ELSEVIER Computational Statistics & Data Analysis 41 (2003) 645–655

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Mixtures of distance-based models for ranking data

Thomas Brendan Murphy[a,*], Donal Martin[b]

[a] *Department of Statistics, Trinity College, Dublin 2, Ireland*
[b] *Division of Statistics, 355 Kerr Hall, University of California, Davis, CA 95616, USA*

Received 1 April 2002; received in revised form 1 April 2002

Abstract

Ranking data arises when judges are asked to rank some or all of a group of objects. Examples of ranking data arise in many areas, including the Irish electoral system and the Irish college admission system. Mixture models can be used to study heterogeneous populations. The study of these populations is achieved by thinking of the population as being composed of a finite number of homogeneous sub-populations. Mixtures of distance-based models are used to analyze ranking data from heterogeneous populations. Results from simulations are included, as well as an application to the well-known American Psychological Association election data set.

Keywords: Mixture models; Ranking data; Maximum likelihood

1. Introduction

Ranking data arises when judges are asked to rank some or all of a group of objects. Some examples of ranking data arise in the Irish college admissions system where students rank the courses that they wish to take and the Irish electoral system where voters rank the candidates in order of preference. Different types of ranking data can arise depending on whether the judges are required to rank all of the objects or not. We will restrict our attention to the case where the judges rank all of the objects, this type of data is called complete ranking data.

Many models have been developed for ranking data and these are extensively discussed in Marden (1995). Many of the standard ranking models assume that we have

[*] Corresponding author.
E-mail address: murphybt@tcd.ie (T.B. Murphy).

PII: S0167-9473(02)00165-2

a homogeneous population of judges, and this assumption is frequently valid. However, in many examples like the Irish college admissions and Irish electoral system the population of judges is more likely to be heterogeneous than homogeneous.

Mixture models provide a method for modeling heterogeneous populations by assuming that the population is composed of a finite number of homogeneous sub-populations. The distribution of rankings within the sub-populations can be modeled using one of the standard models for rankings. Some previous applications of mixture models to ranking data are described in Marden (1995, Chapter 10) and references therein.

In Section 3, we propose using mixtures of distance-based models for modeling heterogeneous populations in ranking data analysis. The distance-based models for rankings have two parameters, a central ranking and a measure of precision; the probability of a ranking occurring is large for rankings close to the central ranking and is small for rankings far away from the central ranking.

The use of mixtures of distance-based models has some appealing features: the resulting fitted model is easy to interpret because the parameters that describe the centers of the components are themselves rankings, there are many ways of constraining the precision parameters which gives good modeling flexibility, and the models can be fitted quite quickly using the EM algorithm (Dempster et al., 1977).

One question of interest, when fitting mixture models, is the number of mixture components that should be used. In Section 3.3, we compare two methods of determining how many components are appropriate.

The use of mixtures of distance-based models for ranking data is demonstrated on simulated data (Section 4) and on the American Psychological Association 1980 presidential election data (Section 5).

2. Distance-based models

Distance-based models for ranking data have two parameters, a central ranking R and a precision parameter λ. The probability of a ranking r occurring is $f(r|R,\lambda) = C(\lambda)\exp[-\lambda d(r,R)]$, where $d(\cdot,\cdot)$ is a distance between rankings and $C(\lambda)$ is chosen so that the $f(r|R,\lambda)$ is a valid probability mass function.

Many distances between rankings have been proposed and these are described in Critchlow (1985), Diaconis (1988), and Marden (1995). We investigate the use of Kendall, Spearman, and Cayley distances (see Eq. (1)) between rankings.

If $r = (r_1, r_2, \ldots, r_M)$ and $s = (s_1, s_2, \ldots, s_M)$ are two rankings of M objects, where the jth number in the ranking records the rank given to object j; then we can define distances between rankings as follows:

$$\begin{aligned}
d_{\mathrm{K}}(r,s) &= \#\{(i,j)\colon i<j, (r_i - r_j)(s_i - s_j) < 0\}, \quad \text{Kendall,}\\
d_{\mathrm{S}}(r,s) &= \sqrt{\sum_{i=1}^{M}(r_i - s_i)^2}, \quad \text{Spearman,}\\
d_{\mathrm{C}}(r,s) &= M - \text{number of cycles in } s \circ r^{-1}, \quad \text{Cayley.} \qquad (1)
\end{aligned}$$

Cayley's distance $d_C(r,s)$ is the minimum number of transpositions required to transform the ranking r into s. Note that Spearman's distance is sometimes defined to be the square of the Spearman's distance shown above.

Interpretations of these (and other) distances are given in Marden (1995). The distance-based model that uses Kendall's distance is frequently called Mallow's ϕ model (Mallows, 1957).

3. Mixtures of distance-based models

3.1. Mixture models

Distance-based models can provide a reasonable model for rankings produced by a homogeneous population of judges, but we sometimes have a heterogeneous population of judges. Mixture models provide a method of developing models for heterogeneous populations by modeling the population as a collection of homogeneous sub-populations.

Suppose that a population consists of G sub-populations. Suppose further that the probability that an observation comes from sub-population g is p_g and given that the observation belongs to sub-population g it is generated from a distance-based model with central ranking R_g and precision λ_g.

Then, the model for rankings from this population is

$$f(r) = \sum_{g=1}^{G} p_g f(r|R_g, \lambda_g) = \sum_{g=1}^{G} p_g C(\lambda_g) \exp[-\lambda_g d(r, R_g)], \tag{2}$$

i.e., the model is a mixture of distance-based models. Therefore, the log-likelihood of a data set $\mathbf{r}_n = (r^{(1)}, r^{(2)}, \ldots, r^{(n)})$ consisting of n rankings is

$$l(\underline{R}, \underline{\lambda}, \underline{p}|\mathbf{r}_n) = \sum_{i=1}^{n} \log\left\{ \sum_{g=1}^{G} p_g C(\lambda_g) \exp[-\lambda_g d(r^{(i)}, R_g)] \right\}.$$

There are a few ways in which we can constrain the precision parameters in the distance-based mixture model and this gives us a large range of modeling flexibility; this is analogous to the flexibility available when using mixtures of Gaussian distributions with different constraints on the variance parameters (see, for example, Fraley and Raftery, 2000).

We consider mixtures with the following constraints on the precision parameters:

- All components have unrestricted precision parameters.
- All components, except one, have unrestricted precision parameters and one component has precision equal to zero; this forces the mixture to have a component which is of uniform distribution. The uniform (or noise) component can be used to pick up outlying "noise" rankings.
- All components have identical precision parameters.

- All components, except one, have identical precision parameters and one component has precision equal to zero; this forces one component to be a uniform distribution.

The inclusion of a noise component which is uniform on the set of possible rankings is analogous to the use of a Poisson noise term in Gaussian model-based clustering (Fraley and Raftery, 1998, 2000).

3.2. Fitting mixture models

The mixtures of distance-based models were fitted by maximum likelihood using the EM algorithm. In order to implement the EM algorithm we introduce latent variables z which record the component membership of each observation. The latent (membership) variable $z = (z_1, z_2, \ldots, z_G)$ is defined such that $z_g = 1$ if the observation belongs to component g and $z_g = 0$ otherwise. We write $\mathbf{z}_n = (z^{(1)}, z^{(2)}, \ldots, z^{(n)})$ for the latent variables for all of the observations. Hence, the complete-data log-likelihood is

$$l_C(\underline{R}, \underline{\lambda}, \underline{p} | \mathbf{r}_n, \mathbf{z}_n) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_g^{(i)} [\log p_g + \log C(\lambda_g) - \lambda_g d(r^{(i)}, R_g)].$$

The EM algorithm is easy to implement for distance-based mixture models and it involves the following two steps:

E-*Step*: Compute the values of $\hat{z}_g^{(i)}$ as follows:

$$\hat{z}_g^{(i)} = \frac{p_g f(r^{(i)} | R_g, \lambda_g)}{\sum_{g'=1}^{G} p_{g'} f(r^{(i)} | R_{g'}, \lambda_{g'})}.$$

M-*Step*:

- Compute the values of R_g as follows:

$$R_g = \underset{R}{\operatorname{argmin}} \sum_{i=1}^{n} \hat{z}_g^{(i)} d(r^{(i)}, R).$$

- Compute the p_g values as follows: $p_g = \sum_{i=1}^{n} \hat{z}_g^{(i)} / n$.
- For the clusters with unrestricted λ_g values, let λ_g be the solution of

$$\sum_{r} d(r, R_g) f(r | R_g, \lambda_g) = \frac{\sum_{i=1}^{n} \hat{z}_g^{(i)} d(r^{(i)}, R_g)}{\sum_{i=1}^{n} \hat{z}_g^{(i)}},$$

  where the left-hand side summation is taken over all possible rankings r.
  For the clusters which have identical $\lambda_g = \lambda$ values, let λ be the solution of

$$\sum_{r} d(r, R) f(r | R, \lambda) = \frac{\sum_g \sum_{i=1}^{n} \hat{z}_g^{(i)} d(r^{(i)}, R_g)}{\sum_g \sum_{i=1}^{n} \hat{z}_g^{(i)}},$$

  where we sum over those g for which clusters are restricted to have equal precision. Note that, the value of the LHS (above) is independent of the choice of R.

The EM algorithm for the mixture of distance-based models converged very quickly, but can easily get stuck in local maxima. To alleviate this problem the EM algorithm was run from 30 randomly chosen starting points and the best solution was recorded. The use of the SEM algorithm (Celeux and Govaert, 1992) as a method of avoiding local maxima was also investigated; this algorithm replaces the E-step with a stochastic classification step.

3.3. Model choice

The problem of choosing which mixture model to use can be a difficult one. For the current problem we have to decide:

- Which distance should be used?
- How do we constrain the precision parameters?
- How many components are in the mixture?

We use two different criteria for choosing which model is appropriate. These are the Bayesian information criterion (BIC) (Fraley and Raftery, 1998) and integrated complete likelihood (ICL) (Biernacki et al., 2000). A review of many methods for choosing the appropriate number of components is given in McLachlan and Peel (2000).

The BIC provides an approximation to the Bayes factor for model selection; it involves the maximized log-likelihood minus a penalty term.

$$\mathrm{BIC}(G) = 2l(\hat{\theta}_G) - v(G)\log(n),$$

where $v(G)$ is the number of free parameters and $l(\hat{\theta}_G)$ is the maximized log-likelihood for the G component mixture. When using this criterion we choose the model with the largest BIC value. This criterion tends to produce models that fit the data very well.

The ICL is a classification variant of the BIC. This criterion is used to find a mixture of well-separated components that fits the data well; this makes it a good criterion for clustering applications. When using this criterion we choose the model with the highest ICL value.

$$\mathrm{ICL}(G) = \mathrm{BIC}(G) + 2\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_g^{(i)}\log\hat{z}_g^{(i)} = \mathrm{BIC}(G) - 2\sum_{i=1}^{n}\mathrm{Entropy}(\hat{z}^{(i)}),$$

where the $\hat{z}$ values are the latent membership variables generated in the E-step of the EM algorithm; the extra entropy term in the ICL criterion penalizes for cases where the membership of an observation is uncertain.

4. Simulation results

Artificial ranking data sets were generated using the 14 models described in Table 1. For each model 10 data sets (five with 200 observations and five with 2000 observations) were generated and the various mixture models (as described in Section 3.1) were fitted using maximum likelihood. In each case, the best mixture model was chosen using the BIC and ICL criteria.

The results of the study (Table 2) show that the two-model selection criteria give quite different results. The simulation results indicate that the ICL criterion tends to

Table 1
Parameters for the mixture models used in the simulation study

Model	Distance	Center					Precision	Proportion
1	Kendall	1	2	3	4	5	1	0.5
	Kendall	5	4	3	2	1	1	0.5
2	Spearman	1	2	3	4	5	1	0.5
	Spearman	5	4	3	2	1	0.5	0.5
3	Cayley	1	2	3	4	5	1	0.9
	Cayley	5	4	3	2	1	1	0.1
4	Kendall	1	2	3	4	5	1	0.9
	Kendall	5	4	3	2	1	0.5	0.1
5	Spearman	1	2	3	4	5	0.5	0.9
	Spearman	5	4	3	2	1	1	0.1
6	Cayley	1	2	3	4	5	1	0.5
	Cayley	1	2	5	4	3	1	0.5
7	Kendall	1	2	3	4	5	1	0.5
	Kendall	1	2	5	4	3	0.5	0.5
8	Spearman	1	2	3	4	5	1	0.9
	Spearman	1	2	5	4	3	1	0.1
9	Cayley	1	2	3	4	5	1	0.9
	Cayley	1	2	5	4	3	0.5	0.1
10	Kendall	1	2	3	4	5	0.5	0.9
	Kendall	1	2	5	4	3	0.1	0.1
11	Spearman	1	2	3	4	5	1	1.0
12	Cayley	1	2	3	4	5	3	1.0
13	Kendall	1	2	3	4	5	0.3	1.0
14	Uniform	1	2	3	4	5	0	1.0

underestimate the number of components and the BIC criterion tends to provide a more accurate estimate of the number of components but has a tendency to overestimate the number of components.

One possible reason that the ICL criterion may underestimate the number of components is that the range of possible values of the distances between rankings is quite small. The small range of distance values means that the probability of component membership values tends to be large for more than one component of the mixture and

Table 2
Frequencies that each type of mixture was selected using the BIC and ICL criteria for simulated data

Model number	Sample size	BIC Mixture components						ICL Mixture components					
		$0+N$	1	$1+N$	2	$2+N$	3	$0+N$	1	$1+N$	2	$2+N$	3
1	200				4	1		1	3		1		
	2000				5				4	1			
2	200			5				4		1			
	2000	4				1		5					
3	200		2	3					5				
	2000			5					5				
4	200		1	4					4		1		
	2000			4			1		4				1
5	200	3		2				5					
	2000			4		1		4	1				
6	200				5				5				
	2000				5				5				
7	200			4	1			2	2	1			
	2000				2	2	1		5				
8	200		4	1					4	1			
	2000			4		1			4	1			
9	200		1	4					5				
	2000			5					5				
10	200			5				3	1	1			
	2000			4	1			4	1				
11	200		4	1					4	1			
	2000			5					4	1			
12	200			5						5			
	2000			5						5			
13	200	4		1				5					
	2000			5				4	1				
14	200	5						5					
	2000	5						5					

We show that if one of the components in the selected mixture was a noise component, for example, the $2+N$ column represents a mixture with two components plus a noise term.

this leads to a large value for the Entropy($\hat{z}^{(i)}$) term. Therefore, the entropy penalty provides a large penalty for almost all of the fitted mixtures. That said, the ICL has been developed to find well-separated clusters and not necessarily to determine the correct number of components.

Although the BIC criterion tends to get the number of components correct it has a tendency to include a noise component instead of an extra non-uniform component. This may be because a noise component requires no center or precision parameters to be estimated, while a non-uniform component involves parameter estimation which is penalized by the BIC criterion. The BIC criterion performs better when selecting amongst those mixtures without noise components (Table 3).

5. Application: APA election data

The 1980 American Psychological Association (APA) presidential election had five candidates. In this election, voters were asked to rank the candidates in order of preference. A total of 15449 votes were cast in the election and of these 5738 ranked all five candidates. The data set which records the 5738 complete votes is available in Diaconis (1988) and Hand et al. (1994).

Mixtures of distance-based models, with up to 10 components, were fitted to data. Mixtures using the three distances (Section 2) and the various restrictions on the precision parameters (Section 3.1) were fitted.

When using BIC as a model selection criterion, the best mixture was found to be a five component Cayley-based mixture with unrestricted precision parameters (Table 4).

Including a noise component in the model was not required for this data. Interestingly, in the best mixture model, all the three smallest components all put candidate B in the first place; it may be that these three components are trying to model a sub-population for which the Cayley-based model provides a poor fit. Also, the largest component is also the one with least precision, this feature occurred amongst many of the mixtures chosen using BIC.

The model chosen when using the ICL criterion is a single component Cayley-based model (Table 5). Therefore, even though the BIC criterion suggests that there are five components in the population the ICL criterion indicates that there are no well-separated components in the population.

In almost all of the cases, the models involving Cayley's distance are superior, in terms of BIC values, to those using the other distances. When the ICL criterion is used to compare mixtures with the same number of components it almost always selected Kendall-based mixtures. The value of Kendall's distance is strictly greater than Cayley's distance, so the distances between component centers tend to be larger, leading to well-separated centers and a higher ICL value.

It is worth noting that candidates A and C are research psychologists, candidates D and E are clinical psychologists and candidate B is a community psychologist. The components of the fitted mixtures indicate that the voters tend to give high rankings to candidates within a sub-category of psychologists. The mixture components also

Table 3
Frequencies that each type of mixture was selected using the BIC for simulated data when only models without noise components are considered

Model number	Sample size	Components		
		1	2	3
1	200		4	1
	2000		5	
2	200	1	4	
	2000	5		
3	200	5		
	2000		5	
4	200	4	1	
	2000		5	
5	200		5	
	2000		4	1
6	200		5	
	2000		5	
7	200	2	3	
	2000		2	3
8	200	4	1	
	2000		5	
9	200	5		
	2000		5	
10	200	1	4	
	2000		5	
11	200	4	1	
	2000	2	3	
12	200		2	
	2000		5	
13	200	3	2	
	2000		5	
14	200	5		
	2000	5		

Table 4
Parameters of best mixture model selected using BIC; the best model is a mixture of Cayley-based models with unrestricted precision parameters

Cayley-based mixture						
Center					Precision	Proportion
A	B	C	D	E		
5	2	4	1	3	0.16	0.42
4	5	1	2	3	0.79	0.31
3	1	2	4	5	1.52	0.12
3	1	2	5	4	1.81	0.08
3	1	5	2	4	1.72	0.07

Table 5
Parameters of best mixture model selected using ICL; this model consists of a single Cayley-based model component

Cayley-based mixture						
Center					Precision	Proportion
A	B	C	D	E		
3	1	2	5	4	0.25	1.00

give insight, for example, into how people who support candidate B, a community psychologist, give their lower preferences to the other candidates.

6. Conclusions

Mixtures of distance-based models provide a relatively straightforward method of modeling heterogeneous populations of judges who are ranking. They provide models with good modeling flexibility which are easy to interpret and that are easy to fit.

The selection of the best mixture amongst the many possible mixtures can be done using BIC, and the results achieved using this criterion are reasonable. The BIC criterion has some problems when mixtures involving a noise component are considered, and no noise component exists in the population. Therefore, the noise component should not be included in the mixture unless it is considered to be necessary.

The selection of mixtures using ICL is more problematic for the ranking analysis problems shown. We would expect its performance to improve in cases where the number of objects being ranked is large because it is easier for clusters to be well separated in these cases.

The methods can be extended to other types of ranking data, including partial ranking data and ranking with ties using the methods of extending distances to these data types as described in Critchlow (1985).

Mixtures of other standard models for rankings could be considered. Initial investigations suggest that some of these models are more difficult to fit because the EM algorithm can converge very slowly and also because the likelihood function may be unbounded. That said, for some data sets, these mixtures can provide a better fit than the distance-based mixtures.

Acknowledgements

I would like to thank the anonymous referees and the associate editor for comments which greatly improved this paper.

References

Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. 22 (7), 719–725.

Celeux, G., Govaert, G., 1992. A classification EM algorithm for clustering and two stochastic versions. Comput. Statist. Data Anal. 14 (3), 315–332.

Critchlow, D.E., 1985. Metric Methods for Analyzing Partially Ranked Data. . Lecture Notes in Statistics, Vol. 34. Springer, Berlin.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39 (1), 1–38.

Diaconis, P., 1988. Group Representations in Probability and Statistics. Institute of Mathematical Statistics, Hayward, CA.

Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method?—answers via model-based cluster analysis. Comput. J. 41 (8), 578–588.

Fraley, C., Raftery, A.E., 2000. Model-based clustering, discriminant analysis, and density estimation. Technical Report 380, Department of Statistics, University of Washington.

Hand, D.J., Daly, F., McConway, K., Lunn, D., Ostrowski, E., 1994. Handbook of Small Data Sets. Chapman & Hall/CRC Press, Boca Raton, FL.

Mallows, C.L., 1957. Non-null ranking models. I. Biometrika 44 (1/2), 114–130.

Marden, J.I., 1995. Analyzing and Modeling Rank Data. Chapman & Hall, London.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.