J. R. Statist. Soc. B (2002)
64, *Part* 3, *pp.* 433–445

Conjugacy class prior distributions on metric-based ranking models

Jayanti Gupta

Merck Research Laboratories, Rahway, USA

and Paul Damien

University of Michigan Business School, Ann Arbor, USA

[Received May 2001. Revised February 2002]

Summary. A new class of prior distributions for metric-based models in the analysis of fully and partially ranked data is developed. This class is attractive because it provides a meaningful way to encapsulate prior information about the parameters of the model. Three examples illustrate the ideas developed in the paper.

Keywords: Equivalence class; Mallows model; Symmetric group

1. Introduction

In the context of ranked data, given a set of n items, a full ranking of these items is simply an ordering of all these items by choice or preference. Any such ranking forms an element π of the permutation group S_n such that $\pi(i)$ is the rank given to item i and $\pi^{-1}(i)$ is the item assigned the rank i.

Given a set of n items, a partial ranking of k out of these n items is a ranking where only the first k choices are specified. A partial ranking of this type forms an element of the coset space S_n/S_{n-k}, where S_{n-k} is the subgroup of S_n consisting of all permutations which leave the first k integers fixed:

$$S_{n-k} = \{\pi \in S_n : \pi(i) = i, 1 \leqslant i \leqslant k\}.$$

Historically, models for random rankings grew out of the literature on paired comparisons. For example, the Thurstone (1927) model specifies that item i would be preferred to item j if $X_i > X_j$ where the Xs are independent and identically distributed normal random variables with different means and equal variance. Mosteller (1951) provided simple forms for the least square estimators of the means of the Xs under this model. MacKay and Chaiy (1982) used Monte Carlo methods to compare estimators of the means of Xs in the above model with those under the unequal variance model. These paired comparison models were extended to rankings by letting π_i equal the rank of X_i.

Mallows (1957) also started with models for paired comparisons and used a conditional argument to extend these to models for rankings. His two-parameter models are unimodal with the probability of a ranking π decreasing as the distance in a certain metric between π and the mode

Address for correspondence: Paul Damien, University of Michigan Business School, 701 Tappan Street, Ann Arbor, MI 48109-1234, USA.
E-mail: pdamien@umich.edu

 1369–7412/02/64433

increases. These models were popularized by Feigin and Cohen (1978) and Schulman (1979) who provided tables for their use. Other models for random rankings have been introduced by Luce (1959), Plackett (1975), Fienberg and Larantz (1976), Henery (1981), Berry (1979) and Tallis and Dansie (1983). Gordon (1979) introduced a model based on Ulam's distance, whereas Fligner and Verducci (1986, 1988) investigated Cayley's distance and Kendall's τ-distance. Fligner and Verducci (1990) did a Bayesian analysis of the generalized Mallows model by introducing prior distributions on the parameters of the model. Diaconis (1988) developed a second-order analysis for ranked data.

In this paper, we discuss the notion of conjugacy classes on the space of permutations and use it to define a new class of prior distributions on this space. It is called the conjugacy class prior and it uses properties of the permutation group and the notion of metrics on conjugacy classes to define the prior distributions. We focus on the Mallows model throughout the paper purely for illustration, noting that the ideas developed here—with minor modifications—can be used with other models as well. We use the Gibbs sampling algorithm to generate random variates from the distributions of the parameters of the Mallows model, leading to a full Bayesian analysis using this prior.

The paper is organized in the following manner. Section 2 describes the Mallows model and defines some of the different metrics on fully ranked data. Section 3 discusses the notion of conjugacy classes of the permutation group and describes the conjugacy class prior distribution for the Mallows model. Metric-based priors for partially ranked data are developed in Section 4. Section 5 illustrates these priors via examples.

2. Models and metrics

Mallows (1957) proposed a non-null probability model, i.e. a model distinct from the uniform model (the model where all $k!$ possible rankings of the k items are equally likely) for fully ranked data. The model specifies a particular ranking $\pi_0 \in S_k$, the permutation group on k objects, which can be interpreted as the most likely or the modal ranking of the k items, and states that the probability of any other ranking π decreases exponentially according to the distance from π to π_0. So, $P(\pi) = K(\lambda) \exp\{-\lambda\, d(\pi, \pi_0)\}$, for all $\pi \in S_k$, $d(\cdot, \cdot)$ is a metric on S_k, π_0 is the location parameter and $\lambda \geqslant 0$ is a dispersion parameter. The normalizing constant $K(\lambda)$ is defined as

$$K(\lambda)^{-1} = \sum_{\pi \in S_k} \exp\{-\lambda\, d(\pi, \pi_0)\}$$

and is independent of the choice of π_0. The model is centred about the ranking π_0, and as λ increases the distribution becomes increasingly peaked about π_0.

2.1. Some metrics on fully ranked data

By a suitable choice of a metric on S_k, some well-known measures of association of two permutations have been obtained; see, for example, Diaconis (1987). These are

$$R(\pi, \sigma) = \left[\sum_{i=1}^{k} \{\pi(i) - \sigma(i)\}^2\right]^{1/2},$$

Spearman's ρ-distance,

$$F(\pi, \sigma) = \sum_{i=1}^{k} |\pi(i) - \sigma(i)|,$$

Spearman's footrule, $T(\pi, \sigma)$, the number of pairs of items (i, j) such that $\pi(i) < \pi(j)$ and $\sigma(i) > \sigma(j)$ is Kendall's τ, $H(\pi, \sigma) = \#\{i = 1, \ldots, k : \pi(i) \neq \sigma(i)\}$, Hamming distance, $U(\pi, \sigma) = k$, the maximal number of items ranked in the same order by π and σ is Ulam's distance, and $C(\pi, \sigma)$, Cayley's distance, the minimum number of transpositions needed to transform π into σ.

Since the labels $1, \ldots, k$ are arbitrarily assigned to the items being ranked, it is natural to insist that the distances between rankings do not depend on the labelling. Mathematically, this implies that the metric should be right invariant: $d(\pi, \sigma) = d(\pi\tau, \sigma\tau)$, for all $\pi, \sigma, \tau \in S_k$. All these six metrics are right invariant, i.e. they remain unchanged under arbitrary relabelling of the items. Some of these metrics will be used to exemplify the prior distributions that are developed in this paper.

2.2. Conjugacy classes of the symmetric group

We collect together some relevant facts from algebra; see, for example, Diaconis (1987). The following facts are central to the ideas in the rest of the paper.

Definition 1. For two permutations, $\pi_1, \pi_2 \in S_k$, π_1 is said to be a conjugate of π_2 in S_k (or $\pi_1 \sim \pi_2$), if there is an element σ in S_k such that $\pi_1 = \sigma\pi_2\sigma^{-1}$.

Conjugacy is an equivalence relationship on S_k and so splits the group into equivalence classes, called conjugacy classes.

Definition 2. Given an integer k, we say that the sequence of positive integers $k_1 \leqslant \ldots \leqslant k_r$ constitutes a partition of k if $k = k_1 + \ldots + k_r$.

Definition 3. The set of integers $(i_1, \ldots, i_r)$ is said to be a cycle of the permutation $\pi \in S_k$, if π sends i_1 into i_2, i_2 into $i_3, \ldots, i_{r-1}$ into i_r and i_r into i_1, and leaves all other items fixed.

Definition 4. A permutation $\pi \in S_k$ has the cycle decomposition $\{k_1, \ldots, k_r\}$ if it can be written as the product of disjoint cycles of lengths $k_1, \ldots, k_r$, $k_1 \leqslant \ldots \leqslant k_r$.

For example, in S_9,

$$\pi = \begin{pmatrix} 123456789 \\ 132564798 \end{pmatrix} = (1)(2, 3)(4, 5, 6)(7)(8, 9)$$

has cycle decomposition $\{1, 1, 2, 2, 3\}$ and $1 + 1 + 2 + 2 + 3 = 9$.

Let $p(k)$ denote the number of partitions of k. Each time that we break a given permutation in S_k into a product of disjoint cycles, we obtain a partition of k; for if the cycles appearing have lengths $k_1, \ldots, k_r$ respectively, $k_1 \leqslant \ldots \leqslant k_r$, then $k = k_1 + \ldots + k_r$.

A well-known result in algebra states that two permutations in S_k are conjugate if and only if they have the same cycle decomposition. The reason is the following formula for computing the conjugate: if π sends $i \to j$ and σ sends $i \to s$ and $j \to t$, then $\sigma\pi\sigma^{-1}$ sends $s \to t$. Every symbol in π is replaced by its image in σ to compute $\sigma\pi\sigma^{-1}$. So, if π written in cycle notation is $(a \ldots b)(c \ldots d) \ldots (e \ldots f)$, then

$$\sigma\pi\sigma^{-1} = (\sigma(a) \ldots \sigma(b))(\sigma(c) \ \ldots \ \sigma(d)) \ \ldots \ (\sigma(e) \ \ldots \ \sigma(f)).$$

This results in a one-to-one correspondence between the conjugacy classes of S_k and the partitions of k. This correspondence will be used to develop a prior distribution on rankings based on the conjugacy classes to which they belong.

3. Prior distribution on S_k

Before delving into the mathematical details of the actual construction of the prior distribution, we would first like to motivate the proposed prior.

3.1. Motivating the prior distribution

The Mallows model

$$P(\pi) = K(\lambda) \exp\{-\lambda\, d(\pi, \pi_0)\}$$

has two parameters: π_0 the location parameter and $\lambda \geqslant 0$ the scale parameter. It is well known (Serre (1977), pages 32–33) that a natural choice for a measure on a topological group is the Haar measure of that group. However, for the finite permutation group S_k, the Haar measure on this group with the discrete topology is simply the uniform distribution on the group.

In the Mallows model, a prior density for the scale parameter λ could be exponential, i.e. $P(\lambda) = \alpha_0 \exp(-\alpha_0 \lambda),\ \lambda \in \mathcal{R}^+$. Interest here, however, is on developing a prior distribution for the modal ranking π_0.

Our approach to constructing a prior distribution for π_0 must have two features:

(a) it must exploit the structure of the permutation group in which the data are observed to encapsulate prior information about the random quantity π_0;
(b) it must give the prior distribution a 'sensible' interpretation.

Since the notion of conjugacy classes is central in the study of permutation groups, we wish to exploit this feature of the group to construct a prior. Thus the prior distribution on π_0 will be taken to be constant on conjugacy classes. Is this sensible? We think so because we are assigning equal prior probabilities to all permutations that permute an equal number of items and leave the remainder unchanged.

A simple example illustrates this. The group S_4 has 24 elements. Since four can be partitioned in five ways, there are five conjugacy classes in S_4. These classes are listed in Table 1 along with the number of elements in each class.

In this example, the first conjugacy class consists of the identity element in which the items were ranked in the same order in which they were observed, i.e. the item that was observed first received the first rank, the item observed second was ranked second, and so on.

The second conjugacy class, $\{(1)(1)(2)\}$, consists of those permutations in which two of the items were assigned the same rank as the order in which they were observed, corresponding to the two 1s in this partition, whereas the remaining two had their ranks interchanged, corresponding to the 2 in the partition. For example, in the permutation (4231) which

Table 1. Conjugacy classes of S_4

Partition	*Conjugacy class*	*Number of elements*
(1,1,1,1)	{(1)(1)(1)(1)}	1
(1,1,2)	{(1)(1)(2)}	6
(1,3)	{(1)(3)}	8
(2,2)	{(2)(2)}	3
(4)	{(4)}	6

belongs to this class, the items that were observed second and third were assigned ranks 2 and 3 respectively, whereas the items appearing first and fourth received ranks 4 and 1 respectively.

In addition to this feature shared by elements within a conjugacy class, these classes are also invariant to a relabelling of the items, i.e., if all the items were observed in a different order, the conjugacy classes would remain unchanged. This motivates us to assign equal prior probabilities to all the rankings in the same conjugacy class.

For a moderately large value of k, say 10, S_k is enormous (of the order of 3 million). Conjugacy classes thus form a useful way of developing prior distributions on the group by dividing it into smaller classes and constructing the prior on these classes.

Yet another interesting fact about considering prior distributions using conjugacy classes as the support space stems from considering the 'metric' feature in the Mallows model stated earlier. It is clear from the model that the probability of a ranking π from the modal ranking π_0 decreases exponentially on the basis of the distance from π to π_0.

In the development of the prior distribution to be described later this property is preserved when one constructs the prior on the space of conjugacy classes. We next investigate the properties of metrics on conjugacy classes that are required in developing the prior.

3.2. Metrics on conjugacy classes

For any set X, endowed with a bounded metric d, the induced Hausdorff distance d^* between any two closed non-empty subsets of X is well defined and satisfies the axioms for a metric on such subsets of X (Kuratowski (1966), pages 214–215). Since conjugacy classes are subsets of S_k, each of the metrics on S_k can be extended to compute the induced Hausdorff metrics on conjugacy classes given by

$$d^*(C_\pi, C_\sigma) = \max(\max_{\beta \in C_\sigma}[\min_{\alpha \in C_\pi}\{d(\alpha, \beta)\}],\ \max_{\alpha \in C_\pi}[\min_{\beta \in C_\sigma}\{d(\alpha, \beta)\}]),$$

where C_π and C_σ are the conjugacy classes of π and σ respectively. The induced metric d^* is right invariant if d is.

Now, from a computational perspective, having a simpler form to compute the distances could be quite useful. Also, any metric can be made invariant by averaging it (Diaconis (1987), pages 114–115). Hence theorem 1 stated below can be used to compute induced Hausdorff distances for bivariant metrics, or metrics that have both right and left invariance.

Theorem 1. If a metric d on S_k is bivariant, then its induced Hausdorff metric d^* on the conjugacy classes may be computed according to the simpler formula

$$d^*(C_\pi, C_\sigma) = \min_{\substack{\alpha \in C_\pi \\ \beta \in C_\sigma}}\{d(\alpha, \beta)\} = \min_{\beta \in C_\sigma}\{d(\pi, \beta)\}\,.$$

Proof. We invoke properties of conjugacy classes to prove the theorem.

$$\max_{\beta \in C_\sigma}[\min_{\alpha \in C_\pi}\{d(\alpha, \beta)\}] = \max_{\tau_2 \in S_k}[\min_{\tau_1 \in S_k}\{d(\tau_1 \pi \tau_1^{-1}, \tau_2 \sigma \tau_2^{-1})\}] \qquad \text{by definition of } C_\pi, C_\sigma$$

$$
\begin{aligned}
&= \max_{\tau_2 \in S_k}[\min_{\tau_1 \in S_k}\{d(\tau_1\pi, \tau_2\sigma\tau_2^{-1}\tau_1)\}] && \text{by right invariance}\\
&= \max_{\tau_2 \in S_k}[\min_{\tau_1 \in S_k}\{d(\pi, \tau_1^{-1}\tau_2\sigma\tau_2^{-1}\tau_1)\}] && \text{by left invariance}\\
&= \max_{\tau_2 \in S_k}[\min_{\tau_3 \in S_k}\{d(\pi, \tau_3\sigma\tau_3^{-1})\}] && \text{where } \tau_3 = \tau_1^{-1}\tau_2\\
&= \min_{\tau_3 \in S_k}\{d(\pi, \tau_3\sigma\tau_3^{-1})\}\\
&= \min_{\beta \in C_\sigma}\{d(\pi, \beta)\}\,.
\end{aligned}
$$

By the same argument as above,

$$\max_{\alpha \in C_\pi}[\min_{\beta \in C_\sigma}\{d(\alpha, \beta)\}] = \min_{\beta \in C_\sigma}\{d(\pi, \beta)\}\,.$$

Hence,

$$d^*(C_\pi, C_\sigma) = \min_{\beta \in C_\sigma}\{d(\pi, \beta)\}\,.$$

Also,

$$
\begin{aligned}
\min_{\beta \in C_\sigma}\{d(\pi, \beta)\} &= \min_{\tau \in S_k}\{d(\pi, \tau\sigma\tau^{-1})\} && \text{by definition of } C_\sigma\\
&= \min_{\tau_1, \tau_2 \in S_k}[d\{\pi, \tau_1^{-1}\tau_2\sigma(\tau_1^{-1}\tau_2)^{-1}\}]\\
&= \min_{\tau_1, \tau_2 \in S_k}\{d(\tau_1\pi\tau_1^{-1}, \tau_2\sigma\tau_2^{-1})\}\\
&= \min_{\substack{\alpha \in C_\pi\\ \beta \in C_\sigma}}\{d(\alpha, \beta)\}\,.
\end{aligned}
$$

□

Only two of the six metrics defined in Section 2.2 are bivariant: Hamming's and Cayley's distances. We shall later use the above simplified versions of these two metrics to illustrate our methods.

3.3. Prior distributions via metrics on conjugacy classes

On the basis of the development in the last two sections, we are ready to define a class of prior distributions on the symmetric group. Let us specify a ranking π^* which we believe *a priori* to be the modal ranking. For any $\pi \in S_k$ let C_π be the conjugacy class containing π.

The prior distribution on the modal ranking is

$$P(\pi) = K^*(\lambda^*)\,\exp\{-\lambda^* d^*(C_\pi, C_{\pi^*})\}, \qquad \lambda^* \geqslant 0,$$

where λ^* is a scalar that determines how peaked the distribution is around C_{π^*}. The choice $\lambda^* = 0$ corresponds to a uniform prior on all rankings, and larger values reflect stronger beliefs in π^*. The normalizing constant $K^*(\lambda^*)$, as before, is independent of the choice of π^*.

As discussed before, this prior assigns equal probability to all permutations within a conjugacy class. It is to be noted that, although two permutations π_1 and π_2 in a class are not necessarily 'close' with respect to any of the metrics $d(\cdot, \cdot)$ discussed in Section 2.1, when each of these metrics induce the corresponding Hausdorff metric on the space of conjugacy classes, then $d^*(C_{\pi_1}, C_{\pi_2}) = 0$. Although this implies that other rankings in the modal conjugacy class, for which we may not have prior information, receive the same prior weight, we feel that the

advantage that is gained by reducing the (possibly very large) group into fewer conjugacy classes in constructing the prior far outweighs this potential drawback, particularly in the light of the desirable properties resulting from this construction; these properties are discussed a little later.

We express our prior belief about the population of rankings through our choice of π^*. The nature of the prior helps us to incorporate this belief to develop a prior distribution on the entire group by using the conjugacy class C_{π^*} of π^*.

Suppose, however, as one referee suggested, that we wish to express equally strong prior beliefs on two different rankings, π_1^* and π_2^*. If $\pi_1^* \in C_{\pi_2^*}$ then this prior is still suitable. If $\pi_1^* \notin C_{\pi_2^*}$ our prior can be modified by defining a new modal class composed of $C_{\pi_1^*}$ and $C_{\pi_2^*}$. This would, of course, change the prior distribution on the other conjugacy classes, but the new distance metric between the modal class and the other conjugacy classes can be easily computed using the formula for $d^*(\cdot, \cdot)$ in Section 3.2.

Suppose that a group of k items is being ranked by n people. Let $\pi_1, \ldots, \pi_{k!}$ be the set of all possible rankings of these k items.

Let the data be $\sigma = (\sigma_1, \ldots, \sigma_n)$, the rankings of these items by the n people. Assume the Mallows model. The joint likelihood is

$$P(\sigma_1, \ldots, \sigma_n | \pi, \lambda) = K(\lambda)^n \exp\left\{-\lambda \sum_{i=1}^{n} d(\sigma_i, \pi)\right\}.$$

Assigning λ an exponential prior, the prior joint distribution of the parameters is

$$(\pi, \lambda) = P(\pi)\, P(\lambda) = \frac{\exp\{-\lambda^* d^*(C_\pi, C_{\pi^*})\}}{\sum_{i=1}^{k!} \exp\{-\lambda^* d^*(C_{\pi_i}, C_{\pi^*})\}} \alpha_0 \exp(-\lambda \alpha_0).$$

Clearly, a closed form solution to the posterior (obtained by multiplying the likelihood and the prior) is impossible. However, a Gibbs sampling algorithm to simulate from the full conditional distributions of the location and scale parameters for the above model is easy to implement. The full conditional distributions are

$$\begin{aligned} P(\lambda | \sigma, \pi) &= \frac{P(\sigma | \pi, \lambda)\, P(\lambda)}{\int_0^\infty P(\sigma | \pi, \lambda')\, P(\lambda')\, \mathrm{d}(\lambda')} \\ &\propto K(\lambda)^n \exp\left[-\lambda \left\{\alpha_0 + \sum_{i=1}^{n} d(\sigma_i, \pi)\right\}\right], \\ P(\pi | \sigma, \lambda) &= \frac{P(\sigma | \pi, \lambda)\, P(\pi)}{\sum_{\pi_j \in S_k} P(\sigma | \pi_j, \lambda)\, P(\pi_j)} \\ &= \frac{\exp\left[-\left\{\lambda \sum_{i=1}^{n} d(\sigma_i, \pi) + \lambda^* d^*(C_\pi, C_{\pi^*})\right\}\right]}{\sum_{j=1}^{k!} \exp\left[-\left\{\lambda \sum_{i=1}^{n} d(\sigma_i, \pi_j) + \lambda^* d^*(C_{\pi_j}, C_{\pi^*})\right\}\right]}. \end{aligned}$$

For the examples in Section 5, the acceptance–rejection algorithm is used to sample from the full conditional density of λ with an exponential dominating density. To sample π, a Metropolis–Hastings algorithm is used to define a Markov chain with the above full conditional as its target density.

Even with a uniform prior on π (by choosing $\lambda^* = 0$), the form of the full conditional for π is such that, even for moderately small values of λ, the denominator in the last expression could be much smaller than 1, for a reasonably large data set. This could inflate the posterior probability of the modal ranking in the observed data much beyond its observed proportion. This drawback of the Mallows model can be overcome if we can make a more informed choice for the prior parameters π^* and λ^*.

From a perspective of inference, some interesting questions must be addressed. Is there a relationship between λ, the scale parameter in the model, and λ^*, the scaling factor in the prior distribution, that would determine which ranking receives the highest posterior distribution? In the following, the proofs of the corollaries are omitted since they are straightforward.

Theorem 2. Let $D(\pi_j) = \Sigma_{i=1}^n d(\sigma_i, \pi_j)$, the sum of the distances of the observed rankings from π_j, and let $D^*(\pi_j) = d^*(C_{\pi_j}, C_{\pi^*})$, the induced Hausdorff distance between C_{π_j} and C_{π^*}. For $\pi_1, \pi_2 \in S_k$, and given $\lambda, \lambda^* > 0$, π_1 will have a higher posterior probability than π_2 if and only if

$$D(\pi_1) - D(\pi_2) < \gamma\{D^*(\pi_2) - D^*(\pi_1)\}$$

where $\gamma = \lambda^*/\lambda$.

Proof.

$$P(\pi_1|\sigma, \lambda) > P(\pi_2|\sigma, \lambda)$$

if and only if

$$K(\lambda)^n \exp\{-\lambda\, D(\pi_1)\}\, K^*(\lambda^*)\, \exp\{-\lambda^*\, D^*(\pi_1)\} > K(\lambda)^n \times \exp\{-\lambda\, D(\pi_2)\}\, K^*(\lambda^*)\, \exp\{-\lambda^*\, D^*(\pi_2)\}$$

if and only if

$$\exp[-\lambda\{D(\pi_1) - D(\pi_2)\}] > \exp[-\lambda^*\{D^*(\pi_2) - D^*(\pi_1)\}]$$

if and only if

$$\lambda\{D(\pi_1) - D(\pi_2)\} < \lambda^*\{D^*(\pi_2) - D^*(\pi_1)\}$$

if and only if

$$D(\pi_1) - D(\pi_2) < \gamma\{D^*(\pi_2) - D^*(\pi_1)\}. \qquad \square$$

When would the ranking with the highest proportion in the observed data also have the highest posterior probability?

Corollary 1. Let $\hat{\pi}$ be the maximum likelihood estimator of the modal ranking in the Mallows model, i.e. $\hat{\pi}$ minimizes $\Sigma_{i=1}^n d(\sigma_i, \pi)$. If $D^*(\pi_i) \geqslant D^*(\hat{\pi})$, π_i will have a lower posterior probability than $\hat{\pi}$.

Consider $\pi_1, \pi_2 \in S_k$. What are the conditions under which π_1 would have a higher posterior probability than π_2?

Corollary 2. If $\pi_2 \in C_{\pi_1}$, i.e. π_1 and π_2 belong to the same conjugacy class, π_1 has a higher posterior probability than π_2 if and only if $D(\pi_1) < D(\pi_2)$.

Corollary 3. If $D(\pi_1) < D(\pi_2)$ and $D^*(\pi_1) < D^*(\pi_2)$, then π_1 will have a higher posterior probability than π_2.

4. Extension to partially ranked data

It turns out that with some minor mathematical modifications the ideas developed thus far also carry over to the context of partially ranked data. We omit most of the details and only discuss some key points.

4.1. Partially ranked data

Given a set of n items, a partial ranking of k out of these n items is a ranking where only the first k choices are specified. An example would be when 10 candidates are contesting for an election and people are asked to rank only their five most favoured candidates. A partial ranking of this type forms an element of the coset space S_n/S_{n-k}, where S_{n-k} is the subgroup of S_n consisting of all permutations which leave the first k integers fixed:

$$S_{n-k} = \{\pi \in S_n : \pi(i) = i, 1 \leqslant i \leqslant k\}.$$

The equivalence relation '$\sim$', defined on S_n, is given by the following. For $\pi,\sigma \in S_n$, $\pi \sim \sigma \iff \pi\sigma^{-1} \in S_{n-k}$, partitions S_n into equivalence classes such that for any $\pi \in S_n$ the equivalence class containing π, denoted by $S_{n-k}\pi$, is $\{\tau\pi : \tau \in S_{n-k}\}$; it is called a right coset of S_{n-k}. It follows that, to each partial ranking of k out of n items, there corresponds a unique right coset of S_{n-k}, and two full permutations $\pi, \sigma \in S_n$ belong to the same right coset of S_{n-k} if and only if π and σ induce the same partial ranking of k out of n items, i.e. $\pi^{-1}(i) = \sigma^{-1}(i),\ 1 \leqslant i \leqslant k$.

All the metrics on fully ranked data discussed in Section 2.1 can be extended to form metrics on partially ranked data and the Mallows model for such data (Critchlow (1985), pages 100–101) can be written as

$$P(\pi^p) = C(\lambda) \exp\{-\lambda\, d_p(\pi^p, \pi_0^p)\}$$

for all partial rankings $\pi^p \in S_n/S_{n-k}$. Here, π_0^p is a location parameter representing the modal partial ranking and $\lambda \geqslant 0$ is a dispersion parameter. $d_p(\cdot,\cdot)$ is the induced Hausdorff metric on the coset space S_n/S_{n-k} and $C(\lambda)$ is the normalizing constant.

4.2. Prior distributions via coset classes on partially ranked data

Recall that, for fully ranked data, we divided the space of all rankings into conjugacy classes, because we believed that rankings within a conjugacy class were similar to each other as these rankings were assigned by permuting in a similar manner the order in which the items were observed.

In the case of partially ranked data, we wish to argue that, among all the people who rank k out of the n items, those who choose to rank the same set of k items are similar in some sense as they have the same choice of k favourite items but may choose to rank them differently. With this in mind, let us partition the coset space S_n/S_{n-k} into coset classes, where each coset class consists of all partial rankings having the same set of k items out of the n items which are ranked differently.

The metric-based prior for fully ranked data can be extended in a similar manner to form the analogous prior distribution for partially ranked data. Since the coset classes are bounded non-empty subsets of S_n/S_{n-k}, all the metrics defined on partially ranked data can be extended to form the corresponding induced Hausdorff metrics on the coset classes (see Section 3.2). Then, following the same argument as in the case of fully ranked data, if π_*^p is our choice of the modal partial ranking and $C_{\pi_*^p}$ is the corresponding coset class, the prior distribution on any partial ranking, π^p, is given by

$$P(\pi^p) = C_p(\lambda^*)\ \exp\{-\lambda^*\, d_p^*(C_{\pi^p}, C_{\pi_*^p})\},$$

where $d_p^*(\cdot,\cdot)$ is the induced metric on coset classes and $C_p(\lambda^*)$ is the normalizing constant.

The full conditional densities can be derived as in the previous case and the Gibbs sampler algorithm can be implemented. We omit the details.

5. Illustrative analyses

We first discuss an example simulated to illustrate the behaviour of the posterior distributions in the case of fully ranked data. The second example provides a comparative illustration with the data set analysed by Fligner and Verducci (1990). The third example illustrates the prior on partial rankings using the data in Diaconis (1988).

5.1. Example 1

A sample of 80 rankings is generated uniformly on S_4. For the data, the maximum likelihood estimator of π_0 in the Mallows model, using both Hamming's and Cayley's distance, is (2413). Choosing the prior modal parameter as $\pi^* = (4321)$ and the prior scale parameter as $\lambda^* = 0$, a Gibbs sampler is developed by using the method outlined in Section 3.3. For Hamming's and Cayley's distances, the posterior means of all the rankings are computed. They are listed in Table 2 and are classified by the conjugacy class to which they belong. Then λ^* is changed to 0.1 and the corresponding mean posterior probabilities are evaluated.

The maximum likelihood estimator of the modal ranking is $\hat{\pi} = (2413)$, the ranking with the minimum value of $D(\pi)$, shown in bold in Table 2. For all conjugacy classes with $D^*(\hat{\pi}) \leqslant D^*(\pi)$, the rankings in these classes have a lower posterior mean than $\hat{\pi}$, under both the metrics and for both values of λ^*, by corollary 1.

Within each conjugacy class, if $D(\pi_1) < D(\pi_2)$, then π_1 has a higher mean posterior density than π_2 has, by corollary 2.

Similarly by corollary 3, for any $\pi_1, \pi_2 \in S_4$, when $D(\pi_1) < D(\pi_2)$ and $D^*(\pi_1) < D^*(\pi_2)$, π_1 has a higher posterior mean probability than π_2 has.

However, the posterior means vary somewhat for the two metrics. This is mainly because the metrics differ greatly in the way in which they are defined and hence in the way in which they measure distances. This is evident from the differences in the $D(\pi)$ columns for the two metrics.

Next consider a sensitivity analysis. On increasing the value of λ^* to 0.1, the conjugacy class (2,2), containing the prior mode, $\pi^* = (4321)$ (for which $D^*(\pi) = 0$), has the largest increase in the posterior means whereas the conjugacy class (1,1,2), that is furthest away from this class (for which $D^*(\pi) = 3$ by Hamming and 2 by Cayley), has the largest decrease in the posterior means. So this prior parameter can be used to reflect our strength of belief in the prior mode.

The posterior means of λ from the Gibbs samplers corresponding to the fourth, fifth, eighth and ninth columns of Table 2 were 0.038, 0.037, 0.055 and 0.056 respectively.

5.2. Example 2

This example is taken from Fligner and Verducci (1990) in which the Graduate Record Examination Board sampled 98 college students who were asked to rank five words according to the strength of association with a target word. For the target word 'idea', the five choices were A, thought, B, play, C, theory, D, dream, and E, attention. If we can assume that all students were presented with these five choices in the same order as the above, then we can use our method of developing prior distributions to analyse these data.

Table 2. Posterior means of rankings in example 1

π	*Results for Hamming's distance*				*Results for Cayley's distance*			
	$D(\pi)$	$D^*(\pi)$	$\lambda^* = 0$	$\lambda^* = 0.1$	$D(\pi)$	$D^*(\pi)$	$\lambda^* = 0$	$\lambda^* = 0.1$
Conjugacy class (1,1,1,1)								
(1234)	242	4	0.0378	0.0313	155	2	0.0365	0.0240
Conjugacy class (1,1,2)								
(1243)	233	2	0.0531	0.0536	149	1	0.0512	0.0551
(1432)	233	2	0.0531	0.0536	147	1	0.0573	0.0618
(1324)	242	2	0.0378	0.0383	157	1	0.0327	0.0352
(4231)	242	2	0.0378	0.0383	153	1	0.0408	0.0439
(3214)	245	2	0.0338	0.0343	157	1	0.0327	0.0352
(2134)	249	2	0.0291	0.0296	159	1	0.0293	0.0315
Conjugacy class (2,2)								
(3412)	236	0	0.0474	0.0584	155	0	0.0365	0.0648
(2143)	240	0	0.0407	0.0503	155	0	0.0365	0.0648
(4321)	242	0	0.0378	0.0467	153	0	0.0408	0.0724
Conjugacy class (1,3)								
(1342)	233	3	0.0531	0.0485	149	2	0.0512	0.0334
(1423)	233	3	0.0531	0.0485	145	2	0.0643	0.0420
(2431)	235	3	0.0492	0.0450	155	2	0.0365	0.0238
(3241)	239	3	0.0423	0.0387	153	2	0.0408	0.0266
(4213)	239	3	0.0423	0.0387	153	2	0.0408	0.0266
(2314)	241	3	0.0392	0.0359	153	2	0.0408	0.0266
(4132)	247	3	0.0314	0.0288	157	2	0.0327	0.0213
(3124)	253	3	0.0252	0.0232	163	2	0.0236	0.0154
Conjugacy class (4)								
(2413)	**232**	**2**	**0.0552**	**0.0556**	**143**	**1**	**0.0722**	**0.0779**
(2341)	235	2	0.0492	0.0497	149	1	0.0512	0.0551
(3421)	239	2	0.0423	0.0428	151	1	0.0457	0.0492
(4312)	239	2	0.0423	0.0428	153	1	0.0408	0.0439
(3142)	244	2	0.0351	0.0355	153	1	0.0408	0.0439
(4123)	247	2	0.0314	0.0318	163	1	0.0236	0.0254

For the Mallows model, Fligner and Verducci assumed a uniform prior for the modal ranking π_0 and an independent conjugate prior for the scale parameter λ.

Suppose that we have reason to believe that (ADCBE) is the true order of association of the given words. Let us choose $\pi^* = (\text{ADCBE})$ as the prior mode. The choice of λ^* should reflect our strength of belief in π^*. To compare our method of analysis with the previous one, we shall also choose a uniform prior for π^* by using $\lambda^* = 0$. With the Hamming distance as the metric, a Gibbs sampling algorithm was used to obtain posterior estimates for the rankings. The results are provided in Table 3.

The first seven ranks with the highest observed frequencies are listed in Table 3 along with their mean posterior probabilities under the two methods; Mallows(uniform) corresponds to Fligner and Verducci's analysis. The numbers in parentheses in the third and fourth columns denote the ranks of the corresponding permutations in the posterior models. The posterior

Table 3. Posterior means of rankings in example 2

Rank	*Means from the following methods:*		
	Frequency (proportion)	*Mallows (uniform)*	*Mallows (conjugacy class)*
ACDEB	33 (0.337)	0.032 (1)	0.7591 (1)
ADCEB	18 (0.184)	0.026 (2)	0.1261 (2)
ACDBE	12 (0.122)	0.022 (3)	0.0217 (3)
ADCBE	8 (0.082)	0.019 (4)	0.0067 (7)
ACEDB	6 (0.061)	0.019 (5)	0.0144 (4)
CADEB	5 (0.051)	0.019 (6)	0.0040 (10)

Table 4. Posterior means in example 3

Ranking	*Proportion*	*Means for the following priors:*		
		$\lambda^* = 0$	$\lambda^* = 0.01$	$\lambda^* = 0.1$
00001	0.1988	0.20040	0.20023	0.23437
00010	0.2227	0.20357	0.20024	0.19493
00100	0.2330	0.20468	0.20015	0.19599
01000	0.1714	0.19538	0.20018	0.18707
10000	0.1741	0.19596	0.19969	0.18763

mean for λ using the conjugacy class prior is 0.053. Although our method picks the correct modal ranking and some of the subsequent rankings, the posterior mean of the modal ranking is much larger than its proportion in the observed data. As described in Section 3.3, even for small values of λ (0.053), an observed proportion of 0.337 is sufficiently large for the Mallows model to put most of the mass on the maximum likelihood estimator and to penalize other rankings strongly.

5.3. Example 3

In this example we illustrate our prior on partially ranked data. The American Psychological Association elects its President every year by asking each member to rank a slate of five candidates. Out of the 15000 members who cast their vote in 1980, 5141 selected their best candidate only. These rankings form elements of the coset spaces S_5/S_4 which we analyse below. Using the prior on partially ranked data discussed in this paper, the full Bayesian analysis was performed on these data. We choose the prior modal ranking to be $\pi_*^p = (00001)$, the ranking where the fifth candidate was selected as the favourite. Three values of λ^* are used, 0, 0.1 and 0.01, and the results are compared.

Recall that to define our prior distribution on the coset space we need to divide the space into coset classes, where each coset class consists of all rankings where the same candidates are selected. In the case of S_5/S_4, where only one candidate is being ranked, each ranking is a coset class. The observed proportions of each of these rankings along with their posterior means for the above choice of the prior parameters are provided in Table 4. We see that for a

non-informative prior ($\lambda^* = 0$) the posterior means are in the same order as the proportions in the data. On increasing λ^*, the highest mean shifts towards (00001), the prior modal ranking.

Acknowledgements

We thank the Associate Editor and two referees for their insightful comments that enhanced the paper.

References

Berry, D. A. (1979) Detecting trends in the arrangements of ordered objects: a likelihood approach. *Scand. J. Statist.*, **6**, 169–174.
Critchlow, D. E. (1985) *Metric Methods for Analyzing Partially Ranked Data*. New York: Springer.
Diaconis, P. (1987) *Group Representations in Probability and Statistics*. Haywood: Institute of Mathematical Statistics.
———(1988) A generalization of spectral analysis with applications to ranked data. *Ann. Statist.*, **17**, 949–979.
Feigin, P. D. and Cohen, A. (1978) On a model for concordance between judges. *J. R. Statist. Soc.* B, **40**, 203–213.
Fienberg, S. E. and Larntz, K. (1976) Log-linear representation for paired and multiple comparison of models. *Biometrika*, **63**, 345–354.
Fligner, M. A. and Verducci, J. S. (1986) Distance based ranking models. *J. R. Statist. Soc.* B, **48**, 359–369.
———(1988) Multistage ranking models. *J. Am. Statist. Ass.*, **83**, 892–901.
———(1990) Posterior probabilities for a consensus ordering. *Psychometrika*, **55**, 53–63.
Gordon, A. D. (1979) A measure of agreement between rankings. *Biometrika*, **66**, 7–15.
Henery, R. J. (1981) Permutation probabilities as models for horse races. *J. R. Statist. Soc.* B, **43**, 86–91.
Kuratowski, K. (1966) *Topology*, vol. I. New York: Academic Press.
Luce, R. D. (1959) *Individual Choice Behavior*. New York: Wiley.
MacKay, D. B. and Chaiy, S. (1982) Parametric estimation for the Thurstone case III model. *Psychometrika*, **47**, 353–359.
Mallows, C. L. (1957) Non null ranking models I. *Biometrika*, **44**, 114–130.
Mosteller, F. (1951) Remarks on the method of paired comparisons, I: the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**, 3–9.
Plackett, R. L. (1975) The analysis of permutations. *Appl. Statist.*, **24**, 193–202.
Schulman, R. S. (1979) Ordinal data: an alternative distribution. *Psychometrika*, **44**, 3–20.
Serre, J. P. (1977) *Linear Representations of Finite Groups*. New York: Springer.
Tallis, G. M. and Dansie, B. R. (1983) An alternative approach to the analysis of permutations. *Appl. Statist.*, **32**, 110–114.
Thurstone, L. L. (1927) A law of comparative judgment. *Psychol. Rev.*, **34**, 273–286.