta\lambda_j},$$

  where $\vartheta > 1$.
- Paired comparison of teams, Huang et al. (2006): for team I, $\lambda_I = \sum_{i \in I} \lambda_i$, where λ_i are assigned to the individual players. Home field advantage and ties can also be introduced.

Methods in the literature

Numerous methods have been proposed, e.g.

- Ad hoc iterative algorithms for finding the MLE.
- MM algorithms for finding the MLE, Hunter (2004): several previous ad hoc algorithms were shown to fit in the MM scenario.
- Bayesian inference via message passing algorithms, Guiver and Snelson (2009).

Hunter also proved the convergence of the iterates $\lambda^{(t)}$ to the unique maximum likelihood estimate, under mild conditions.

MM algorithms

We briefly describe MM (minorization-maximization) algorithms.
(M) Given the current iterate $\lambda^{(t)}$, define a function $Q_t(\lambda)$, which minorizes the log-likelihood at $\lambda^{(t)}$:

$$Q_t(\lambda) \leq \ell(\lambda), \text{ with equality if } \lambda = \lambda^{(t)}.$$

(M) Define $\lambda^{(t+1)}$ to be the maximizer of $Q_t(\lambda)$.
The minorizing function should allow for an explicit maximization. Hunter uses the inequality

$$-\ln x \geq 1 - \ln y - (x/y), \quad x, y > 0$$

to define $Q_t(\lambda)$ such that the components of λ are separated.
EM algorithms are a special case of MM algorithms (e.g. Heiser, 1995).

Complete observations: exponential rv's

Lemma

Let X_i be independent exponentially distributed rv's with parameter λ_i. Then

$$P(X_{\pi(1)} < \ldots < X_{\pi(|I|)}) = \prod_{k=1}^{|I|} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{|I|} \lambda_{\pi(j)}}$$

$$P(X_i < X_j/\vartheta) = \frac{\lambda_i}{\lambda_i + \vartheta\lambda_j}$$

$$P(\min_{i\in I} X_i < \min_{j\in J} X_j) = \frac{\sum_{i\in I} \lambda_i}{\sum_{i\in I} \lambda_i + \sum_{j\in J} \lambda_j}$$

Complete observations: draws from an urn

Lemma

Suppose an urn contains n balls with weights λ_i. For a subset I, we draw from the urn with replacement until every ball $i \in I$ appears at least once. Let π denote the order in which the elements of I appeared. Then

$$p(\pi) = \prod_{k=1}^{|I|} \frac{\lambda_{\pi(k)}}{\sum_{j=k}^{|I|} \lambda_{\pi(j)}}.$$

If we draw until we first see a ball from $I \cup J$, then the probability that this ball is from I is

$$\frac{\sum_{i \in I} \lambda_i}{\sum_{i \in I} \lambda_i + \sum_{j \in J} \lambda_j}.$$

EM algorithms

According to the above lemmas, the observed permutations can be viewed as incomplete observations, and the EM algorithm can be derived straightforwardly.

- In the Plackett-Luce model and in the “teams” model, we obtain two different EM algorithms.
- In the “home team advantage” model, we get a GEM algorithm, where we update ϑ and λ separately.
- The EM algorithm also works for the “ties” model, when ϑ is known.

The algorithms have the same convergence properties as the MM algorithms of Hunter.

Comparison for the Plackett-Luce model

Suppose we have m observations, the rth being the ordering π_r of players I_r. For $i \in I_r$, denote by $\alpha_r(i)$ the rank of i in the ordering π_r. Suppose i appears in m_i orderings.
Iteration of the "exponential" EM algorithm:

$$\lambda_i^{(t+1)} = m_i \left[\sum_{r:i\in I_r} \sum_{k=1}^{\alpha_r(i)} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} \right]^{-1} \qquad 1 \le i \le n.$$

Hunter's algorithm is similar, we must subtract u_i from the numerator, and $u_i/\lambda_i^{(t)}$ from the denominator, where u_i is the number of orderings in which i is the loser.

For the “urn” EM algorithm, we have the quite different update

$$\lambda_i^{(t+1)} = m_i + \lambda_i^{(t)} \left[\sum_{r=1}^{m} \sum_{k=1}^{|I_r|} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} - \sum_{r:i \in I_r} \sum_{k=1}^{\alpha_r(i)} \frac{1}{\sum_{j=k}^{|I_r|} \lambda_{\pi_r(j)}^{(t)}} \right].$$

The results are similar in the other cases: for the “home team advantage” and “ties” models, we get algorithms similar to Hunter’s MM algorithms, and for the “teams” model, we get two EM algorithms, of which the “urn” type is similar to the MM algorithm developed by Huang, Weng and Lin (2006).

Comparing the speed of convergence

We show some numerical results for the Plackett-Luce model.

- Let $n = 5$, $m = 40$, and $\lambda^{(0)} = (0.2, \ldots, 0.2)$. The average number of steps until convergence were found to be 10 for the MM algorithm, 20 for the "exponential" EM, and around 100 for the "urn" EM.
- Let $n = 20$, $m = 200$. The quotient

$$q_t = \frac{\max_i \|\lambda_i^{(t)} - \lambda_i^{(t-1)}\|}{\max_i \|\lambda_i^{(t-1)} - \lambda_i^{(t-2)}\|}$$

can be used to measure the speed of convergence. We compared the MM and the "exponential" EM algorithms. In both cases, q_t was nearly constant (after the first few steps), with 0.16 for the MM, and 0.35 for the EM.