Conjugacy Class Prior Distributions on Metric Based Ranking Models

JAYANTI GUPTA
University of Michigan, Ann Arbor, MI.
PAUL DAMIEN
University of Michigan, Ann Arbor, MI.

SUMMARY

A new class of prior distributions for metric based models in the analysis of fully ranked data is developed. This class has two attractive features: first, it provides a meaningful way to encapsulate prior information about the parameters of the model; second, a full Bayesian solution is made possible via stochastic simulation methods. The ideas are exemplified using illustrative analyses.

Key words: Ranked data, Mallows model, Permutation group, Bayesian inference, Gibbs sampling.

1 Introduction

Suppose that each person in a random sample of n people is asked to rank his preferences among a fixed set of k items. Each person produces a ranking $\pi = [\pi(1), \ldots, \pi(k)]$, where $\pi(i)$ denotes the rank assigned to item $i, (i = 1, \ldots, k)$. The problem here is to characterize the population based on a random sample $\pi_j, j = 1, \ldots, n$ of rankings.

Historically, models for random rankings grew out of the literature on paired comparisons. For example, the Thurstone (1927) model specified that item i would be preferred to item j if $X_i > X_j$ where the $X's$ are i.i.d. normal random variables with different means and equal variance. Mosteller (1951) provided simple forms for the least square estimators of the means of $X's$ under this model. MacKay and Chaiy (1982) used Monte Carlo methods to compare estimators of the means of $X's$ in the above model with those under the unequal variance model. Clearly, these paired comparison models can be extended to rankings by letting $\pi_i =$ 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 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), Tallis and Dansie (1983). Gordon (1979) introduced a model based on Ulam's distance, while Fligner and Verducci (1986) 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 two classes of prior distributions on the space of rankings. The first one, called the conjugacy class prior distribution uses properties of the permutation group to define the nature of the prior. The second one, called the metric based prior distribution uses the notion of metrics on conjugacy classes to define prior distributions on rankings. 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 both the priors.
The paper is organised 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. Section 4 defines metrics between conjugacy classes, and discusses the metric based prior distribution. Section 5 illustrates these priors via examples and Section 6 concludes the paper with a discussion of the ideas proposed here.

2 Models and Metrics

2.1 The Mallows Model

Colin Mallows(1957) proposed a non-null probability model, that is 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, $P(\pi) = K(\lambda)e^{-\lambda d(\pi,\pi_0)}$, for all $\pi \in S_k$, $d(,)$ is a metric on S_k, π_0 is the

location parameter and $\lambda \geq 0$ is a dispersion parameter. The normalizing constant, $K(\lambda)$, is defined as $K(\lambda)^{-1} = \sum_{\pi \in S_k} e^{-\lambda d(\pi, \pi_0)}$, and is independent of the choice of π_0. The model is centered about the ranking π_0, and as λ increases the distribution becomes more and more peaked about π_0.

2.2 Some metrics on fully ranked data

By 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:

$R(\pi, \sigma) = (\sum_{i=1}^{k} (\pi(i) - \sigma(i))^2)^{1/2}$ is Spearman's rho distance.

$F(\pi, \sigma) = \sum_{i=1}^{k} |\pi(i) - \sigma(i)|$ is Spearman's footrule.

$T(\pi, \sigma)$ = number of pairs of items, (i, j), such that $\pi(i) < \pi(j)$ and $\sigma(i) > \sigma(j)$ is Kendall's τ.

$H(\pi, \sigma) = \#\{i = 1, \ldots, k : \pi(i) \neq \sigma(i)\}$ is Hamming distance.

$U(\pi, \sigma) = k-$ the maximal number of items ranked in the same order by π and σ is Ulam's distance.

$C(\pi, \sigma)$ = minimum number of transpositions needed to transform π into σ is Cayley's distance.

It should be noted that all these metrics are right invariant, that is they remain unchanged under arbitrary relabeling of the items.

3 Conjugacy Class Prior on S_k

Fligner and Verducci (1990) performed a Bayesian analysis on ranked data by introducing prior distributions on the parameters of the Mallows model. They assume a uniform prior for the modal ranking, π_0 and an independent conjugate prior for the scale parameter given π_0. In this section, we develop a new class of prior distributions for the modal ranking in the Mallows model;

the uniform prior, of course, is obtained as a special case.

3.1 Conjugacy Classes of the Permutation Group

Definition 3.1.1 *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 exists an element σ in S_k such that $\pi_1 = \sigma\pi_2\sigma^{-1}$.*

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

Definition 3.1.2 *Given an integer k, we say the sequence of positive integers, $k_1, k_2, \ldots, k_r$, $k_1 \leq k_2 \leq \cdots \leq k_r$ constitute a partition of k if $k = k_1 + k_2 + \cdots + k_r$.*

Definition 3.1.3 *The set of integers $(i_1, i_2, \ldots, i_r)$ 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, ..., i_{r-1} into i_r and i_r into i_1, and leaves all other items fixed.*

Definition 3.1.4 *A permutation $\pi \in S_k$ has the cycle decomposition $\{k_1, k_2, \ldots, k_r\}$ if it can be written as the product of disjoint cycles of lengths $k_1, k_2, \ldots, k_r$, $k_1 \leq k_2 \leq \cdots \leq k_r$.*

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 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 $k_1, k_2, \ldots, k_r$ respectively,

$k_1 \leq k_2 \leq \cdots \leq k_r$, then $k = k_1 + k_2 + \cdots + k_r$.

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 written in cycle notation is :

$(a \cdots b)(c \cdots d) \cdots (e \cdots f)$,

then,

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

This results in the following

Lemma: The total number of conjugacy classes in S_k, is $p(k)$, the number of partitions of k.

Proof: There is one conjugacy class for each partition of k: thus the identity forms a class, the transpositions or 2 cycles {(i,j)} form a class, the 3 cycles {(ijk)}, products of 2-2 cycles {(ij)(kl)}, and so on; each form a conjugacy class, whence the lemma.

3.2 Choice of prior distribution on S_k

The Mallows model:

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

has two parameters, π_0 the location parameter and $\lambda \geq 0$ the scale parameter. It is well known [Serre, 1977; pp 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 distribution for the scale parameter λ could be :

$P(\lambda) \propto exp^{-\alpha_0\lambda}$, $\lambda \in \Re^+$, i.e. Exponential(α_0). Interest here is on developing a prior distribution for the modal ranking π_0.

We develop a new approach to constructing a prior distribution for π_0. This approach must satisfy two features: (a) can one exploit the structure of the permutation group in which the data is observed to encapsulate prior information about the random quantity π_0? (b) does the prior distribution have 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. In particular, the prior distribution on π_0 will be taken to be constant on conjugacy classes, *and* proportional to the number of elements in the conjugacy class it belongs to. 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 remaining unchanged.

Let us consider a simple example to illustrate 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 here along with the number of elements in each class.

Partition	Conjugacy Class	# 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

The elements within a conjugacy class are *similar* as they represent rankings of items that were assigned by permuting the order in which the items were observed in a similar manner.

In the above 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 1's in this partition while the remaining two had their ranks interchanged, corresponding to the 2 in the partition. For example, in the permutation (4231) which belongs to this class, the items that were observed second and third were assigned ranks 2 and 3 respectively, while the items appearing first and fourth received ranks 4 and 1 respectively.

Due to this similarity between rankings within a conjugacy class we assign equal prior probabilities to all these rankings. Also, conjugacy classes are invariant to relabeling of the items, i.e. if all the items were observed in a different order, the conjugacy classes would remain unchanged. Further, a conjugacy class that has more elements is more likely to be chosen, hence these probabilities are also proportional to the number of elements in the conjugacy classes they belong to or, more generally, to some function of the number of elements in these classes.

Recall that in the Mallows model the probability of any ranking decreases exponentially as its distance from the modal ranking π_0 increases. We may argue that the prior probability of the modal ranking increases exponentially as the number of elements in its conjugacy class increases. This would give rise to another rule for assigning prior probabilities to rankings given by:

$$P(\pi) \propto e^{\beta c(\pi)}$$

where $c(\pi)$ = number of elements in the conjugacy class of π, and $\beta \geq 0$ represents the strength of our prior belief. Larger values of β indicate stronger prior beliefs about the distribution of the modal ranking, and vice versa, with $\beta = 0$ denoting the uniform distribution of the modal ranking. Since

the space of all rankings S_k is finite, the prior distribution on π_0 is proper, which implies that the posterior on π_0 will also be proper.

3.3 A Gibbs sampler for the Mallows model

Here, we develop a Gibbs sampler algorithm to sample from the posterior distribution of the parameters of the Mallows model using the prior distribution developed earlier.

Suppose a group of k items are being ranked by n people. Let $\pi_1, \pi_2, \ldots, \pi_{k!}$ be the set of all possible rankings of these k items.
Let the data be $(\sigma_i, i = 1, \ldots, n)$, the rankings of these items by the n people.
The joint likelihood is given by :

$$\begin{aligned} P(\sigma_1, \ldots, \sigma_n | \pi_0, \lambda) &= (K(\lambda))^n e^{-\lambda \sum_{i=1}^{n} d(\sigma_i, \pi_0)} \\ &= (K(\lambda))^n e^{-\lambda \sum_{i=1}^{r} n_i d(\sigma_i, \pi_0)} \end{aligned}$$

where

r= # of distinct rankings in the data

n_i= # of people with ranking σ_i

The prior distribution of the parameters is :

$$\begin{aligned} P(\pi_0, \lambda) &= P(\pi_0) * P(\lambda) \\ &= \frac{e^{\beta c(\pi_0)}}{\sum_{j=1}^{k!} e^{\beta c(\pi_j)}} \alpha_0 e^{-\lambda \alpha_0} \end{aligned}$$

where

$c(\pi_0)$ = # of elements in the conjugacy class of π_0.

The conditional density of $\lambda | data, \pi_0$ is :

$$\begin{aligned} P(\lambda | data, \pi_0) &= \frac{P(data | \pi_0, \lambda) P(\lambda)}{\int_0^\infty P(data | \pi_0, \lambda') P(\lambda') d(\lambda')} \\ &\propto (K(\lambda))^n e^{-\lambda(\alpha_0 + \sum_{i=1}^{r} n_i d(\sigma_i, \pi_0))} \end{aligned}$$

The conditional density of $\pi_0|data, \lambda$ is :

$$P(\pi_0|data,\lambda) = \frac{P(data|\pi_0,\lambda)P(\pi_0)}{\sum_{\pi_j \in S_k} P(data|\pi_j,\lambda)P(\pi_j)}$$

$$= \frac{e^{-\lambda \sum_{i=1}^{r} n_i d(\sigma_i,\pi_0)} e^{\beta c(\pi_0)}}{\sum_{j=1}^{k!} e^{-\lambda \sum_{i=1}^{r} n_i d_{ij}} e^{\beta c_j}}$$

where $d_{ij} = d(\pi_i, \pi_j)$,
$c_j = c(\pi_j)$, j runs over all $k!$ permutations in S_k.
Using the full conditional densities as described above, a hybrid Gibbs/Metropolis Hastings algorithm can be used to obtain posterior estimates of the parameters. Details of the algorithm are given in the appendix.

Computation of $c(\pi_0)$:
Let π_0 correspond to the partition $\lambda_1 + \lambda_2 + \cdots + \lambda_r = k$, i.e. π_0 can be written as the product of r disjoint cycles of lengths $\lambda_1, \lambda_2, \ldots, \lambda_r$ respectively. Suppose that r_1 of these are distinct. Let $\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_{r_1}$ denote these distinct values and let the number of $\lambda's$ equal to $\overline{\lambda}_i$ be $k_i, i = 1, \ldots, r_1$.
Then $c(\pi_0)$, the # of elements of S_k that belong to the above partition is :

$$\frac{k!}{\overline{\lambda}_1^{k_1}\overline{\lambda}_2^{k_2}\cdots\overline{\lambda}_{r_1}^{k_{r_1}}}\left(\prod_{i=1}^{r_1}(k_i!)^{-1}\right.$$

With our choice of prior distribution on the parameters, we would now be interested in finding the modal ranking in the posterior distribution. Or for $\pi_1, \pi_2 \in S_k$, what are the conditions under which π_1 would have a higher posterior probability than π_2? When would the ranking with the highest proportion in the observed data also have the highest posterior probability? Is there a relationship between λ the scale parameter in the model and β the scaling factor in the prior distribution that would determine which rank-

ing gets the highest posterior distribution? The following theorem and its corollaries attempt to answer some of these questions.

Theorem 3.1 *Let* $D(\pi_j) = \sum_{i=1}^{n} d(\sigma_i, \pi_j)$, *the sum of the distances of the observed rankings from* π_j. *For* $\pi_1, \pi_2 \in S_k$, *and given* $\lambda, \beta > 0$, π_1 *will have a higher posterior probability than* π_2 *iff*

$$c(\pi_1) - c(\pi_2) > \gamma(D(\pi_1) - D(\pi_2)) \quad \textit{where} \quad \gamma = \frac{\lambda}{\beta}.$$

Proof:

$$\begin{aligned}
P(\pi_1|data, \lambda) &> P(\pi_2|data, \lambda) \\
\text{iff} \quad (K(\lambda))^n e^{-\lambda \sum_{i=1}^{n} d(\sigma_i, \pi_1)} \frac{e^{\beta c(\pi_1)}}{\sum_{j=1}^{k!} e^{\beta c(\pi_j)}} &> (K(\lambda))^n e^{-\lambda \sum_{i=1}^{n} d(\sigma_i, \pi_2)} \frac{e^{\beta c(\pi_2)}}{\sum_{j=1}^{k!} e^{\beta c(\pi_j)}} \\
\text{iff} \quad e^{-\lambda D(\pi_1) + \beta c(\pi_1)} &> e^{-\lambda D(\pi_2) + \beta c(\pi_2)} \\
\text{iff} \quad e^{\beta(c(\pi_1 - \pi_2))} &> e^{\lambda(D(\pi_1) - D(\pi_2))} \\
\text{iff} \quad \beta(c(\pi_1) - c(\pi_2)) &> \lambda(D(\pi_1) - D(\pi_2)) \\
\text{iff} \quad c(\pi_1) - c(\pi_2) &> \gamma(D(\pi_1) - D(\pi_2))
\end{aligned}$$

Corollary 3.1 *Let* $\hat{\pi}$ *be the m.l.e. of the modal ranking in the Mallows model, i.e.* $\hat{\pi}$ *minimizes* $\sum_{i=1}^{n} d(\sigma_i, \pi)$. *If* $c(\pi_i) \leq c(\hat{\pi})$, π_i *will have a lower posterior probability than* $\hat{\pi}$.

Corollary 3.2 *For* $\pi_1, \pi_2 \in S_k$, *if* $c(\pi_1) = c(\pi_2)$, *then* π_1 *will have a higher posterior probability than* π_2 *iff* $D(\pi_1) < D(\pi_2)$.

Corollary 3.3 *For* $\pi_1, \pi_2 \in S_k$, *if* $c(\pi_1) > c(\pi_2)$ *and* $D(\pi_1) < D(\pi_2)$, *then* π_1 *will have a higher posterior probability than* π_2.

4 A Metric-Based Prior Distribution for π_0

4.1 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: pp 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^* in right-invariant if d is.

Theorem 4.1 *If a metric d on S_k is bi-variant, 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:

$$\begin{aligned}
\max_{\beta\in C_\sigma}\min_{\alpha\in C_\pi} d(\alpha,\beta) &= \max_{\tau_1\in S_k}\min_{\tau_2\in S_k} d(\tau_1\pi\tau_1^{-1}, \tau_2\sigma\tau_2^{-1}) \\
&= \max_{\tau_1\in S_k}\min_{\tau_2\in S_k} d(\tau_1\pi, \tau_2\sigma\tau_2^{-1}\tau_1) \\
&= \max_{\tau_1\in S_k}\min_{\tau_2\in S_k} d(\pi, \tau_1^{-1}\tau_2\sigma\tau_2^{-1}\tau_1) \\
&= \max_{\tau_1\in S_k}\min_{\tau_3\in S_k} d(\pi, \tau_3\sigma\tau_3^{-1}) \qquad \text{where } \tau_3 = \tau_1^{-1}\tau_2 \\
&= \min_{\tau_3\in S_k} d(\pi, \tau_3\sigma\tau_3^{-1}) \\
&= \min_{\beta\in C_\sigma} d(\pi,\beta)
\end{aligned}$$

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

$$
\begin{aligned}
\text{Hence,} \quad d^*(C_\pi, C_\sigma) &= \min_{\beta \in C_\sigma} d(\pi, \beta) \\
\text{Also,} \quad \min_{\beta \in C_\sigma} d(\pi, \beta) &= \min_{\tau \in S_k} d(\pi, \tau\sigma\tau^{-1}) \\
&= \min_{\tau_1, \tau_2 \in S_k} d(\pi, \tau_1^{-1}\tau_2\sigma(\tau_1^{-1}\tau_2)^{-1}) \\
&= \min_{\tau_1, \tau_2 \in S_k} d(\tau_1\pi\tau_1^{-1}, \tau_2\sigma\tau_2^{-1}) \\
&= \min_{\substack{\alpha \in C_\pi \\ \beta \in C_\sigma}} d(\alpha, \beta)
\end{aligned}
$$

Among the six metrics defined in Section 2.2, only two are bi-variant: Hamming distance and Cayley's distance. The above theorem can thus be used to derive simpler forms for these two metrics. However, any metric can be made invariant by averaging it [Diaconis, 1987 pp 114-115]. Hence the above theorem can be used to compute induced Hausdorff distances for all of the metrics. The corollaries below provide the simplified forms of the metrics induced by Cayley's and Hamming distance.

Corollary 4.1 *The Hausdorff metric between two conjugacy classes induced by Cayley's distance is the minimum number of transpositions required to transform a permutation of one of the conjugacy classes into a permutation of the other class.*

This transformation occurs in the following manner:
Let the ranking π belong to the conjugacy class having the cycle decomposition $\{k_1, k_2, k_3, \ldots, k_r\}$. Suppose π is given by :

$$\pi = (a_1, \ldots, a_{k_1})(b_1, \ldots, b_{k_2}) \cdots (p_1, \ldots, p_{k_r})$$

A transposition on π can be one of two possible types: the two numbers being permuted could either belong to the same cycle of π, (say a_i and a_j), or they could belong to different cycles of π, (say a_i and b_j).
A transposition of the first type results in a ranking with cycle decomposition $\{(j-i), k_1-(j-i), k_2, \ldots, k_r\}$, i.e. the cycle containing a_i and a_j is split

into two cycles with lengths depending on the positions of a_i and a_j in the cycle, while a transposition of the second type gives the cycle decomposition $\{(k_1+k_2), k_3, \ldots, k_r\}$ where the two cycles containing a_i and b_j are combined to form one cycle of length $k_1 + k_2$.

Definition 4.1 *For two conjugacy classes C_π and C_σ in S_k, with partitions $\{k_1, k_2, \ldots, k_l\}$ and $\{p_1, p_2, \ldots, p_m\}$ respectively, the relative partition of C_π w.r.t. C_σ denoted by $C_\pi|C_\sigma$ is the set of $k_i's$ without those that are equal to some distinct p_j.*

For example, consider the conjugacy classes $C_\pi = \{2, 2, 1\}$ and $C_\sigma = \{3, 2\}$ of S_5. Then, the relative partition of C_π w.r.t. C_σ is given by $C_\pi|C_\sigma = \{2, 1\}$, while the relative partition of C_σ w.r.t. C_π is given by $C_\sigma|C_\pi = \{3\}$.

Definition 4.2 *For two conjugacy classes C_π and C_σ in S_k, with partitions $\{k_1, k_2, \ldots, k_l\}$ and $\{p_1, p_2, \ldots, p_m\}$ respectively, C_π and C_σ are said to be divisible if $\exists\ \mathcal{K} \subsetneqq C_\pi|C_\sigma$ and $\mathcal{P} \subsetneqq C_\sigma|C_\pi$, such that $\mathcal{K}, \mathcal{P} \neq \phi$ and*

$$\Sigma\{k_i : k_i \in \mathcal{K}\} = \Sigma\{p_j : p_j \in \mathcal{P}\}$$

Then, $\{\mathcal{K}, \mathcal{P}\}$ are called a pair of divisible subpartitions of $\{C_\pi, C_\sigma\}$.

For example, the conjugacy classes of S_{11} given by $C_\pi = \{5, 4, 1, 1\}$ and $C_\sigma = \{3, 2, 2, 2, 2\}$ are divisible. A pair of divisible subpartitions would be $\mathcal{K} = \{5\}$ and $\mathcal{P} = \{3, 2\}$. $\mathcal{K}_1 = \{4\}, \mathcal{P}_1 = \{2, 2\}$ form a second pair of divisible subpartitions which is disjoint from the first pair.

Corollary 4.2 *The Hausdorff metric induced by Hamming distance between two conjugacy classes C_π and C_σ that are not divisible is the number of integers in $C_\pi|C_\sigma$ plus the number of integers in $C_\sigma|C_\pi$ minus 1.*

$$i.e.\quad d^*(C_\pi, C_\sigma) = \#\{C_\pi|C_\sigma\} + \#\{C_\sigma|C_\pi\} - 1$$

Proof: Let $\{k_1, k_2, \ldots, k_l\}$ and $\{p_1, p_2, \ldots, p_m\}$ be the partitions of C_π and C_σ respectively. Without loss of generality, suppose $k_1 < p_1$. Then of the k_1 elements corresponding to the first cycle of π, atmost $k_1 - 1$ of them would be equal to the elements in the corresponding positions of σ. Hence the minimum contribution to Hamming distance from these k_1 elements would be 1.

Now consider k_2 and $(p_1 - k_1)$. As before, if $k_2 < (p_1 - k_1)$, minimum contribution to Hamming distance from these k_2 elements of π is 1. If $k_2 > (p_1 - k_1)$, the minimum contribution from the $(p_1 - k_1)$ elements would again be 1. Since C_π and C_σ are not divisible, $k_2 \neq (p_1 - k_1)$.

Continuing this pairwise comparison of cycle lengths of C_π and C_σ, we see that for each comparison, the minimum contribution to Hamming distance is 1. Since the number of integers in the partitions of C_π and C_σ are l and m respectively, there would be $l + m - 1$ such comparisons, hence the corollary.

Corollary 4.3 *When C_π and C_σ are divisible, let r be the maximum number of disjoint pairs of divisible subpartitions of $\{C_\pi, C_\sigma\}$. Then the induced Hamming distance between them is given by*

$$d^*(C_\pi, C_\sigma) = \#\{C_\pi | C_\sigma\} + \#\{C_\sigma | C_\pi\} - r$$

Proof: The proof of this result is the same as the previous one, except that now since C_π and C_σ are divisible, it is possible that $k_2 = (p_1 - k_1)$ when $\mathcal{K} = \{k_1, k_2\}$ and $\mathcal{P} = \{p_1\}$ form a pair of divisible subpartitions. Then the minimum contribution to Hamming distance from these elements would still be 1 but the next pairwise comparison would now be between k_3 and p_2.

So we now lose one pairwise comparison, hence one contribution to the distance.

If r is the maximum number of disjoint pairs of divisible subpartitions of $\{C_\pi, C_\sigma\}$, then in the pairwise comparison of cycle lengths of C_π and C_σ, we would lose a maximum of $r - 1$ such comparison s, hence the result.

4.2 Prior Distribution via Metrics on Conjugacy Classes

In this section we introduce another prior distribution on rankings that is based on the distance between conjugacy classes. Let us specify a ranking, π^* which we believe apriori to be the modal ranking. Let C_{π^*} be the conjugacy class containing π^*.

We argue that the prior distribution on a conjugacy class decreases exponentially as its distance from C_{π^*} increases, i.e.

$$P(C_\pi) \propto e^{-\lambda^* d^*(C_\pi, C_{\pi^*})}, \qquad \lambda^* \geq 0$$

λ^* is a scalar that determines how peaked the distribution is around C_{π^*}. This prior distribution on conjugacy classes induces a prior distribution on rankings, if we assume that all rankings within a conjugacy class have the same prior distribution. The induced prior distribution on rankings is given by:

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

Let us develop the Gibbs sampler algorithm for the Mallows model using the above prior on π_0, the modal ranking :

Let the prior distribution on , λ be $Exp(\alpha_0)$.

The conditional density of $\lambda | data, \pi_0$ is :

$$\begin{aligned} P(\lambda | data, \pi_0) &= \frac{P(data|\pi_0, \lambda)P(\lambda)}{\int_0^\infty P(data|\pi_0, \lambda')P(\lambda')d(\lambda')} \\ &\propto (K(\lambda))^n e^{-\lambda(\alpha_0 + \sum_{i=1}^{r} n_i d(\sigma_i, \pi_0))} \end{aligned}$$

The conditional density of $\pi_0 | data, \lambda$ is :

$$P(\pi_0 | data, \lambda) = \frac{P(data|\pi_0, \lambda)P(\pi_0)}{\sum_{\pi_j \in S_k} P(data|\pi_j, \lambda)P(\pi_j)}$$

$$= \frac{e^{-\lambda \sum_{i=1}^{r} n_i d(\sigma_i, \pi_0)} \frac{e^{-\lambda^* d^*(C_{\pi_0}, C_{\pi^*})}}{c(\pi_0)}}{\sum_{j=1}^{k!} e^{-\lambda \sum_{i=1}^{r} n_i d_{ij}} \frac{e^{-\lambda^* d^*(C_{\pi_j}, C_{\pi^*})}}{c(\pi_j)}}$$

5 Prior Distributions on Partially Ranked Data

5.1 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 5 of their most favorite 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:

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

The equivalence relation $\sim$, defined on S_n by:
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 $\{\tau\pi : \tau \in S_{n-k}\}$ and 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} iff π and σ induce the same partial ranking of k out of n items, i.e. $\pi^{-1}(i) = \sigma^{-1}(i), 1 \leq i \leq k$.

All the metrics on fully ranked data discussed in section 2.2 can be extended to form metrics on partially ranked data and the Mallows model for such data [Critchlow,1985 pp 100-101] can be written as :

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

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 $\lambda \geq 0$ is a dispersion parameter. $d_p(,)$ is the induced Hausdorff metric on the coset space S_n/S_{n-k} and $C(\lambda)$ is the normalizing constant.

5.2 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 the order in which the items were observed in a similar manner.
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 that choose to rank the same set of k items are similar in some sense as they have the same choice of k favorite 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 that choose the same set of k items out of the n items but rank them differently.

On fully ranked data, the conjugacy class prior on π_0 assigned equal probabilities to rankings within a conjugacy class that was proportional to the number of elements in the class, $c(\pi_0)$, and was given by:

$$P(\pi_j) \propto e^{\beta c(\pi_j)}$$

In the case of partially ranked data, each coset class has the same number of elements, so the analog of the conjugacy class prior here would be :

$$P(\pi_j^p) \propto \beta_j$$

where β_j is a scalar quantity representing our prior belief on the coset class containing π_j^p. So our choice of β_j for each coset class should be proportional

to the strength of our belief in the k items being ranked in that class.

Using this prior on the modal partial ranking, and an independent exponential prior for the scale parameter, λ, the conditional density of $\pi_0^p|data, \lambda$ is calculated to be :

$$P(\pi_0^p|data, \lambda) = \frac{e^{-\lambda \sum_{i=1}^{m} d_p(\sigma_i^p, \pi_0^p)} \beta_0}{k! \sum_{j=1}^{n_k} e^{-\lambda \sum_{i=1}^{m} d_p(\sigma_i^p, \pi_j^p)} \beta_j}$$

where $\sigma_i^p, i = 1, \ldots, m$ are the observed partial rankings and $n_k = \binom{n}{k}$, the number of coset classes in S_n/S_{n-k}.

The metric based prior on fully ranked data can be extended in a similar manner to form the analogous prior distribution on 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 4.1]. Then following the same argument as in the case of fully ranked data, if $C_{\pi_*^p}$ is our choice of the modal coset class, the prior distribution on any coset class, C_{π^p} is given by :

$$P(C_{\pi^p}) \propto e^{-\lambda^* d_p^*(C_{\pi^p}, C_{\pi_*^p})}$$

where $d_p^*(,)$ is the induced metric on coset classes. Assuming that all partial rankings in a coset class have the same prior distribution, this induces a prior distribution on partially ranked data given by :

$$P(\pi^p) = \frac{e^{-\lambda^* d_p^*(C_{\pi^p}, C_{\pi_*^p})}}{k! \sum_{i=1}^{n_k} e^{-\lambda^* d_p^*(C_{\pi_i^p}, C_{\pi_*^p})}}$$

The full conditional densities can be calculated as in the previous case and the Gibbs sampler algorithm can be developed in a similar manner to simulate from the posterior densities of π_0^p and λ.

6 Illustrative Analyses

The first example discussed here is a simulated example: the motivation here is to illustrate the posterior distributions obtained under different prior/metric combinations. The second example provides a comparative illustration with the data set analyzed by Fligner and Verducci(1990).

Example 1:

Let us illustrate the conjugacy class prior on S_4 with a simulated example. With the modal ranking, $\pi_0 = (2143)$ and $\lambda = 0.4$, samples of size 80 were generated according to the Mallows model using four different metrics and for different prior settings, the posterior probabilities of all the rankings were computed.

β	Kendall's τ	Footrule	Hamming	Spearman's ρ
0	0.9901 (2143)	0.9999 (2143)	0.9997(2143)	0.9952 (2143)
0.5	0.9569 (2143)	0.9999 (2143)	0.9987 (2143)	0.9789 (2143)
1	0.8321 (2143)	0.9998 (2143)	0.9940 (2143)	0.9119 (1243)
1.5	0.5250 (1243)	0.9991 (2143)	0.9717 (2143)	0.6973 (1243)
5	0.4287 (3142)	0.5824 (1243)	0.9581 (4132)	0.3444 (4132)
10	0.6026 (3241)	0.7700 (4132)	0.9645 (1243)	0.4399 (4132)

The conjugacy class containing the true modal ranking, (2143), corresponds to the partition (2,2) and contains only 3 elements.

When $\beta = 0$, the prior is uniform over all rankings, hence the ranking with the highest posterior probability is the true modal ranking, using all four metrics.

Changing β to 0.5 gives similar results.

When $\beta = 1.0$, the highest posterior probability is now given to the ranking (1243) by one of the metrics, which belongs to the conjugacy class corre-

sponding to the partition (1,1,2) and has a larger (6) number of elements. When $\beta = 5$, all metrics give the highest posterior probability to some ranking other than the true modal ranking, which belong to conjugacy classes that have larger number (6 or 8) of elements. Thus, the choice of β can be used to define the strength of our belief on the prior distribution.

Using the same data, but the metric based prior, the analysis was repeated using the Hausdorff metric induced by Hamming distance. Two different choices of π^* were used, $\pi^* = (2143)$, the true modal ranking, and $\pi^* = (2134)$. For six different choices of λ^* the modal rankings in the posterior models are tabulated below along with their probabilities.

λ^*	$\pi^* = (2143)$	$\pi^* = (2134)$
0	0.9998 (2143)	0.9998 (2143)
0.5	0.9999 (2143)	0.9996 (2143)
1	0.9999 (2143)	0.9989 (2143)
1.5	0.9999 (2143)	0.9973 (2143)
5	1.0000 (2143)	0.6228 (2134)
10	1.0000 (2143)	0.8320 (2134)

So for the correct choice of π^*, the posterior model puts most of the mass at the true modal ranking, whereas with the incorrect choice of π^*, on increasing the value of λ^*, maximum posterior probability is given to this incorrect ranking.

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.

They fit both the Mallows and generalized Mallows model to the data. For the Mallows model, they assume a uniform prior for the modal ranking, π_0 and an independent conjugate prior for the scale parameter, λ.

The conjugacy class prior for the Mallows model with $\beta = 0$ and Kendall's τ metric was used to analyze this data, and a Gibbs sampling algorithm was used to obtain posterior estimates for the rankings. The results are tabulated below along with the results from the previous analyses.

The first 7 ranks with the highest observed frequencies are listed along with their mean posterior probabilities by the two methods. The numbers in parentheses in the 3rd and 4th columns denote the ranks of the corresponding permutations in the posterior models.

Obs ranks	Freq(prob)	Mallows(unif)	Mallows(conj. class)
15234	33 (.337)	.032 (1)	.8658 (1)
15324	18 (.184)	.026 (2)	.1039 (2)
14235	12 (.122)	.022 (3)	.0220 (3)
14325	8 (.082)	.019 (4)	.0032 (4)
15243	6 (.061)	.019 (5)	.0021 (5)
25134	5 (.051)	.019 (6)	.0021 (6)
15423	5 (.051)	.016 (7)	.0003 (7)

The posterior mean for lambda is 0.083. So our method not only picks the correct modal ranking and most of the subsequent rankings, the posterior mean of the modal ranking is a lot higher than that obtained by the previous analysis, reflecting its high proportion in the observed data.

7 Discussion

In this paper, we discussed the concept of conjugacy classes in permutation groups and introduced the use of these classes to define two types of prior distributions on metric based ranking models. The conjugacy class prior is useful when we do not have any prior knowledge of what the modal ranking could be, whereas the metric-based prior is useful when we have some idea about what the most popular ranking is. Each of these priors can control the strength of our prior belief through scale parameters, β and λ^* respectively.

References

Berry,D.A. (1979), "Detecting Trends in the Arrangements of Ordered Objects: A Likelihood Approach," *Scandinavian Journal of Statistics*, 6,169-174.

Critchlow,D.E.(1985), Metric Methods for Analyzing Partially Data, Springer-Verlag: New York

Diaconis,P. (1987), Group Representations in Probability and Statistics, Haywood, CA: Institute of Mathematical Statistics.

Diaconis,P. (1988), "A Generalization of Spectral Analysis with Applications to Ranked Data," *Annals of Statistics,* 17,949-979.

Feigin,P., and Cohen,A. (1978), "On a Model for Concordance Between Judges," *Journal of the Royal Statistical Society,* Ser.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," *Journal of the Royal Statistical Society,* Ser.B, 48,859-869.

Fligner,M.A., and Verducci,J.S. (1988), "Multistage Ranking Models," *Journal of the American Statistical Association* 83,892-901.

Fligner,M.A., and Verducci,J.S. (1990), "Posterior Probabilities for a Con-

sensus Ordering," *Psychometrika* 55,No. 1,53-63.

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," *Journal of the Royal Statistical Society,* Ser.B, 43,86-91.

Kuratowski,K. (1966), Topology: Volume I, New York: Academic Press.

Luce,R.D. (1959), *Individual Choice Behavior,* New York: John 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.

Michael,E. (1951), "Topologies on Spaces of Subsets," *Transactions of the American Mathematical Society,* 71,152-182.

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," *Applied Statistics,* 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-Verlag.

Tallis,G.M.,and Dansie,B.R. (1983),"An Alternative Approach to the Analysis of Permutations," *Applied Statistics,* 32,110-114.

Thurstone,L.L. (1927),"A Law of Comparative Judgment," *Psychological Reviews,* 34,273-286.