Lecture Notes in Statistics 80

Michael A. Fligner
Joseph S. Verducci (Eds.)

Probability Models and Statistical Analyses for Ranking Data

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

Michael A. Fligner
Joseph S. Verducci
Department of Statistics
The Ohio State University
Columbus, OH 43210
USA

Mathematics Subject Classifications: 62-06, 60-06

Library of Congress Cataloging-in-Publication Data
Probability models and statistical analyses for ranking data / Michael
A. Fligner, Joseph S. Verducci (eds.).
p. cm. — (Lecture notes in statistics ; 80)
Includes bibliographical references.

1. Ranking and selection (Statistics)—Congresses. I. Fligner,
Michael A. II. Verducci, Joseph S. III. Series: Lecture notes in
statistics (Springer-Verlag) ; v. 80.
QA278.75.P76 1992
519.5—dc20 92-30305

Printed on acid-free paper.

Softcover reprint of the hardcover 1st edition 1993

Camera ready copy provided by the editors.

9 8 7 6 5 4 3 2 1

ISBN-13: 978-0-387-97920-5 e-ISBN-13: 978-1-4612-2738-0
DOI: 10.1007/978-1-4612-2738-0

Editorial Policy

for the publication of proceedings of conferences
and other multi-author volumes

Lecture Notes aim to report new developments - quickly, informally, and at a high level. The following describes criteria and procedures for multi-author volumes. For convenience we refer throughout to "proceedings" irrespective of whether the papers were presented at a meeting.

The editors of a volume are strongly advised to inform contributors about these points at an early stage.

§1. One (or more) expert participant(s) should act as the scientific editor(s) of the volume. They select the papers which are suitable (cf.§§2-5) for inclusion in the proceedings, and have them individually refereed (as for a journal). It should not be assumed that the published proceedings must reflect conference events in their entirety. The series editors will normally not interfere with the editing of a particular proceedings volume - except in fairly obvious cases, or on technical matters, such as described in §§2-5. The names of the scientific editors appear on the cover and title-page of the volume.

§2. The proceedings should be reasonably homogeneous i.e. concerned with a limited and well defined area. Papers that are essentially unrelated to this central topic should be excluded. One or two longer survey articles on recent developments in the field are often very useful additions. A detailed introduction on the subject of the congress is desirable.

§3. The final set of manuscripts should have at least 100 pages and preferably not exceed a total of 400 pages. Keeping the size below this bound should be achieved by stricter selection of articles and NOT by imposing an upper limit on the length of the individual papers.

§4. The contributions should be of a high mathematical standard and of current interest. Research articles should present new material and not duplicate other papers already published or due to be published. They should contain sufficient background and motivation and they should present proofs, or at least outlines of such, in sufficient detail to enable an expert to complete them. Thus summaries and mere announcements of papers appearing elsewhere cannot be included, although more detailed versions of, for instance, a highly technical contribution may well be published elsewhere later. Contributions in numerical mathematics may be acceptable without formal theorems/proofs provided they present new algorithms solving problems (previously unsolved or less well solved) or develop innovative qualitative methods, not yet amenable to a more formal treatment.

Surveys, if included, should cover a sufficiently broad topic, and should normally not just review the author's own recent research. In the case of surveys, exceptionally, proofs of results may not be necessary.

§5. "Mathematical Reviews" and "Zentralblatt für Mathematik" recommend that papers in proceedings volumes carry an explicit statement that they are in final form and that no similar paper has been or is being submitted elsewhere, if these papers are to be considered for a review. Normally, papers that satisfy the criteria of the Lecture Notes in Statistics series also satisfy this requirement, but we strongly recommend that each such paper carries the statement explicitly.

§6. Proceedings should appear soon after the related meeting. The publisher should therefore receive the complete manuscript (preferably in duplicate) including the Introduction and Table of Contents within nine months of the date of the meeting at the latest.

§7. Proposals for proceedings volumes should be sent to one of the editors of the series or to Springer-Verlag New York. They should give sufficient information on the conference, and on the proposed proceedings. In particular, they should include a list of the expected contributions with their prospective length. Abstracts or early versions (drafts) of the contributions are helpful.

To our parents

Preface

In June of 1990, a conference was held on Probablity Models and Statistical Analyses for Ranking Data, under the joint auspices of the American Mathematical Society, the Institute for Mathematical Statistics, and the Society of Industrial and Applied Mathematicians. The conference took place at the University of Massachusetts, Amherst, and was attended by 36 participants, including statisticians, mathematicians, psychologists and sociologists from the United States, Canada, Israel, Italy, and The Netherlands.

There were 18 presentations on a wide variety of topics involving ranking data. This volume is a collection of 14 of these presentations, as well as 5 miscellaneous papers that were contributed by conference participants.

We would like to thank Carole Kohanski, summer program coordinator for the American Mathematical Society, for her assistance in arranging the conference; M. Steigerwald for preparing the manuscripts for publication; Martin Gilchrist at Springer-Verlag for editorial advice; and Persi Diaconis for contributing the Foreword. Special thanks go to the anonymous referees for their careful readings and constructive comments. Finally, we thank the National Science Foundation for their sponsorship of the AMS-IMS-SIAM Joint Summer Programs.

Contents

Miscellanea

Conference Participants

June 8 - 13, 1990

Laura J. Adkins	Dept. of Statistics, Marshall University, Huntington, West Virginia
Mayer Alvo	Dept. of Mathematics, University of Ottawa, Ottawa, Canada
Douglas M. Andrews	Depts. of Mathematics and Computer Science, Wittenburg University, Springfield, Ohio
N. Balakrishnan	Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
Laurel A. Beckett	Department of Medicine, Harvard University, Boston, Massachusetts
Ulf Böckenholt	Dept. of Psychology, University of Illinois, Champaign, Illinois
Paul Cabilio	Dept. of Mathematics, Acadia University, Nova Scotia, Canada
Subhabrata Chakraborti	Dept. of Management Science and Statistics, University of Alabama, Tuscaloosa, Alabama
Ching-Yuan Chiang	Dept. of Mathematics and Computer Science, James Madison University, Harrisonburg, Virginia

Ayala Cohen	Department of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa, Israel
Hans Colonius	Dept. of Psychology, Purdue University, West Lafayette, Indiana
Douglas E. Critchlow	Dept. of Statistics, The Ohio State University, Columbus, Ohio
Marcel A. Croon	Dept. of Social Services, Tilburg University, The Netherlands
Edwin L. Crow	Institute for Telecommunication Sciences, U.S. Dept. of Commerce Boulder Laboratories, Boulder, Colorado
Persi W. Diaconis	Dept. of Mathematics, Harvard University, Cambridge, Massachusetts
E. Jacquelin Dietz	Department of Statistics, North Carolina State University, Raleigh, North Carolina
Paul D. Feigin	Department of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa, Israel
Michael A. Fligner	Dept. of Statistics, The Ohio State University, Columbus, Ohio
Rudy A. Gideon	Dept. of Mathematical Sciences, University of Montana, Missoula, Montana
S. Rao Jammalamadaka	Dept. of Statistics, University of California, Santa Barbara, California
Harry Joe	Dept. of Statistics, University of British Columbia, Vancouver, British Columbia, Canada
K. R. Kadiyala	Krannert Graduate School of Management, Purdue University, West Lafayette, Indiana

Peter McCullagh	Dept. of Statistics, University of Chicago, Chicago, Illinois
John I. Marden	Dept. of Statistics, University of Illinois, Champaign, Illinois
A. A. J. Marley	Dept. of Psychology, McGill University, Montreal, Canada
Sri Gopal Mohanty	Dept. of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
Kandethody M. Ramachandran	Dept. of Mathematics, University of South Florida, Tampa, Florida
Marco Scarsini	Dept. of Science Attvariali, Roma, Italy
Georgia Lee Thompson	Dept. of Statistical Science, Southern Methodist University, Dallas, Texas
Jack D. Tubbs	Dept. of Mathematics Science, University of Arkansas, Fayetteville, Arkansas
Adrina R. VanBlokland	Data Theory, Leiden University, Leiden, Netherlands
Joseph S. Verducci	Dept. of Statistics, The Ohio State University, Columbus, Ohio
John G. Wilson	Operations Research, Case Western Reserve University, Cleveland, Ohio
Tommy Wright	Mathematical Science Section, Oak Ridge National Laboratory, Oak Ridge, Tennessee
Damzngr Xu	Department of Mathematics, University of Oregon, Eugene, Oregon

Foreword

This book represents a coming of age for the statistical analysis of permutations. I would like to recapture some of the excitement and sheer amazement at the Amherst meeting as the participants realized that there were other people seriously working on the same problems with good new ideas. I'll also chronical an attempt to justify two models (The Luce Model and the Mallows Model through the Cayley distance) and suggest some open problems.

Over the years, there have been sporadic efforts to develop models and data analytic techniques for permutations. Noteworthy lines begin with psychological working models and the paired comparisons literature. Mathematical Statisticians have contributed a line through their work on rank tests. These lines formed smallish spirals with little interaction.

I began my work in this subject by realizing that most standard rank correlation methods could be restated as the computation of distances between permutations with naturally defined metrics. This led to a study of metrics on groups and a number of data analytic suggestions summarized in my book.

These metric ideas have been brilliantly developed in a series of papers by Critchlow, Fligner and Verducci. Complete references are included in their articles in this volume. Fresh work along these lines appears here for the first time in the papers of Marden and McCullagh with their co-authors. Paul Feigin explores the natural direction of two sample problems. The development, starting from Mallows' basic work to the present high-tech version, makes a lovely study in statistical evolution. They offer a developed set of tools for day-to-day problems.

One bottom line question: is it just "us" writing about these models or is someone actually using them? Laurie Beckett has been out there getting doctors to use them in Alzheimers studies. In my opinion, the single most important step to making our subject mainstream involves getting out in the world and applying the techniques in real problems.

In keeping with the new realism of statistics, I found some real ranked data sets. Honest analysis of real data can lead in all sorts of strange directions. One of these, Fourier analysis of ranked data, was developed following thesis work of Joe Verducci at Stanford. It formed the basis of my talk at Amherst and is published in Diaconis (1989).

Hal Stern's contribution to the present volume begins with one of these examples and runs with it, making the connection to voting paradoxes and completing the analysis in ways I wish I had done. The data are "data analyzed" by Georgia Thompson who has introduced much needed graphical techniques in her paper here and a more extensive account in Thompson (1992).

The conference had welcome representations of the paired comparison world in the work presented by Crow, by David and Andrews, and by Joe and Verducci. There has been too little interaction between this world and the current swirl of activity. Alvo and Cabilio provided our main link to rank tests in statistics. This is a rich source of possible applications and the frontier seems will open.

I think most of the statisticians present were surprised and delighted at how advanced the work of mathematical psychologists has become. Starting with primitives such as Thurstone latent parameter ranking models and Coombs unfolding hypothesis, these workers have developed versatile packages with associated interpretive languages. These are more than competitive with current versions of statisticians' models due to years of experience and available software. The papers by Bockenholt, Colonius, Croon and Luijkx, Marley, and van Blokland are wonderful introductions to what is available. I hope these authors were as favorably impressed with the statisticians efforts as I was with theirs.

These brief mentions are no substitute for a hands-on experience. Fortunately, the papers lie in wait beyond!

A Brief Study of Two Models

By now we realize it is easy to make up models. To understand and justify them is the next challenge. I want to record two efforts along these lines. Each concerns a family $P_\theta(\pi)$ of probability measures on the permutation group. Each family is shown to be the stationary distribution of a simple Markov chain.

These characterizations suggest natural mechanisms underlying the models. Such representations are also useful in using Stein's Method to study large n limits. For each family a sequential scheme exists for choosing π according to $P_\theta(\pi)$. This gives a second possible generating mechanism. These two generating mechanisms are used to study simple properties such as what is the chance that $\pi(1) = i$? How many cycles, fixed points, etc., can be expected?

I. The Luce Model. Let w_i, $1 \leq i \leq n$ be positive weights adding to 1. Use these to choose a permutation by assigning $\pi(1) = j$ with probability w_j; given $\pi(1) = j$, set $\pi(2) = h$ with probability $\frac{w_h}{1-w_j}$. Continue, sampling from the weights without replacement until a complete permutation has been formed. Here, the parameter space is Θ^+ is the open standard n simplex, and for

$$\theta = (w_i, \ldots, w_n) \in \Theta^+,$$

the Luce model is

$$P_\theta(\pi) = \prod_{i=1}^{n-1} \frac{w_{\pi(i)}}{\sum_{j=i}^{n} w_{\pi(j)}}.$$

This sequential generating mechanism is plausible – each item being ranked has a popularity. This was close to Luce's original motivation. On the other hand, people sometimes choose rankings by voting for popular items and against unpopular items. The reader who tries to calculate $P_\theta\{\pi : \pi(n) = i\}$ will see the difficulty of trying to use the Luce model for data with these characteristics.

It is possible to derive integral expressions for the probability of specified final places $\pi(n)\pi(n-1)$ using the representation of the Luce Model as the relative order of n independent exponential variables (See Yellott (1977)).

The Luce Model arises in a widely-used computer science algorithm for list management. This gives a Markovian description. Suppose you had n folders and used them with different frequencies. You want to arrange them so the most popular folder is on top, the next most popular folder next, and so on. If you don't know the popularity of the folders, it is natural to rearrange them by leaving the last used folder on top each time. If the frequency of use of folders i is w_i, this gives a Markov chain on the symmetric group with the Luce Model as its stationary probability. In Diaconis and Hanlon (1992b) a careful study of the rate of convergence of this chain to its stationary distribution is given. Phataford (1991) gives the eigenvalues and references to the computer science literature.

As a start to studying properties of this model, we note that for $k \leq n$, the chance that π begins $\pi(1), \pi(2), \ldots, \pi(k)$ is clearly

$$\prod_{i=1}^{k} \frac{w_{\pi(i)}}{\sum_{j=i}^{n} w_{\pi(j)}}.$$

This can be used to show that for large n, the first few coordinates are approximately independent with $P_\theta\{\pi(i) = j\} = w_j$. This assumes the weights are "not too wild." Rosen (1972) continues a careful development of asymptotic properties. His work was developed for problems in weighted survey sampling.

It is also possible to show that the P_θ chance that items $i_1, i_2, \ldots, i_k$ appear in the order i_1 before i_2 before ... before i_k is the stationary distribution restricted to this list, namely

$$\prod_{a=1}^{k-1} \frac{w_{i_a}}{\sum_{b=a}^{k} w_{i_b}}.$$

For example, the chance that i_1 appears before i_2 is $w_{i_1}/(w_{i_1} + w_{i_2})$; the chance of i_1 before i_2 before i_3 is $w_{i_1} w_{i_2}/\ (w_{i_1} + w_{i_2} + w_{i_3})(w_{i_2} + w_{i_3})$.

It is easy to show that P_θ is monotone in the weak Bruhat order. This is a partial order on permutations which has $\pi \succ \sigma$ if π can be moved to σ by interchanging pairwise adjacent items which are out of order. See Bjorner (1982). As far as I know even simple functional properties such as number of fixed points, cycles, or the average distance to the identity in various metrics is unknown.

The computer science literature on dynamic list management contains other material of interest. For example, Rivest (1976) modified the move to top heuristic to the following: at each time, an item is chosen with probability w_i and transposed with the item above it. If the top item is chosen, nothing changes. Rivest's work implies that the stationary distribution of π is proportional to $\prod_{i=1}^{n}\ w_i^{-\pi(i)}$. This gives a Markov chain method for generating from densities proportional to $e^{\eta \cdot \pi}$ for $\eta \in \Re^n$. In particular, choosing $\eta = (1, 2, 3 \ldots, n)$, it gives a stochastic interpretation of the Mallows model through the squared Spearman's rho distance. Van Leeuwen (1990) gives an up-to-date overview of the computer science literature.

II. The Mallows Model. Define a metric on permutations by $d(\pi, \sigma) =$ minimum number of transpositions required to bring π to σ. This Cayley distance seems to be the only reasonable distance invariant under relabelling on both sides $d(\pi, \sigma) = d(\pi\eta, \sigma\eta) = d(\eta\pi, \eta\sigma)$ for all π, σ, η. It is easy to calculate because of a relation discovered by Cayley: $d(\pi, \sigma) = n - c(\pi\sigma^{-1})$ with $c(\pi)$ the number of cycles in π.

Following Mallows (1957), a metric can be used to define a family of probabilities $P_\theta(\pi)$ on permutations by

$$P_\theta(\pi) = C(\theta)\theta^{d(\pi,\pi_0)} \qquad 0 < \theta \leq 1$$

with $c(\theta) = \prod_{i=1}^{n-1}\ (\theta_i + 1)^{-1}$ and π_0 a location parameter. This probability is largest when $\pi = \pi_0$ and falls off exponentially. When $\theta = 1$ it is the uniform distribution. These models are easy to write down and picture. In what follows, take $\pi_0 =$ identity without essential loss.

One mechanism that gives rise to P_θ is the Chinese Restaurant story. Picture a Chinese restaurant containing n large circular tables labelled $1, 2, \ldots, n$. People arrive at the restaurant and choose a table according to the following scheme: The first person to arrive sits at table 1. The second person to arrive sits to the right of the first person or at table 2 with respective probabilities $\theta/(1 + \theta)$ and $1/(1 + \theta)$. If j people have been seated, the $j + 1$st chooses to sit at an empty table with probability

$1/(j\theta + 1)$ and to the immediate right of a randomly chosen one of the previous j people with equal probability. This final arrangement around the tables is interpreted as a permutation in cycle notation. For example, if reading clockwise, the first table contains (1,3,5) and the second table contains (2,4) this is the permutation $\begin{smallmatrix}12345\\34521\end{smallmatrix}$.

The partition given by the cycles in this Mallows Model arises in Mathematical population genetics as the Ewens sampling formula. Aldous (1985) gives a splendid treatment.

Some properties of the family of probabilities are easy to derive from the restaurant description. For example $P_\theta\{\pi(i) = i\} = 1/[1 + (n-1)\theta]$, $P_\theta\{\pi(i) = j\} = \theta/[1+(n-1)\theta]$ for $j \neq i$. For a second set of properties, let $a_i(\pi)$ be the number of cycles of length i. Many properties of permutations can be described using the a_i, $1 \leq i \leq n$. For example, the number of fixed points in π is $a_1(\pi)$. The number of cycles in π is $a_1(\pi) + \ldots + a_n(\pi)$. The order of π is the smallest k so π^k = identity. This is the least common multiple of the i such that $a_i > 0$. Define a generating function as

$$f_n(x_1, \ldots, x_n) = E_\theta \prod x_i^{a_i(\pi)}.$$

Then, an easy variant of Polya's cycle index theorem shows that

$$\sum_{n=0}^{\infty} t^n f_n = \frac{1-\theta}{\theta t^{(1-\theta)/\theta}} \int_0^t \prod_{i=1}^{\infty} e^{x_i s^i/i\theta}(t-s)^{(1-2\theta)/\theta} ds.$$

From this, it is a straight forward matter to approximate the moments of a_i and move the following result.

Theorem. *As n tends to infinity, the joint distribution of $a_1, a_2, \ldots, a_k$ under P_θ tends to independent Poisson $(1/i\theta^i)$.*

A second line of development for this Mallows model is very generally applicable. This sees the Mallows model as the stationary distribution of a Markov chain on permutations. The Markov chain will be taken as the result of thinning down the chain resulting from random transposition by the Metropolis algorithm. The chain has a simple description: suppose the chain is currently at π. Randomly transpose two places. If this brings π closer to the identity, make the change. If it brings π further from the identity, flip a θ coin. If the coin comes up heads, make the change. If the coin comes up tails the chain stops at π. This chain has P_θ as its stationary distribution. The eigenvalues and rate of convergence to stationary are derived in Diaconis and Hanlon (1992a). There is a curious connection to the zonal polynomials of multivariate analysis.

Closing Remark

None of the justifications discussed above seems terribly natural. I mention them in the hope that they may trigger someone to do better. The program of developing these ideas for the other standard models seems substantial and worthwhile.

I am grateful to the organizers for making this conference take off.

Persi Diaconis
Dept. of Mathematics
Harvard University

REFERENCES

Aldous, D. (1985). Exchangeability and related topics. *Spring Lecture Notes in Mathematics.* **1117**.

Bjorner, A. (1982). Bruhat order of Coxeter groups and shellability. *Adv. Math* **43**, 87-100.

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

Diaconis, P. (1989) A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17**, 949-979.

Diaconis, P. and Hanlon, P. (1992a). Eigen analysis for some examples of the metropolis alogithm. To appear in (D. Richards, ed.) *Special Functions.*

Diaconis, P. and Hanlon, P. (1992b). Analysis of some dynamic list management algorithms. Technical report, Dept. of Mathematics, Harvard University.

Diaconis, P. and Hanlon, P. (1990). Efficient computation of the finite Fourier transform on finite groups. *Jour. Amer. Math. Society.* **3**, 297-332.

Diaconis, P. and Sturmfels, B. (1992). An application of Grobner bases to Monte Carlo algorithms. Technical report Dept. of Mathematics, Harvard University.

Mallows, C. (1957). Non-null ranking models I. *Biometrika* **44,** 114-130.

Phataford, R. (1991). On the matrix occurring in a linear search problem. *Jour. Appl. Prob.* **28,** 336-346.

Rivest, R. (1976). On self-organizing sequential search heuristics. *Comm. ACM* **19**, 63-67.

Rockmore, D. (1992) Efficient computation of Fourier inversion for finite groups. To appear, *Proc. A. C. M.*

Rosen, B. (1972) Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Am. Math. Statist.* **43**, 373-397; 748-776.

Thompson, G. (1992). Graphical methods for permutations. To appear, *Ann. Statist.*

Van Leeuwen, J. (1990) *Handbook of Theoretical Computer Science*, MIT Press: Cambridge.

Yellott, J. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement and the double exponential distribution. *Jour. Math. Psychol.* **15**, 109-144.

1

Ranking Models with Item Covariates

Douglas E. Critchlow[1]
Michael A. Fligner [2]

ABSTRACT Two parametric classes of ranking models are investigated: the Thurstone order statistics models and the Babington Smith models. Both families are natural extensions of commonly used paired comparison models. The concept of an "item parameter" is introduced and studied in the context of each of these classes of models. This distinction between the item parameters and the remaining parameters in a ranking model is useful not only for the general interpretation of model parameters, but also for the specific problem of introducing covariates in these models. Estimation schemes are described for these models, both with and without covariates, and are implemented in an example.

Key Words and Phrases: Permutation data, paired comparisons, order statistics model, iteratively reweighted least squares, item parameter.

1.1 Introduction

Probability models for the subjective comparison of several items are used in a variety of applications. The most widely studied class of such models has been developed for the paired comparison experiment. In the most basic paired comparison experiment, each judge considers several pairs of items, and states a preference for one item within each pair. Excellent introductions to the paired comparison literature are provided in Bradley [1, 2] and the monograph by David [10].

In this paper, the focus is on the *simultaneous* comparison of all of the items, which results in a complete ranking of the k items by each judge. However, for the case in which the number k of items is large, it may be necessary to rank the items within smaller subsets, yielding triple comparisons, quadruple comparisons, and so on. In either case, probability models for rankings of more than two items are needed. We consider two classes

[1] Department of Statistics, The Ohio State University
[2] Department of Statistics, The Ohio State University

of probability models for rankings that are natural extensions of some of the basic paired comparison models. The necessary background on paired comparison models is presented in the remainder of this section.

Suppose that the paired comparison experiment involves k items, and that n_{ij} independent comparisons are made for the pair of items (i, j). The basic parameters are then the $k(k-1)/2$ quantities p_{ij}, where for $i < j, p_{ij}$ denotes the probability that i is preferred to j in a comparison of these two items. In the simplest paired comparison models, it is assumed that ties are not permitted, and that the order of item presentation is unimportant.

The paired comparison models most commonly used in practice are the Thurstone [31] -Mosteller [22] model and the Bradley-Terry [3] model. Both of these models assume specific functional forms for the p_{ij}. For the Thurstone-Mosteller model, $p_{ij} = P(X_i < X_j)$, where X_i and X_j are independent normal random variables with common standard deviation σ and means μ_i and μ_j, respectively. The random variables X_i and X_j represent a random judge's perceptions of the merits of items i and j, and it is these unobservable random variables that determine the ordering of the items.

In the usual formulation of the Bradley-Terry model,

$$p_{ij} = p_i/(p_i + p_j), \tag{1}$$

where p_i is a non-negative parameter associated with item i, for $i = 1, \ldots, t$. Equivalently, $p_{ij} = P(X_i < X_j)$, where X_i has a Gumbel distribution $F(x) = 1 - \exp(-\exp(x - \mu_i))$, and $\mu_i = -\log p_i$. Both of these forms of the Bradley-Terry model will be used in the extensions to ranking models.

The next section discusses two methods of generalizing paired comparison models to rankings, along with a framework for interpreting the "item parameters" in these models. In Section 1.3 we extend these models to include covariates, and Section 1.4 considers estimation and other inferential procedures for the models. The paper concludes with an example in Section 1.5, and a discussion in Section 1.6.

1.2 Basic Ranking Models and Their Parameters

The two parametric families of ranking models discussed in this section are the Thurstone order statistics models and the Babington Smith models. The necessary notation for rankings is as follows. Once the k items are labelled arbitrarily as item 1 to item k, a *ranking* of the k items corresponds to a permutation function $\pi = (\pi(1), \ldots, \pi(k))$ from $\{1, \ldots, k\}$ onto $\{1, \ldots, k\}$, where $\pi(i)$ is the rank assigned to item $i, i = 1, \ldots, k$, and smaller ranks correspond to the more preferred items. An alternative description of any ranking π is given by the associated *ordering* of the k items, denoted by the bracketed vector $< i_1, i_2, \ldots, i_k >$, where i_r is the item assigned rank $r, r = 1, \ldots, k$. For example, $< 2\ 3\ 1 >$ denotes the ordering in which item 2 is ranked as best, item 3 is ranked as second best, and item 1 is ranked as worst. The composition of two rankings is defined by $(\pi \circ \sigma)(i) = \pi[\sigma(i)]$. A probability mass function $P(\pi)$ represents a probability model on rankings — or, more briefly, a *ranking model* — usually indexed by a finite set of parameters.

Before defining the two classes of ranking models, the concept of an "item parameter" is introduced. In any ranking model, certain parameters may reflect solely properties of the individual items being ranked, while other parameters may not be directly associated with individual items, but rather contain different types of information about the distribution of rankings. This distinction between item parameters and the remaining parameters is useful not only for the general interpretation of model parameters, but also for the specific problem of introducing covariates in these models, as described in Section 1.3.

The idea behind the formal definition of the item parameters $\nu_1, \nu_2, \ldots, \nu_k$ is that the probability distribution of the rankings should possess a natural invariance, with respect to these parameters, under an arbitrary relabeling of the k items. On the other hand, the remaining "non-item" parameters $\eta_1, \eta_2, \ldots, \eta_t$ will not exhibit this invariance. Specifically, assume that the probability distribution $P_{\boldsymbol{\theta}}(\pi)$ is indexed by the vector of $t + k$ parameters $\boldsymbol{\theta} = (\eta_1, \eta_2, \ldots, \eta_t, \nu_1, \nu_2, \ldots, \nu_k)$. Suppose the k items $1, \ldots, k$ are now relabeled as $\alpha(1), \ldots, \alpha(k)$, so that $\alpha(i)$ is the new label for item i. If this implies that the induced probability model $P_{\boldsymbol{\theta}^*}(\pi)$ has the parameter vector $\boldsymbol{\theta}^* = (\eta_1, \eta_2, \ldots, \eta_t, \nu_{\alpha^{-1}(1)}, \nu_{\alpha^{-1}(2)}, \ldots, \nu_{\alpha^{-1}(k)})$, where $(\alpha^{-1} \circ \alpha)(i) \equiv i$, then the parameters $\nu_1, \nu_2, \ldots, \nu_k$ are said to be *item parameters.* Intuitively, when the items are relabeled, their corresponding parameters must be relabeled accordingly. Further interpretations and properties of item parameters will be given later in this section, and will be illustrated in the context of the two classes of ranking models now introduced.

Daniels [9] suggests the following ranking model as a natural extension of the Thurstone paired comparison model. Suppose that the random variables $X_1, \ldots, X_k$ represent a random judge's perceptions of the merits of the k items, and that the relative ordering of these random variables determines his ordering of the items. Formally, let X_i be independent random variables with distributions $F_i(x) = F(x - \mu_i)$, where $i = 1, \ldots, k$ indexes items and F is a fixed but arbitrary continuous c.d.f. A random ranking π is defined by setting $\pi(i)$ equal to the rank, from smallest to largest, of X_i among $\{X_1, \ldots, X_k\}$. Thus, the ranking π with ordering $< i_1, \ldots, i_k >$ is assigned the probability

$$P(\pi) = P\{X_{i_1} < \ldots < X_{i_k}\}. \tag{2}$$

The probability model on rankings defined by (2) is referred to as a *Thurstone order statistics model* (see, for example, Yellott [33]).

The parameters $\mu_1, \mu_2, \ldots, \mu_k$ are clearly item parameters according to the above definition. Since smaller ranks correspond to the more preferred items, note that for a Thurstone order statistics model, those items with the smaller values of μ_i are the "best" items. This convention is maintained for the item parameters in the remaining classes of models.

Two well-studied cases of the Thurstone model are the Thurstone [31] - Mosteller [22] - Daniels [9] model (*TMD model*) where F is the standard normal distribution, and the *Luce model* [16], where F is Gumbel. Brook and Upton [5] use the TMD model to analyze voters' rankings of candidates for public office. The Luce model has a simple expression for the ranking probabilities given in (2) (see, for example, Plackett [23]). Further properties and references for these models can be found in Yellott [33], Stern [30], and Critchlow, Fligner, and Verducci [8].

The second class of ranking models is suggested by Babington Smith [10]. He uses a conditioning argument for inducing a probability model on rankings from a set of arbitrary paired comparison probabilities. For each pair of items $i < j$, let p_{ij} be the probability that item i is preferred to item $j (i \longrightarrow j)$ in a paired comparison of these two items. Imagine a tournament in which all of the $k(k-1)/2$ possible paired comparisons are made independently. If the results of this tournament contain no circular triads $(h \longrightarrow i \longrightarrow j \longrightarrow h)$, then the tournament corresponds to a unique ranking π of the items; otherwise the entire tournament is repeated until a unique ranking is obtained. The probability of any resulting π is thus given by

$$P(\pi) = \text{ constant } \prod_{\{(i,j):\pi(i)<\pi(j)\}} p_{ij}, \tag{3}$$

where p_{ij} for $i > j$ is defined by $p_{ij} \equiv 1 - p_{ji}$, and the constant is chosen to make the probabilities sum to 1. Although the Babington Smith model is indexed by $k(k-1)/2$ parameters, constraints on the $\{p_{ij}\}$ proposed by Mallows [17] lead to three important subclasses of the general Babington Smith model (3), that are called the Mallows-Bradley-Terry model, the rho-based model, and the ϕ-model, respectively.

The general Babington Smith model does not contain item parameters. Mallows [17] suggests a way of both reducing the number of parameters in (3) and introducing item parameters. He assumes that the paired comparison probabilities have the Bradley-Terry form:

$$\text{logit } p_{ij} = \gamma_j - \gamma_i$$

for some nonnegative parameters $\gamma_1, \ldots, \gamma_k$, where γ_i = -ln p_i in (1). This leads to the *Mallows-Bradley-Terry (MBT) ranking model*

$$P(\pi) = C(\gamma) \prod_{r=1}^{k-1} \exp(-(k-r)\gamma_{i_r}), \tag{4}$$

for any ranking π with associated ordering $< i_1, \ldots, i_k >$, where $\gamma = (\gamma_1, \ldots, \gamma_k)$, and $C(\gamma)$ is chosen to make the probabilities sum to 1. The parameters $\gamma_1, \ldots, \gamma_k$ satisfy the defining condition for item parameters. Moreover, for the parameterization used, smaller values of γ_i correspond to more preferred items, just as for the Thurstone order statistics model.

Mallows [17] suggests a further simplification of the MBT model, which assumes that there is a modal ranking $\pi_0 = (\pi_0(1), \ldots, \pi_0(k))$. This modal ranking corresponds to a vector of parameters $(\pi_0(1), \ldots, \pi_0(k))$, where $\pi_0(i)$ is the rank assigned to item i by the modal ranking. Fixing $\theta \in (0,1)$ and letting $\gamma_i = -2\pi_0(i) \ln\theta$ in (4) then gives Mallows' *rho-based model*

$$P(\pi|\theta, \pi_0) = C(\theta)\theta^{R^2(\pi,\pi_0)}, \qquad (5)$$

where

$$R^2(\pi, \pi_0) = \sum_{i=1}^{k} [\pi(i) - \pi_0(i)]^2$$

is the Spearman's rho-distance between the rankings π and π_0, and is related to Spearman's [28] correlation coefficient. The rho-based model has the interpretation that the probability $P(\pi)$ decreases geometrically, as the R^2 distance from π to π_0 increases. According to the item parameter definition, the components $\pi_0(1), \ldots, \pi_0(k)$ of π_0 are item parameters. On the other hand, θ is a dispersion parameter: as $\theta \longrightarrow 1$, the rho-based model approaches the uniform distribution on all $k!$ possible rankings, whereas when $\theta \longrightarrow 0$, the model becomes increasingly concentrated about the modal ranking π_0.

For this model, the item parameters $\pi_0(1), \ldots, \pi_0(k)$ have an important distinction from the item parameters of both the Thurstone order statistics model and the general MBT model. Namely, the item parameters take their values in the *discrete* parameter space consisting of all $k!$ permutations of the integers $1, \ldots, k$. This distinction will be important in the covariate models of Section 1.3, since covariates are more easily incorporated into models that have *continuous* item parameters.

Suppose that π_0 is not restricted to be an actual ranking, but rather that its components are allowed to be arbitrary real numbers in (5). It has been noted by McCullagh [18], among others, that the resulting model is equivalent to the MBT model of (4).

The final simplification of the general Babington Smith model is Mallows' well-known *ϕ-model*. Unlike the rho-based model, the ϕ-model is not an MBT model. To describe the ϕ-model, fix $\phi \in (0,1)$, and suppose that in (3) the corresponding paired comparison probabilities have the simple form $p_{ij} = (\phi + 1)^{-1} > .5$ for $\pi_0(i) < \pi_0(j)$, so that the probability that item i is preferred to item j depends only upon their relative order in the modal ranking. Then the resulting ranking model is

$$P(\pi|\phi, \pi_0) = C(\phi)\phi^{T(\pi,\pi_0)}, \qquad (6)$$

where

$$T(\pi, \pi_0) = \sum_{i<j} I\{[\pi(i) - \pi(j)][\pi_0(i) - \pi_0(j)] < 0\}$$

is the Kendall's tau-distance between the rankings π and π_0, and $I(\cdot)$ is the indicator function: $I(A) = 1$ if the event A occurs, and $= 0$ otherwise. The ϕ-model has the interpretation that the probability $P(\pi)$ decreases geometrically according to increasing tau-distance from π to the modal ranking π_0. The components $\pi_0(1), \ldots, \pi_0(k)$ of π_0 are again discrete item parameters, and ϕ is a dispersion parameter.

In the remainder of this section, several properties of the item parameters $\nu_1, \ldots, \nu_k$ in a ranking model are considered. Specifically, for many models, both the ordering and the spacing of the ν_i contain useful information regarding the items. With regard to the ordering of the ν_i, suppose that the ν_i are distinct, and consider two items i and j for which $\nu_i < \nu_j$. For the given ranking model, suppose that $P[\pi(i) < \pi(j)]$, the probability that item i is preferred to item j, exceeds .5. If this probability continues to exceed .5 given any fixed assignment of ranks to the other $k-2$ items, then item i is *strongly preferred* to item j. If item i is strongly preferred to item j for *every* pair of items i and j with $\nu_i < \nu_j$, then the ranking model is said to have a complete consensus, with consensus ordering determined by the ordering of the ν_i.

To formally define complete consensus, let the transposition permutation τ_{ij} be defined by $\tau_{ij}(i) = j, \tau_{ij}(j) = i$, and $\tau_{ij}(m) = m$ for all $m \neq i, j$. Note that $\pi \circ \tau_{ij}$ is the ranking that agrees with π except that the ranks assigned to items i and j are exchanged. A model has the property of *complete consensus*, with consensus ordering determined by the ν_i, if for every pair of items i and j such that $\nu_i < \nu_j$, and any permutation π such that $\pi(i) < \pi(j), P(\pi) > P(\pi \circ \tau_{ij})$. The notion of complete consensus is discussed by Henery [13] and Fligner and Verducci [11].

Although complete consensus is a property that orders the items in a fairly strong sense, it is satisfied by many ranking models. For a Thurstone order statistics model, if the μ_i are distinct and the likelihood ratio $\frac{F'(x-\mu_i)}{F'(x-\mu_j)}$ is a non-increasing function of x for $\mu_i < \mu_j$, Henery [13] shows that the model has complete consensus (see also Savage [24, 25]). Moreover, the complete consensus property is also shared by the three subclasses of the Babington Smith model that have item parameters: the MBT model, the rho-based model, and the ϕ-model (see Critchlow, Fligner, and Verducci [8]).

A consequence of the complete consensus property is that the items are *ordered in expectation*: $\nu_i < \nu_j$ implies $E[\pi(i)] < E[\pi(j)]$ (Fligner and Verducci [11]). In fact, the ordered in expectation property holds for all Thurstone order statistics models, not just those with a monotone likelihood ratio density.

The complete consensus and ordered in expectation properties are ordinal properties; that is, they continue to hold under an arbitrary, strictly increasing reparameterization of the ν_i. A final property of the item parameters in a ranking model involves the spacing of the ν_i as well as their ordering. The idea is that for any items i and j, the probability that i is ranked ahead of j depends only on the parameters ν_i and ν_j associated with these two items, and further is a strictly increasing function of the signed interval length $\nu_j - \nu_i$.

Formally, a model is said to be *interval scaled* if $P(\pi(i) < \pi(j)) = f(\nu_j - \nu_i)$, for some strictly increasing function f that may depend upon the other "non-item" parameters in the model. Note that the interval scaled property depends not just on the ranking probabilities, but also on the particular choice of parameterization of the model in terms of its item parameters ν_i. Thus, if a model is not interval scaled under a particular parameterization, this suggests finding a strictly increasing reparameterization $\xi_i = g(\nu_i)$ for which it is interval scaled, i.e. $P(\pi(i) < \pi(j)) = f(\xi_j - \xi_i)$. For models having continuous item parameters, if there exists such a reparameterization function g, it is uniquely determined up to an affine transformation.

Both the Thurstone order statistics models and the ϕ-model are interval scaled, under the parameterizations given. For the Thurstone models, this follows from the relation $P(\pi(i) < \pi(j)) = P(X_i < X_j) = D(\mu_j - \mu_i)$, where D is the c.d.f. of $Z_1 - Z_2$, and Z_1 and Z_2 are i.i.d. with c.d.f. F. The interval scaled property of the ϕ-model is proved by Mallows [17], Section 9.

It is now shown that the MBT model is not interval scaled under the given parameterization (4), and in fact is not interval scaled under any reparameterization. The proof is by contradiction: suppose that it is interval scaled under the strictly increasing reparameterization $\xi_i = g(\gamma_i)$. Fix $\xi_1 = g(\gamma_1)$ and $\xi_2 = g(\gamma_2)$ such that $\xi_1 < \xi_2$ (and therefore $\gamma_1 < \gamma_2$). We show that the probability $P(\pi(1) < \pi(2))$ depends not only on ξ_1 and ξ_2, but also on $\xi_3, \ldots, \xi_k$, contradicting the interval scaled property. First, it is easy to check that as $\xi_3, \ldots, \xi_k$ increase in such a way that the corresponding $\gamma_3, \ldots, \gamma_k \to \infty$, then

$$\frac{P(\pi(1) < \pi(2))}{P(\pi(2) < \pi(1))} \to \exp(\gamma_2 - \gamma_1),$$

since in the limit, the probability that item 1 is preferred to item 2 in the full ranking is the same as the probability that 1 beats 2 in a paired comparison. On the other hand, suppose ξ_3 is such that $\gamma_3 = (\gamma_1 + \gamma_2)/2$, and suppose that $\xi_4, \ldots, \xi_k$ are still increasing so that $\gamma_4, \ldots, \gamma_k \to \infty$. Then a straightforward calculation yields

$$\frac{P(\pi(1) < \pi(2))}{P(\pi(2) < \pi(1))} \to \exp(\gamma_2 - \gamma_1) \left\{ \frac{\lambda^2 + 2\lambda}{1 + 2\lambda} \right\},$$

where $\lambda = \exp((\gamma_2 - \gamma_1)/2) > 1$. Thus the odds that item 1 precedes item 2 in the ranking have increased, showing that the probability that item 1 is preferred to 2 depends upon the other parameters.

Finally, the rho-based model (with item parameters $\pi_0(i)$) is interval scaled for $k = 3$, but for $k = 4$ a direct calculation shows that when $\pi_0 =< 1, 2, 3, 4 >$,

$$P(\pi(2) < \pi(3)) - P(\pi(1) < \pi(2)) = C(\theta)\theta^4(1-\theta^2)(1-\theta^4)(1-\theta^6) > 0,$$

contradicting the interval scaled property.

1.3 Ranking Models with Covariates

The models of Section 1.2 allow for a comparison of the items through the values of their item parameters. However, in many instances additional measurements about the items are available, corresponding to the values of one or more independent variables, or *item covariates*. In the example of Section 1.5, the items are four formulations of a salad dressing that are ranked according to tartness. Item covariates are available since the formulations differ with respect to the amounts of acetic and gluconic acid present. It is of primary interest to determine how these variables affect perceived tartness. To analyze the data fully, it is necessary to develop ranking models that incorporate the effects of such item covariates, and to study the associated inferential techniques.

For ranking models with continuous item parameters ν_i, we will assume a specific functional relationship between the ν_i and the item covariates. This introduces a more restrictive model, and allows us to determine the extent to which the value of each item parameter ν_i can be modelled in terms of its item covariates.

A linear model for the ν_i in terms of item covariates is given by

$$\boldsymbol{\nu} = \boldsymbol{C\beta}, \tag{7}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)'$, $\boldsymbol{C}$ is a $k \times p$ matrix whose i-th row is a vector of covariates associated with item i, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. For example, in ranking the four formulations of salad dressing, the two covariates of interest are c_1, the concentration of acetic acid, and c_2, the concentration of gluconic acid. An initial model which incorporates these item covariates is $\nu_i = \alpha + \beta_1 c_{i1} + \beta_2 c_{i2}$, where c_{i1} and c_{i2} are the concentrations of acetic acid and gluconic acid, respectively, associated with the i-th formulation. However, for the models considered in this paper, α can be omitted, since all values of α give the same ranking probabilities. Thus, for this example, model (7) has $\boldsymbol{\beta} = (\beta_1, \beta_2)$, and $\boldsymbol{C}$ is a 4×2 matrix with i-th row $(c_{i1}, c_{i2}), i = 1, \ldots, 4$. In Sections 1.4 and 1.5, inferential procedures are described to investigate the adequacy of the model, as well as the necessity of including both covariates to explain the differences in perceived tartness.

The idea of using a response surface model such as (7) in modelling preferences is considered, among others, by Springall [29] and Kousgaard [15] for paired comparison experiments, and by Hausman and Wise [12], McFadden [20], and Kamakura and Srivastava [14] for probabilistic choice models. However, its application to modelling the continuous item parameters in a ranking model appears to be new. An interesting problem, not treated in this paper, is to develop useful methods for including item covariates in models having *discrete* item parameters, such as (5) and (6).

1.4 Estimation

In this section, procedures are described for the estimation of the model parameters, for both the MBT and the Thurstone order statistics ranking models. Since both of these models have continuous item parameters, item covariates can be easily included in the models, as in (7). Thus, the associated parameter estimation procedures will be considered for models both with and without item covariates. In brief, the MBT model is a generalized linear model (GLM), and maximum likelihood estimates can often be obtained using existing statistical packages. On the other hand, maximum likelihood estimation for the Thurstone order statistics models requires direct evaluation of the log likelihood, usually via a multivariate integration procedure. The subsequent maximization of the log likelihood utilizes a $k-1$ dimensional search algorithm. A simple alternative to maximum likelihood estimation for the Thurstone order statistics models is also described.

In any ranking model, suppose that the $k!$ possible rankings are listed in some definite order, and that $\pi_j = (\pi_j(1), \ldots, \pi_j(k))$ corresponds to the j-th ranking in this list. If n judges independently rank the k items, the data vector is then $\boldsymbol{Y} = (Y_1, \ldots, Y_{k!})$, where Y_j is the number of times the j-th ranking π_j occurs in the data set. The probability distribution of $\boldsymbol{Y}$ is multinomial with n trials. Different ranking models then place varying structures on the multinomial cell probabilities.

First consider the MBT model. The logarithms of the multinomial cell probabilities are linear in the item parameters, so that the MBT model is a GLM (see McCullagh and Nelder [19]). Specifically, $\log(E(\boldsymbol{Y}/n)) = \boldsymbol{X}\boldsymbol{\gamma} + a\boldsymbol{1}$, where a is a scalar normalizing constant, $\boldsymbol{1}$ is a vector of 1's, and by (4), the j-th row of the matrix $\boldsymbol{X}$ has entries $\pi_j(1)-k, \ldots, \pi_j(k)-k$, for $j = 1, \ldots, k!$. In the case of the item covariate model (7), $\log(E(\boldsymbol{Y}/n)) = \boldsymbol{X}\boldsymbol{C}\boldsymbol{\beta} + a\boldsymbol{1}$, and the model is again a GLM. Thus the MBT model, both with and without covariates, can be fitted easily for small to moderate values of k, using the GLIM or S-plus statistical packages, for example. Parameter estimates and their standard errors are included in the output for these packages, as well as likelihood ratio test statistics for hypotheses of interest. In the example of the next section, the numerical results for the MBT model were obtained using GLIM. Further details are provided in Critchlow and Fligner [11].

For the Thurstone order statistics models, two estimation schemes are considered. The first is based on decomposing each observed ranking into its $k(k-1)/2$ induced paired comparisons. The resulting estimators are very simple to compute, and have good efficiency relative to the maximum likelihood estimators. The second scheme utilizes maximum likelihood estimation.

To begin with, note that adding a fixed constant to each of the μ_i does not change the ranking probabilities in a Thurstone order statistics model. Thus, the model actually has only $k-1$ estimable parameters, which can be taken as the linear contrasts $\mu_1 - \mu_k, \ldots, \mu_{k-1} - \mu_k$. Without loss of generality, we set $\hat{\mu}_k = 0$ and estimate the components of the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{k-1})$. For both parameter estimation schemes, it can be checked that the resulting estimates are suitably invariant under an arbitrary reparameterization of the $k-1$ contrasts.

The estimation scheme based on paired comparisons is a large sample method which proceeds as follows. For $i < j$, let N_{ij} count the number of times that item i is ranked before item j in the set of n observed rankings, and let $\hat{p}_{ij} = \frac{N_{ij}}{n}$ be the corresponding sample proportion. As in Section 1.2, let F be the c.d.f. of each $X_i - \mu_i$, and let D be the c.d.f. of $Z_1 - Z_2$, where Z_1 and Z_2 are i.i.d. with c.d.f. F. Since $\hat{p}_{ij}$ is asymptotically normal with mean $D(\mu_j - \mu_i)$, the δ-method shows that $D^{-1}(\hat{p}_{ij})$ has a limiting normal distribution with mean $\mu_j - \mu_i$. The estimation procedure utilizes the vector $\boldsymbol{V} = (D^{-1}(\hat{p}_{12}), D^{-1}(\hat{p}_{13}), \ldots, D^{-1}(\hat{p}_{1k}), D^{-1}(\hat{p}_{23}), \ldots, D^{-1}(\hat{p}_{k-1,k}))'$, which has an approximate ($n \to \infty$) multivariate normal distribution with mean vector $\boldsymbol{X\mu}$, and covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ also depends on the μ_i. The expression for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\mu})$ is given in the Appendix, and the form of the $\boldsymbol{X}$ matrix in the limiting mean vector is as follows. Note that since the entries of $\boldsymbol{V}$ are actually indexed by pairs (i, j), so are the rows of $\boldsymbol{X}$. The matrix $\boldsymbol{X}$ has $k(k-1)/2$ rows and $k-1$ columns, and the m-th entry in the (i, j)-th row of $\boldsymbol{X}$ is -1 if $m = i$, 1 if $m = j$, and 0 otherwise.

Thus, the asymptotic distribution of the random vector $\boldsymbol{V}$ is a linear model in $\boldsymbol{\mu}$, whose covariance structure also depends on $\boldsymbol{\mu}$. The $\boldsymbol{\mu}$ vector can therefore be estimated by iteratively reweighted least squares. Specifically, to implement the $(s+1)$-th stage of the iteratively reweighted least squares procedure, suppose that parameter estimates $\hat{\boldsymbol{\mu}}_{(s)}$ are available from the s-th stage, and let $\hat{\boldsymbol{\Sigma}}_{(s+1)} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\mu}}_{(s)})$ be the covariance matrix determined by $\hat{\boldsymbol{\mu}}_{(s)}$, as described in the Appendix. The new estimate $\hat{\boldsymbol{\mu}}_{(s+1)}$ of $\boldsymbol{\mu}$ is then obtained by substituting $\hat{\boldsymbol{\Sigma}}_{(s+1)}$ into the standard weighted least squares equation:

$$\hat{\boldsymbol{\mu}}_{(s+1)} = (\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{(s+1)}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_{(s+1)}^{-1}\boldsymbol{V}. \tag{8}$$

In the case of the item covariate model (7), the limiting mean vector is just $\boldsymbol{XC\beta}$, and iteratively reweighted least squares can again be used to estimate the β_i, by replacing $\hat{\boldsymbol{\mu}}_{(s+1)}$ with $\hat{\boldsymbol{\beta}}_{(s+1)}$ and $\boldsymbol{X}$ with $\boldsymbol{XC}$ in (8). Preliminary work indicates that this estimation technique has good efficiency relative to maximum likelihood estimation.

Brady [4] uses the estimates $\hat{p}_{ij}$ of the binary choice probabilities directly, rather than the $D^{-1}(\hat{p}_{ij})$, to estimate the parameters in the TMD model. This results in a nonlinear least squares problem, instead of the simpler linear approach considered here. The $\hat{p}_{ij}$ are also utilized by Cohen and Mallows [6], but for the problem of assessing the goodness of fit of a ranking model.

Finally, the second estimation scheme described for Thurstone order statistics models is maximum likelihood estimation. The log likelihood is

$$A + \sum_{j=1}^{k!} Y_j \log(P(\pi_j)),$$

where A does not depend on the parameter vector $\boldsymbol{\mu}$. This requires evaluation of each $P(\pi_j)$, usually by a numerical multiple integration technique. Note that for any ranking π with ordering $< i_1, \ldots, i_k >$, the probability $P(\pi) = P\{X_{i_1} < \ldots < X_{i_k}\}$ can be thought of as the probability that all the components of the $k-1$ dimensional random vector $(X_{i_2} - X_{i_1}, X_{i_3} - X_{i_2}, \ldots, X_{i_k} - X_{i_{k-1}})$ are positive. When F is normal this can be computed using specialized integration routines. For example, the Fortran program by Schervish [26] evaluates multivariate normal probabilities for an arbitrary correlation structure. Note that for small values of k, or for distributions F other than the normal, general multiple integration routines can be employed, such as those in IMSL. Finally, for the Luce model where F is Gumbel, the log likelihood can be expressed in closed form.

The maximization of the likelihood requires a $k-1$ dimensional search, which can be carried out using the quasi-Newton method and a finite difference gradient. Such algorithms are widely available in IMSL, for example, and do not require an evaluation of derivatives.

The overall performance of this estimation scheme for large values of k needs further investigation, especially with regard to its accuracy and required CPU time. In the next section, an example with $k = 4$ items and covariates is presented, that illustrates all of the above procedures.

1.5 Example

The TMD model (the Thurstone order statistics model with F the standard normal distribution), and also the MBT model, are now used to analyze a set of ranking data collected by Vargo [32]. Each of thirty-two judges was asked to rank four salad dressing preparations according to tartness, with a rank of 1 being assigned to the formulation judged to be the least tart. The raw data are given in the first two columns of Table 1, where the observed rank vector π_j in column 1 contains the ranks given to the four salad dressing preparations. Column 2 gives the observed frequencies Y_j, while columns 3 and 4 give the fitted Y values for the MBT and the TMD models, respectively, when the models are fitted using maximum likelihood estimation. (Column 5 gives the fitted Y values under the iteratively reweighted least squares (IRLS) estimation scheme for the TMD model, which is discussed later.)

TABLE 1
Salad dressing rankings with observed and fitted frequencies

Observed π_j	Frequency (Y_j)	$\hat{Y}_j$ for MBT	$\hat{Y}_j$ for TMD (MLE)	$\hat{Y}_j$ for TMD (IRLS)
(4,3,2,1)	2	0.292	0.256	0.230
(3,4,2,1)	1	1.658	1.601	1.576
(3,4,1,2)	2	1.137	1.124	1.089
(3,2,4,1)	1	0.332	0.324	0.290
(2,4,3,1)	2	4.213	4.043	4.144
(2,4,1,3)	1	1.980	1.863	1.832
(2,1,3,4)	1	0.121	0.126	0.107
(1,4,3,2)	11	7.337	8.010	8.400
(1,4,2,3)	6	5.030	5.339	5.469
(1,3,4,2)	3	3.281	3.108	3.083
(1,2,4,3)	1	1.006	1.035	0.966
(1,2,3,4)	1	0.689	0.738	0.680
	32	27.076	27.567	27.866

For both the TMD and MBT models, hypothesis testing can be done using likelihood ratio test statistics. For each model fitted under maximum likelihood, the deviance is reported in Table 2. The deviance is just the likelihood ratio test statistic $-2\log\lambda$, where λ is the likelihood ratio for testing a particular model against the general multinomial model with $k!-1$ parameters. Thus, the deviance equals

$$2\sum_{j=1}^{k} Y_j \log(Y_j/\hat{Y}_j),$$

where $\hat{Y}_j$ is the fitted count for the j-th ranking under a particular model. As illustrated below, differences in deviances give likelihood ratio test statistics for hypotheses of interest.

The first question of interest concerns the overall fit of the models. The deviances for both the TMD and MBT models, with arbitrary item parameters, are given in the second column of Table 2. To discuss these models simultaneously, both the item parameters μ_i for the TMD model, as well as the item parameters γ_i for the MBT model, will be denoted by the generic symbols $\nu_1, \ldots, \nu_k$ for item parameters.

TABLE 2
Deviances for various TMD and MBT models

Model	ν_1, ν_2, ν_3 arbitrary	$\nu_i = \beta_1 c_{i1}$ $+\beta_2 c_{i2}$	$\nu_i = \beta_1 c_{i1}$	$\nu_i = \beta_2 c_{i2}$
TMD	20.634(20 df)	21.026(21 df)	68.730(22 df)	59.390(22 df)
MBT	22.249(20 df)	22.247(21 df)	68.333(22 df)	59.642(22 df)

The deviances in the first column show that either of the two models adequately explains the data. Due to the sparseness of the data, the deviances may need to be interpreted with caution. However, the fitted values in Table 1 also suggest a satisfactory fit of each of the models, which can be confirmed by an examination of the standardized residuals. Since the TMD and MBT models are not nested, they cannot be compared easily, although the TMD model appears to fit slightly better.

The deviance for the uniform model, corresponding to $\nu_1 = \ldots = \nu_k$, is 70.753. Thus, for the TMD model, the likelihood ratio statistic for testing $H_0 : \nu_1 = \ldots = \nu_k$ versus $H_1 : \nu_i$ arbitrary is 70.753 - 20.634 = 50.119, with $23 - 20 = 3$ df. This indicates that the formulations are perceived differently by the judges with regard to their tartness ($p < .001$). The same conclusion is reached in the context of the MBT model.

In this example, the differences among the four salad dressing formulations are actually due to varying concentrations of acetic and gluconic acid. The acetic and gluconic acid concentrations for the four salad dressings are $(.5, 0)$, $(.5, 10.0)$, $(1.0, 0)$, and $(0, 10.0)$, respectively, where the first number in each pair is the percentage of acetic acid and the second is the percentage of gluconic acid. The deviances for the models incorporating both of these covariates as in (7) are given in the second column of Table 2, while the last two columns correspond to models with only one covariate.

The TMD model including both covariates has a deviance of 21.026 with 21 df. Hence, the likelihood ratio test statistic for testing the more restrictive model with covariates versus the model without covariates is 21.026 - 20.634 = 0.392 with 21 - 20 = 1 df. Therefore, the model with both covariates is not significantly worse than the model with arbitrary item parameters, and provides a suitable model for the data. As can be seen from the deviances for each of the models with a single covariate, neither covariate can be dropped from the model. Using the last line in Table 2, similar conclusions are reached for the MBT model.

The remainder of this section compares the two estimation schemes that were proposed in Section 1.4 for the TMD model. Table 3 provides the parameter estimates using both schemes, for the TMD model with the

ν_i arbitrary, as well as for the model containing both covariates. The parameter estimates in Table 3 appear quite similar. Moreover, the resulting models are also very close, as can be seen from a comparison of the fitted Y values under both schemes, given in the last two columns of Table 1.

TABLE 3
Parameter estimates for two TMD models, using two estimation schemes

	Maximum likelihood estimates	IRLS estimates using paired comparisons
$(\hat{\nu}_1, \hat{\nu}_2, \hat{\nu}_3)$ when ν_i arbitrary	(-0.755, 1.527, 0.498)	(-0.774, 1.599, 0.521)
$(\hat{\beta}_1, \hat{\beta}_2)$ when $\nu_i = \beta_1 c_{i1} + \beta_2 c_{i2}$	(2.773, 0.228)	(2.913, 0.238)

For the TMD model, the equation obtained from fitting the covariate model (7), via maximum likelihood estimation, is

$$\hat{\nu}_i = 2.773 A_i + 0.228 G_i, \tag{9}$$

where A_i and G_i are the concentrations of acetic and gluconic acid for the i-th formulation. Similar results are obtained using the MBT model, as described in Critchlow and Fligner [7].

1.6 Discussion

Although the experimental situation described in this paper results in data in the form of rankings, the approach is parametric; specific probability models are developed for the ranking process. The classical "nonparametric" approach for the comparison of items in such a ranking experiment utilizes the average of the ranks received by each item. These two points of view are now compared, and some advantages of the more complicated parametric approach are discussed.

Let $\bar{\pi}(i)$ denote the average rank received by the i-th item, and let $\bar{\pi} = (\bar{\pi}(1), \ldots, \bar{\pi}(k))$ be the vector of average ranks. For the MBT model, the vector $\bar{\pi}$ is a sufficient statistic. Moreover, there is an interesting relationship between the average rank vector and $\hat{\gamma}$, the vector of maximum likelihood estimates of the item parameters – namely, the ordering of the items according to $\bar{\pi}$ and $\hat{\gamma}$ is the same. Indeed, for the MBT model, the

$$\text{log likelihood } = B(\gamma_1, \ldots, \gamma_k) + n \sum_{j=1}^{k} \bar{\pi}(i)\gamma_i,$$

where $B(\gamma_1, \ldots, \gamma_k)$ is a symmetric function of $\gamma_1, \ldots, \gamma_k$, from which it follows directly that the MLE's $\hat{\gamma}_i$ and the $\bar{\pi}(i)$ must be in the same order.

On the other hand, for the TMD model, $\bar{\pi}$ is not a sufficient statistic, and for small samples, the previous correspondence between the MLE's of the item parameters and the vector of average ranks does not necessarily hold. However, this correspondence does hold asymptotically. This is an easy consequence of the complete consensus property of the TMD model, which, as mentioned in Section 1.2, ensures that the items are ordered in expectation, i.e. the ordering of the ν_i agrees with that of the $E[\pi(i)]$.

For either the TMD or MBT models, in order to achieve a direct numerical comparison of the average rank vector and the vector of parameter estimates, it is best to have these vectors satisfy the same linear restriction. Such a linear constraint can be imposed, because for either model, the ranking probabilities do not change when a fixed constant is added to each of the item parameters. Since $\sum_{j=1}^{k} \bar{\pi}(i) = k(k+1)/2$, it is assumed that the item parameters also sum to this constant. For example, imposing this constraint in (9) yields

$$\hat{\nu}_i = 2.773A_i + 0.228G_i - .027.$$

The vector of average ranks and vector of estimated $\hat{\nu}_i$ are then (1.56, 3.56, 2.69, 2.19) and (1.36, 3.64, 2.75, 2.25), respectively.

Although the two vectors are quite close, it is difficult to use the average rank vector in a nonparametric manner to make suitable inferences about the items. Standard nonparametric theory would declare two items different whenever $|\bar{\pi}(i) - \bar{\pi}(j)| > k$ (see, for example, Miller [21], page 174). However, the classical distribution theory for average ranks is developed under the uniform distribution, namely the constant k is chosen so that $P(|\bar{\pi}(i) - \bar{\pi}(j)| \leq k \text{ for all } i \text{ and } j) = 1 - \alpha$. In contrast, the inferences developed under the parametric approach, although requiring specific model assumptions, allow for inferences away from the uniform. This type of inference is necessary in the covariate model, when deciding whether a term should be dropped from the model. Such a question cannot be answered easily using the classical nonparametric approach.

1.7 Appendix

In this appendix, the limiting ($n \to \infty$) covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}$ of the vector $\boldsymbol{V}$ is computed explicitly, in terms of the parameters μ_i of the underlying Thurstone order statistics model. As described in Section 1.4, $\boldsymbol{V}$ is the $k(k-1)/2$ dimensional vector consisting of the $D^{-1}(\hat{p}_{ij})$, where D is the c.d.f. of $Z_1 - Z_2$, and Z_1 and Z_2 are i.i.d. with c.d.f. F. Since the entries of $\boldsymbol{V}$ are indexed by pairs (i, j), so are both the rows and columns of its limiting covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}$.

By the δ-method, the (i, j), (m, q) element of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma}_{ij,mq} = \lim_{n \to \infty} n \operatorname{Cov}(\hat{p}_{ij}, \hat{p}_{mq}) / [D'(D^{-1}(\hat{p}_{ij}))D'(D^{-1}(\hat{p}_{mq}))].$$

The denominator of the right hand side is just $d(\mu_j - \mu_i)\, d(\mu_q - \mu_m)$, where $d = D'$ is the density corresponding to D. To evaluate the numerator, let D_2 be the bivariate distribution function defined by $D_2(t_1, t_2) = P(Z_1 - Z_2 \le t_1, Z_2 - Z_3 \le t_2)$, where Z_1, Z_2, Z_3 are i.i.d. with c.d.f. F. Then a straightforward calculation shows:

$$\lim_{n\to\infty} n \ \mathrm{Var}(\hat{p}_{ij}) = D(\mu_j - \mu_i)[1 - D(\mu_j - \mu_i)] \text{ if } i \neq j,$$

$$\lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ij}, \hat{p}_{mq}) = 0 \text{ if } i, j, m, \text{ and } q \text{ are distinct },$$

and

$$\begin{aligned}\lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ij}, \hat{p}_{jm}) &= -\lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ij}, \hat{p}_{mj}) = \lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ji}, \hat{p}_{mj}) \\ &= -\lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ji}, \hat{p}_{jm}) \\ &= D_2(\mu_j - \mu_i, \mu_m - \mu_j) - D(\mu_j - \mu_i)D(\mu_m - \mu_j)\end{aligned}$$

if i, j, and m are distinct. The details of the calculation of $\lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ij}, \hat{p}_{jm})$ with i, j, and m distinct are as follows:

$$\begin{aligned}\lim_{n\to\infty} n \ \mathrm{Cov}(\hat{p}_{ij}, \hat{p}_{jm}) &= P(X_i < X_j, X_j < X_m) - P(X_i < X_j)P(X_j < X_m) \\ &= P(Z_i + \mu_i < Z_j + \mu_j, Z_j + \mu_j < Z_m + \mu_m) \\ &\quad -P(Z_i + \mu_i < Z_j + \mu_j)P(Z_j + \mu_j < Z_m + \mu_m) \\ &= P(Z_i - Z_j < \mu_j - \mu_i, Z_j - Z_m < \mu_m - \mu_j) \\ &\quad -P(Z_i - Z_j < \mu_j - \mu_i)P(Z_j - Z_m < \mu_m - \mu_j) \\ &= D_2(\mu_j - \mu_i, \mu_m - \mu_j) - D(\mu_j - \mu_i)D(\mu_m - \mu_j),\end{aligned}$$

where Z_i, Z_j, Z_m are i.i.d. with c.d.f. F.

Acknowledgements:

The authors thank the referees for helpful comments that improved the presentation. The work of both authors was supported by a grant from the National Science Foundation.

1.8 References

[1] R. A. Bradley. Science, statistics and paired comparisons. *Biometrics*, **32**:213-232, 1976.

[2] R. A. Bradley. Paired comparisons: some basic procedures and examples. In P.R. Krishnaiah and P.K. Sen, (editors). *Handbook of Statistics*, **4**:299-326, 1984. Amsterdam: North-Holland.

[3] R. A. Bradley and M. A. Terry. Rank analysis of incomplete block designs. I. *Biometrika*, **39**:324-345, 1952.

[4] H. Brady. Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika*, **54**:181-202, 1989.

[5] D. Brook and G. J. G. Upton. Biases in local government elections due to position on the ballot paper. *Applied Statistics*, **23**:414-419, 1974.

[6] A. Cohen and C. L. Mallows. Assessing goodness of fit of ranking models to data. *The Statistician*, **32**:361-373, 1983.

[7] D. E. Critchlow and M. A. Fligner. Paired comparison triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. To appear in *Psychometrika*, 1991.

[8] D. E. Critchlow, M. A. Fligner, and J. S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, **35**:294-318, 1991.

[9] H. E. Daniels. Rank correlation and population models. *Biometrika*, **33**:129-135, 1950.

[10] H. A. David. *The Method of Paired Comparisons.* Charles Griffin and Company, London, second edition, 1988.

[11] M. A. Fligner and J. S. Verducci. Multi-stage ranking models. *Journal of the American Statistical Association*, **83**:892-901, 1988.

[12] J. A. Hausman and D. A. Wise. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica*, **46**:403-426, 1978.

[13] R. J. Henery. Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society, Series B*, **43**:86-91, 1981.

[14] W. A. Kamakura and R. K. Srivastava. An ideal-point probabilistic choice model for heterogeneous preferences. *Marketing Science*, **5**:199-218, 1986.

[15] N. Kousgaard. Analysis of a sound field experiment by a model for paired comparisons with explanatory variables. *Scandinavian Journal of Statistics*, **11**:51-57, 1984.

[16] R. D. Luce. *Individual Choice Behavior.* Wiley, New York, 1959.

[17] C. L. Mallows. Non-null ranking models. I. *Biometrika*, **44**:114-130, 1957.

[18] P. McCullagh. Models on spheres and models for permutations, 1990. This volume.

[19] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, New York, second edition, 1989.

[20] D. McFadden. Modelling the choice of residential location. In
A. Karlquist et al. (Editors). *Spatial Interaction Theory and Planning Models*, pages. 75-96, 1978. Amsterdam: North-Holland.

[21] R. G. Miller. *Simultaneous Statistical Inference*. Springer-Verlag, New York, second edition, 1981.

[22] F. Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**:3-9, 1951.

[23] R. L. Plackett. The analysis of permutations. *Applied Statistics*, **24**: 193-202, 1975.

[24] I. R. Savage. Contributions to the theory of rank order statistics - the two-sample case. *Annals of Mathematical Statistics*, **27**:590-615, 1956.

[25] I. R. Savage. Contributions to the theory of rank order statistics - the 'trend' case. *Annals of Mathematical Statistics*, **28**:968-977, 1957.

[26] M. Schervish. Algorithm AS 195. multivariate normal probabilities with error bound *Applied Statistics*, **33**:81-94, 1984.

[27] B. Babington Smith. Discussion on Professor Ross's paper. *Journal of the Royal Statistical Society, Series B*, **12**:153-162, 1950.

[28] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, **15**:72-101, 1904.

[29] A. Springall. Response surface fitting using a generalization of the Bradley-Terry paired comparison model. *Applied Statistics*, **22**:59-68, 1973.

[30] H. Stern. Models for distributions on permutations. *Journal of the American Statistical Association*, **85**:558-564, 1990.

[31] L. L. Thurstone. A law of comparative judgement. *Psychological Reviews*, **34**:273-286, 1927.

[32] M. D. Vargo. Microbiological Spoilage of a Moderate Acid Food System Using a Dairy-Based Salad Dressing Model. Master's thesis, Department of Food Science and Nutrition, The Ohio State University, 1989.

[33] J. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, **15**:109-144, 1977.

2

Nonparametric Methods of Ranking from Paired Comparisons

H. A. David [1]
D. M. Andrews [2]

ABSTRACT Ranking by row-sum scores in the case of balanced paired-comparison experiments was generalized to unbalanced experiments in David [7]. Statistical properties of the proposed scores and associated tests of significance are developed in Andrews and David [2], where extensions to unbalanced ranked data are also treated. A brief account of this work is given and a possible generalization is introduced and examined. The simple methods here advanced make no assumptions on the pairwise preference probabilities. A secondary aim of this paper is to provide a critical review of competing methods also involving no such assumptions as well as of related methods requiring only mild assumptions. Many of the procedures discussed are illustrated on a worked example.

2.1 Introduction and Literature Review

Basic Considerations

Suppose that t objects $C_1, \ldots, C_t$ are judged in pairs, $n_{ij} (\geq 0)$ comparisons being made of C_i and $C_j (i, j = 1, \ldots, t, i \neq j)$. We are concerned with nonparametric methods for obtaining scores, and hence a ranking, of the t objects from such paired comparisons and more generally from partial rankings in subsets of two or more.

It will be convenient in this paper to use the language of sports. If $C_i \rightarrow C_j$ (player C_i defeats player C_j), then ordinarily we allot 1 point to C_i and zero to C_j. As will be seen, our estimation procedure works equally well for more general methods of splitting the point (e.g., example in Section 2.2); in particular, a draw can be treated as half a win plus half a loss. The distribution theory to be developed does, however, require a

[1] Iowa State University, Ames, Iowa
[2] Wittenberg University, Springfield, Ohio

1:0 or 0:1 outcome, with no draws permitted. We assume that in each of their n_{ij} encounters $\Pr\{C_i \rightarrow C_j\} = \pi_{ij}$, thus ruling out any replication or judge effect. Then $\pi_{ji} = 1 - \pi_{ij}$, so that there are $\binom{t}{2}$ free parameters.

Much of the paired-comparison literature is based on (paired-comparison) linear models that express π_{ij} in terms of parameters $\theta_1, \ldots, \theta_t$ denoting the strength or merit of the players. In fact, since the location of the θ's is irrelevant, the number of parameters is reduced to $t - 1$. Such "parametric" models have been studied thoroughly and will not be considered here. Instead, we focus on a method that makes no assumptions on the π_{ij} and is nonparametric in this sense. This approach provides a simple estimation procedure (David [7]) and permits the development of distribution theory and tests of significance (Andrews and David [2]). Apart from an integrated account of these two papers, we present a new generalization (equation (9) and sequel).

A secondary purpose of this paper is to survey the literature on methods of ranking from paired comparisons that make weaker than linear model assumptions on the π_{ij}, with special emphasis on those making no assumptions. In some instances more details are given in David [8] and Andrews [1]. A worked example illustrates many of the procedures discussed.

Balanced Data

The results of a paired-comparison experiment can be recorded in a tournament or preference matrix $\boldsymbol{A} = (\alpha_{ij})$, where α_{ij} is the number of wins of C_i over C_j, and $\alpha_{ii} = 0$. In a balanced tournament or round robin, ($n_{ij} = n \geq 1 \,\forall_{i,j}\, i \neq j$) it is natural to estimate the players' strengths or merits by the row-sum score vector or vector of wins $\boldsymbol{w} = \boldsymbol{A1}$, where $\mathbf{1}$ is a column vector of t 1's. But suppose that there are tied scores. One simple procedure for separating them, long used in chess tournaments, is to replace w_i by the sum of the scores of players defeated by C_i. This is given by the column headed $\boldsymbol{w}^{(2)}$ for the small tournament of Table 2.1, where, e.g., the score $w_1^{(2)}$ of C_1 at this second stage is $3 + 2 + 1 = 6$. Alternatively, $\boldsymbol{w}^{(2)} = \boldsymbol{Aw} = \boldsymbol{A}^2\mathbf{1}$.

Table 2.1. Various score vectors for the tournament with matrix **A**

	$\boldsymbol{A}$	$\boldsymbol{w}$	$\boldsymbol{w}^{(2)}$	$\boldsymbol{w}^{(3)}$	$\boldsymbol{w}^{(4)}$	$\boldsymbol{v}$
C_1	0 1 1 1 0	3	6	7	13	.6382
C_2	0 0 1 1 1	3	4	6	13	.5400
C_3	0 0 0 1 1	2	2	4	9	.3415
C_4	0 0 0 0 1	1	1	3	6	.2159
C_5	1 0 0 0 0	1	3	6	7	.3712

Note that not only have all ties been broken at this stage but C_5 has moved ahead of C_3 although $w_5 < w_3$. Neither of these events (breaking of all ties, reversal of some rankings) need happen and one may wish to continue the reallocation process to obtain $\boldsymbol{w}^{(3)} = \boldsymbol{A}^3\mathbf{1}$, etc. This is a slight simplification (Moon [13]) of the Kendall-Wei procedure (Kendall [12]) in which $\boldsymbol{B} = \boldsymbol{A} + \frac{1}{2}\boldsymbol{I}$ rather than $\boldsymbol{A}$ is powered, where $\boldsymbol{I}$ is the identity matrix. The above tournament is *strong*, meaning that for any two players C_i, C_j either $C_i \rightarrow C_j$ or there exist other players $C_{i_1}, C_{i_2}, \ldots$ such that $C_i \rightarrow C_{i_1} \rightarrow C_{i_2} \cdots \rightarrow C_j$. For strong tournaments (with $t > 3$) it follows from Perron-Frobenius theory (e.g., Seneta [19]) that

$$\lim_{r\to\infty} (\boldsymbol{A}/\lambda)^r\, \mathbf{1} = \boldsymbol{v},$$

where λ is the unique positive eigenvalue of $\boldsymbol{A}$ with the largest absolute value and $\boldsymbol{v}$ is a vector of positive terms. In Table 2.1 $\boldsymbol{v}$ has been obtained from $\boldsymbol{Av} = \lambda\boldsymbol{v}$, with $\boldsymbol{v}'\boldsymbol{v} = \mathbf{1}$, and happens to give the same ranking as $\boldsymbol{w}^{(2)}$.

Interesting related methods have been put forward by Daniels [6] and Moon and Pullman [14]. However, it is dubious whether the resulting ranking as well as those corresponding to $\boldsymbol{w}^{(2)}$ and $\boldsymbol{v}$ are really an improvement over the simple row-sum score $\boldsymbol{w}$. The common feature of the more elaborate methods is to give more credit to a player for defeating a high scoring than a low scoring opponent, but this means, of course, that a loss to the latter is punished less than a loss to the former. Also it is easy to show that an interchange of all wins and losses does not necessarily reverse a ranking. Thus use of these methods is perhaps best reserved for breaking ties among high row-sum scores.

Another major but very different area of research has resulted from Slater's [20] principle of seeking a ranking that minimizes the number of *inconsistencies*, i.e., individual -comparison results opposite to what would be expected from the ranking. Slater's premise of giving equal weight to mild and gross inconsistencies is debatable. However, it has been shown by Thompson and Remage [22] that for $n_{ij} \leq 1$ Slater's ranking results if weak stochastic transitivity of the π_{ij} is assumed, i.e., if for any triple (C_i, C_j, C_k)

$$\pi_{ij} \geq \frac{1}{2}, \ \pi_{jk} \geq \frac{1}{2}, \quad \text{imply } \pi_{ik} \geq \frac{1}{2}.$$

Such a restriction on the π_{ij} takes us outside the methods considered here. However, it is worth noting that the approach is applicable for unbalanced data although difficult to implement except in fairly small tournaments. Further references and a very brief review of Slater's and related methods are given in David ([8], pp. 23-25).

Unbalanced Data

Suppose now that the n_{ij} are not all equal. This includes the case when some of the n_{ij} are zero, corresponding to empty cells in the preference matrix. Clearly, row-sum scores are no longer satisfactory. The basic problems are to take into account (a) the varied caliber of the opposition encountered by each player, and (b) possibly different numbers of matches played by the contestants.

The first approach to handling unbalanced data seems to have been through unweighted least squares (Gulliksen [10]). Let θ_i $(i = 1, \ldots, t)$ denote the merit or worth of C_i. Then the estimated merits $\hat{\theta}_i$ are the values of the θ_i minimizing $\Sigma^*(d_{ij} - \theta_i + \theta_j)^2$, where Σ^* ranges over all pairs (i, j) for which $n_{ij} \geq 1$, $d_{ij} = H^{-1}(p_{ij})$, H is the cdf of a rv symmetric about zero, and $p_{ij} = \alpha_{ij}/n_{ij}$. This approach corresponds to a paired-comparison linear model for which $\pi_{ij} = H(\theta_i - \theta_j)$ and gives estimates $\hat{\pi}_{ij} = H(\hat{\theta}_i - \hat{\theta}_j)$. With the reduction in parameters it is outside our scope. However, Kaiser and Serlin [11] note that the essential property of the d_{ij} for a sensible analysis is merely that $d_{ji} = -d_{ij}$; it is not *necessary* to relate the d_{ij} to an H-function. For example, for an incomplete tournament with $n_{ij} \leq 1$ they simply take $d_{ij} = 1$ if $C_i \rightarrow C_j$, $d_{ij} = -1$ if $C_j \rightarrow C_i$, and $d_{ij} = 0$ in case of a draw. The θ_i can then be estimated provided that the tournament is connected (every player meets every other player either directly or through intermediaries).

Conceptually closer to the Kendall-Wei method is Cowden [4] proposing the following iterative procedure for arriving at a set of scores $\boldsymbol{p}^{(k)}$:

$$\boldsymbol{p}^{(k)} = \boldsymbol{A}\boldsymbol{u}^{(k-1)}, \ \boldsymbol{q}^{(k)} = \boldsymbol{A}'\boldsymbol{v}^{(k-1)}, \tag{1}$$

where $\boldsymbol{A}'$, the transpose of $\boldsymbol{A}$, is the matrix of "losses" and

$$u_i^{(k)} = \frac{p_i^{(k)}}{p_i^{(k)} + q_i^{(k)}}, \ v_i^{(k)} = 1 - u_i^{(k)}, \ i = 1, \ldots, t. \tag{2}$$

With $u_i^{(0)} = v_i^{(0)} = \frac{1}{2}$ the rankings usually stabilize within a few iterations and the scores themselves shortly thereafter (provided $\boldsymbol{A}$ is a strong matrix). The method usually gives fairly sensible results, although even in the balanced case it often breaks tied row-sum scores in favor of the player(s) whose wins were over *weaker* opposition. Consider, for example, the following balanced tournament:

$$\begin{array}{cccc} & \text{Row-sum} & \text{Cowden's score} & \text{Kendall-Wei} \\ \boldsymbol{A} & \boldsymbol{A1} & \boldsymbol{u} & \boldsymbol{A}^2\mathbf{1} \\ \begin{pmatrix} 0 & 6 & 6 & 6 \\ 4 & 0 & 5 & 9 \\ 4 & 5 & 0 & 9 \\ 4 & 1 & 1 & 0 \end{pmatrix} & \begin{pmatrix} 18 \\ 18 \\ 18 \\ 6 \end{pmatrix} & \begin{pmatrix} .576 \\ .584 \\ .584 \\ .257 \end{pmatrix} & \begin{pmatrix} 252 \\ 216 \\ 216 \\ 108 \end{pmatrix} \end{array}$$

Note that half of the wins of C_2 and C_3 come at the expense of the inferior player C_4, whereas C_1 has won 6 of 10 from each of the others; yet Cowden's scores rank C_1 below C_2 and C_3.

Nishisato [15] shows that his dual scaling procedure can be applied to unbalanced paired-comparison data. However, his approach is inappropriate in our situation since it depends fundamentally on the presence of a judge effect which we have excluded.

In a recent paper Chebotariov [3] notes that any tournament can be thought of as an aggregation of m, say, possibly incomplete subtournaments or rounds in which any pair of players meet at most once. If C_i meets C_j in the k-th round, the result may be denoted by $d_{ij}^{(k)}$, with the only requirement that $d_{ji}^{(k)} = -d_{ij}^{(k)}$. The simple row-sum for C_i is then

$$u_i = \sum_{k=1}^{m} \sum_{j}^{(i:k)} d_{ij}^{(k)}, \tag{3}$$

where the inner sum is over all C_j that have met C_i in the k-th round. Chebotariov seeks a score x_i for C_i which reduces to u_i under complete balance, and which takes the general form

$$x_i = \sum_{k=1}^{m} \sum_{j}^{(i:k)} f_{ij}^{(k)}, \tag{4}$$

where f_{ij} is a reward function for C_i from its encounter with C_j in round k, viz.,

$$f_{ij}^{(k)} = d_{ij}^{(k)} + \epsilon(x_i - x_j + mtd_{ij}^{(k)}), \tag{5}$$

where the constant $\epsilon \geq 0$ determines the extent to which the x_i-scores depend on the relative strengths of the players.

Scores are calculated by solving the linear system (4) of t equations in t unknowns. Although the reward function (5) seems a reasonable way of taking the strength of a player's opponents into account, little guidance is provided on the choice of ϵ.

We may also mention here a review paper of research in the U.S.S.R. on paired comparisons by Prigarina, Chebotariov, and Schmerling [17].

Another recent addition to the literature is Crow [5], although its origins go back to Thompson [21]. Thompson makes the simplifying assumption that π_{ij} is a function solely of the difference D_{ij} in the ranks of C_i and C_j. Crow is led to seeking the ranking that maximizes what he terms the *net difference in ranks* (NDR), viz.,

$$\sum_{i<j} D_{ij} \ (\alpha_{ij} - \alpha_{ji}).$$

In so far as the π_{ij} are restricted by the simplifying assumption, this approach is, strictly speaking, outside our scope.

2.2 The Proposed Method of Scoring

Basic Properties

In dealing with balanced data Ramanujacharyulu [18] considers powering $\boldsymbol{B}'$ rather than $\boldsymbol{B} = \boldsymbol{A} + \frac{1}{2}\boldsymbol{I}$ to obtain the "iterated weakness" vector $(\boldsymbol{B}')^r\mathbf{1}$. The best player is now the one with the fewest iterated losses and not necessarily with the largest number of iterated wins. A balance between rewarding beating strong players and punishing losing to weak players is struck by the difference vector (David [7])

$$\boldsymbol{s}^{(r)} = \boldsymbol{B}^r\mathbf{1} - (\boldsymbol{B}')^r\mathbf{1}. \tag{6}$$

Consider now an incomplete tournament in which each pair of players has met at most once. It is seen immediately from (6) that

$$\sum_{i=1}^{t} s_i^{(r)} = \mathbf{1}' \ \boldsymbol{s}^{(r)} = 0$$

and that $\boldsymbol{s}^{(r)}$ becomes $-\boldsymbol{s}^{(r)}$ when wins and losses are interchanged (i.e., $\boldsymbol{B}$ is replaced by $\boldsymbol{B}'$). We recommend $\boldsymbol{s}^{(2)}$ (henceforth just $\boldsymbol{s}$) for general use since it is equivalent to row-sum scoring in the case of a balanced tournament (David [7]). In other words, $\boldsymbol{s}$ (unlike $\boldsymbol{s}^{(r)}$ for $r > 2$) cannot serve as a tie breaker for a balanced tournament, which makes it more attractive in the absence of balance.

For $r = 2$, (6) gives

$$\boldsymbol{s} = \boldsymbol{A}^2\mathbf{1} - (\boldsymbol{A}')^2\mathbf{1} + \boldsymbol{A}\mathbf{1} - \boldsymbol{A}'\mathbf{1} \tag{7}$$

or, in obvious notation,

$$\boldsymbol{s} = \boldsymbol{w}^{(2)} - \boldsymbol{\ell}^{(2)} + \boldsymbol{w} - \boldsymbol{\ell} \tag{8}$$

In words, s_i is the total number of (a) wins of players defeated by C_i minus losses of players to whom C_i lost, plus (b) C_i's wins minus C_i's losses. Clearly, (b) could be omitted without changing any of the preceding properties. However, in its absence, C_i after beating an opponent with no wins would be worse off than before, since the win adds nothing to part (a) of C_i's score but adds 1 to the score of each player who defeated C_i.

We now give a new, purely algebraic, proof of the equivalence of $\boldsymbol{s}$ and $\boldsymbol{w}$ for unbalanced tournaments. Since $\boldsymbol{B} + \boldsymbol{B}' = \boldsymbol{J}$, where $\boldsymbol{J}$ is the $t \times t$ matrix of 1's, we have

$$\begin{aligned} \boldsymbol{s} &= \boldsymbol{B}^2\mathbf{1} - (\boldsymbol{B}')^2\mathbf{1} \\ &= \boldsymbol{B}^2\mathbf{1} - \boldsymbol{J}^2\mathbf{1} + \boldsymbol{J}\boldsymbol{B}\mathbf{1} + \boldsymbol{B}\boldsymbol{J}\mathbf{1} - \boldsymbol{B}^2\mathbf{1} \\ &= -\boldsymbol{J}^2\mathbf{1} + \boldsymbol{J}\boldsymbol{A}\mathbf{1} + \boldsymbol{A}\boldsymbol{J}\mathbf{1} + \boldsymbol{J}\mathbf{1} \\ &= -t^2\mathbf{1} + \frac{1}{2}\, t(t-1)\mathbf{1} + t\boldsymbol{w} + t\mathbf{1} \\ &= t(\boldsymbol{w} - \frac{1}{2}\,(t-1)\mathbf{1}). \end{aligned}$$

For numerical examples illustrating the proposed method, see David [7], Andrews and David [2]), and Section 2.5 of this paper.

Formula (8) may also be used in larger tournaments when the number of encounters n_{ij} of A_i and A_j is unrestricted. The most obvious procedure is to take $\boldsymbol{A} = (\alpha_{ij})$ as the matrix of the number of wins of C_i over C_j. But this is often inappropriate since too much weight may then be given to players involved in (relatively) large numbers of comparisons. Indeed, when some n_{ij} are much greater than 1, the effects of the indirect wins and losses $\boldsymbol{w}^{(2)}$ and $\boldsymbol{\ell}^{(2)}$ swamp the effects of $\boldsymbol{w}$ and $\boldsymbol{\ell}$ (Andrews [1]). To avoid this effect one may take $\boldsymbol{A} = (p_{ij})$, where $p_{ij} = \alpha_{ij}/n_{ij}$ for $n_{ij} > 0$ and $p_{ij} = 0$ for $n_{ij} = 0$. This choice, briefly suggested in David [7], is investigated in detail in Andrews and David [2]. More generally, one can take $\boldsymbol{A} = (\alpha'_{ij})$, where for $n_{ij} \geq 1$

$$\alpha'_{ij} = c(n_{ij})\ \alpha_{ij}, \tag{9}$$

$c(n_{ij})$ being a known function of n_{ij}, with $c(1) = 1$. Then,

$$\alpha'_{ij} + \alpha'_{ji} = c(n_{ij})\ (\alpha_{ij} + \alpha_{ji}) = c(n_{ij})\ n_{ij} = n'_{ij}(\text{say}). \tag{10}$$

Important special cases are the previous choices $c(n_{ij}) = 1$ and $c(n_{ij}) = 1/n_{ij}$. An intermediate function is $c(n_{ij}) = 1/\sqrt{n_{ij}}$. If $n_{ij} = 0$, we set $\alpha'_{ij} = \alpha'_{ji} = n'_{ij} = 0$.

FURTHER PROPERTIES

We now explore the consequences of (9) following closely the development given in Andrews and David [2] for $\alpha'_{ij} = p_{ij}$. In preparation for dealing with the distribution of $\boldsymbol{s}$, we express s_i as a linear function of $\alpha'_{12}, \ldots, \alpha'_{n-1,n}$. We have

$$w_i = \sum_j^{(i)} \alpha'_{ij}, \; \ell_i = \sum_j^{(i)} \alpha'_{ji}, \; w_i^{(2)} = \sum_j^{(i)} \alpha'_{ij} w_j, \; \ell_i^{(2)} = \sum_j^{(i)} \alpha'_{ji} \ell_j, \tag{11}$$

where $\sum_j^{(i)}$ denotes the sum over all players C_j that have met C_i. Then from (10) and (11)

$$w_i - \ell_i = \sum_j^{(i)} (2\alpha'_{ij} - n'_{ij}) \tag{12}$$

and

$$\begin{aligned} w_i^{(2)} - \ell_i^{(2)} &= \sum_j^{(i)} [\alpha'_{ij} w_j - (n'_{ij} - \alpha'_{ij})(m'_j - w_j)] \\ &= \sum_j^{(i)} [n'_{ij} w_j - m'_j (n'_{ij} - \alpha'_{ij})], \end{aligned} \tag{13}$$

where

$$m'_j = w_j + \ell_j = \sum_k^{(j)} n'_{jk}. \tag{14}$$

The only quantities in (13) depending on the experimental outcome are w_j and α'_{ij}. Thus (12) and (13) show that s_i in (8) is a linear function of α', $\alpha'_{12}, \ldots, \alpha'_{n-1,n}$.

Next, we express s_i as a linear function of independent elements α'_{gh}, i.e., α'_{gh} and α'_{hg} do not both occur in this form of s_i. Whenever C_i and C_j have met, we have

$$w_j = (n'_{ij} - \alpha'_{ij}) + \sum_{k \neq i}^{(j)} \alpha'_{jk},$$

where $\sum_{k \neq i}^{(j)}$ denotes the sum over all players C_k (excluding C_i) that have met C_j. Substituting this in (13) and adding the result to (12) gives

$$\begin{aligned} s_i = \sum_j^{(i)} [(m'_j + 2 - n'_{ij})\, \alpha'_{ij} + n'_{ij} \sum_{k \neq i}^{(j)} \alpha'_{jk} \\ -n'_{ij}(m'_j + 1 - n'_{ij})] \end{aligned} \tag{15}$$

Both α'_{gh} and α'_{hg} occur in (15) whenever C_g and C_h have met both C_i and each other. To consolidate these complementary quantities, write

$$\sum_{j}^{(i)} \sum_{k\neq i}^{(j)} \alpha'_{jk} = \sum_{j}^{(i)} \sum_{k\neq i}^{(i,j)} \alpha'_{jk} + \sum_{j}^{(i)} \sum_{k\neq i}^{(\sim i,j)} \alpha'_{jk}, \tag{16}$$

where $\sum_{k\neq i}^{(i,j)}$ denotes the sum over the m_{ij}, say, players C_k that have met both C_i and C_j, and $\sum_{k\neq i}^{(\sim i,j)}$ denotes the sum over those C_k (excluding C_i) that have met C_j but not C_i. The first term, T_{i1} on RHS of (16) always includes both α'_{gh} and α'_{hg} or neither, and hence does not depend on the experimental outcome. It may be written $\Sigma n'_{gh}$, where the sum extends over $\frac{1}{2}\sum_{j}^{(i)} m_{ij}$ terms, corresponding to those pairs C_g and C_h that have met C_i and each other. The second term, T_{i2}, consists of elements involving players other than C_i, exactly one of whom has met C_i, and thus contains at most one of α'_{gh} and α'_{hg}. Thus, from (15), s_i has been expressed as required.

Case of No Empty Cells and Discussion

If all $n_{ij} > 0$, then $T_{i2} = 0$ since all players have met C_i. It follows that

$$s_i = \sum_{j\neq i} (m'_j + 2 - n'_{ij})\, \alpha'_{ij} + K_i, \tag{17}$$

where by (14) $m'_j = \sum_{k\neq j} n'_{jk}$, and K_i does not involve any α'_{ij}'s. For the special case $\alpha'_{ij} = p_{ij}$, we have $n'_{ij} = 1$ and $m'_j = t - 1$, giving

$$s_i = t \sum_{j\neq i} (p_{ij} - \frac{1}{2}) = t[w_i - \frac{1}{2}(t-1)]. \tag{18}$$

This relation between s_i and w_i is known for designs that are balanced, which is not a requirement here.

Now $\mathrm{E}\,\frac{\sum_{j\neq i} p_{ij}}{t-1} = \frac{\sum_{j\neq i} \pi_{ij}}{t-1} = \pi_{i.}$ (say) is the probability that C_i defeats a player drawn at random from C_i's opponents. Clearly $\pi_{i.}$ is a measure of strength and is (essentially) estimated by s_i when $\alpha'_{ij} = p_{ij}$. We take this as support for concentrating on this special case in Andrews and David [2] and in the remainder of this paper. Nevertheless, if, e.g., a Bradley-Terry model is appropriate, a different choice of α'_{ij} may give results closer to a Bradley Terry analysis. In a rather special no empty cell situation Groeneveld [9] finds that $\alpha'_{ij} = \alpha_{ij}$ gives rankings closer to the Bradley Terry rankings than does $\alpha'_{ij} = p_{ij}$. Of course, as Groeneveld notes, if a Bradley Terry model holds, the only point in using the present approach is its simplicity and ready comprehensibility. If there is wide variation between the (nonzero) n_{ij}, then the choice $\alpha'_{ij} = p_{ij}$, which ignores this, is inadvisable. Instead the compromise $\alpha'_{ij} = \alpha_{ij}/\sqrt{n_{ij}}$ is recommended.

Further Results for $\alpha'_{ij} = p_{ij}$(with $n_{ij} \geq 0$).

The basic result (15) reduces in this case to

$$s_i = \sum_j^{(i)} [(m_j + 1)p_{ij} - m_j + \frac{1}{2}m_{ij} + \sum_{k \neq i}^{(\sim i,j)} p_{jk}]$$

$$= \sum_j^{(i)} [(m_j + 1)(p_{ij} - \frac{1}{2}) + \sum_{k \neq i}^{(\sim i,j)} (p_{jk} - \frac{1}{2})], \tag{19}$$

where m_j is the number of players met by C_j.

An interesting result can be obtained in an important tournament arrangement when not all matches can be held. If the players are arranged in a group divisible design of m distinct groups of size $a = t/m$, where each player meets once (or an equal number of times) all the players in the other groups only, then (Andrews and David [2])

$$s_i = (t - a + 2)(w_i - \frac{1}{2}(t - a)) - \Sigma(w_k - \frac{1}{2}(t - a)),$$

where the sum extends over the players in C_i's group. Thus s_i is a multiple of C_i's number of wins minus a correction for the strength of C_i's group.

2.3 Distribution Theory and Tests of Significance for $\alpha'_{ij} = p_{ij}$

Moments

From (19) we have at once

$$E(s_i) = \sum_j^{(i)} [(m_j + 1)(\pi_{ij} - \frac{1}{2}) + \sum_{k \neq i}^{(\sim i,j)} (\pi_{jk} - \frac{1}{2})], \tag{20}$$

$$\text{var}(s_i) = \sum_j^{(i)} \left[(m_j + 1)^2 \frac{\pi_{ij}\pi_{ji}}{n_{ij}} + \sum_{k \neq i}^{(\sim i,j)} \frac{\pi_{jk}\pi_{kj}}{n_{jk}} \right]. \tag{21}$$

For the more complex cov(s_i, s_j) see Andrews and David [2]. When there are no empty cells these expressions become simply

$$E(s_i) = t \sum_{j \neq i} (\pi_{ij} - \frac{1}{2}), \ \text{var}(s_i) = t^2 \sum_{j \neq i} \frac{\pi_{ij}\pi_{ji}}{n_{ij}}, \ \text{cov}(s_i, s_j) = -\frac{t^2 \pi_{ij}\pi_{ji}}{n_{ij}} \tag{22}$$

Note that under the "hypothesis of randomness"

$$H_0 : \pi_{ij} = \frac{1}{2} \;\; \forall\,(i,j), i \neq j$$

(20) gives $E(s_i) = 0$, as it should.

Asymptotics

Little can be said about the asymptotic distribution of the scores if the n_{ij} are allowed to grow in an uncontrolled manner. Let $n_{ij}/\sum_k^{(i)} n_{ik} \rightarrow c_{ij}$ as the $n_{ij} \rightarrow \infty$, for all $i \neq j$, where c_{ij} is some constant in (0,1). Since the proportions comprising s_i in (19) are independent, it follows that the standardized score $d_i = [s_i - E(s_i)]\sqrt{\text{var}(s_i)}$ has an asymptotic $N(0,1)$ distribution. Similarly, we can show that any linear function of the standardized scores has a (limiting) normal distribution, and hence that the asymptotic joint distribution of the scores themselves is multivariate normal. Andrews [1] uses this last result to generalize several tests of hypotheses from David [8] for balanced experiments.

Tests for the Equality of the Players

It is also of interest to test whether the players are of equal merit. First note that the covariance matrix $\boldsymbol{\Sigma}$ of the scores s_i is singular, since $\boldsymbol{s}'\mathbf{1} = 0$ and hence $\boldsymbol{\Sigma}\mathbf{1} = \mathbf{0}$. We therefore focus on the covariance matrix $\tilde{\boldsymbol{\Sigma}}$ of any set $\tilde{\boldsymbol{s}}$ of $t-1$ scores, since $\tilde{\boldsymbol{\Sigma}}$ will be of full rank whenever the tournament is connected. In light of the joint asymptotic multivariate normality of $\boldsymbol{s}$, it can be shown that the test statistics $Q = \tilde{\boldsymbol{s}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{s}}$ has an asymptotic χ^2_{t-1} distribution with noncentrality parameter $\lambda = \tilde{\boldsymbol{\mu}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}$, where $\tilde{\boldsymbol{\mu}} = E(\tilde{\boldsymbol{s}})$.

When $n_{ij} = n \,\forall\, i \neq j$, our statistic Q reduces to $Q_{Bal} = 4nt^{-3}\sum_{i=1}^t s_i^2$, with (asymptotic) noncentrality parameter $\lambda_{Bal} = 4nt^{-3}\sum_{i=1}^t \mu_i^2$. An important special case of λ_{Bal} is the following 'one outlier' model:

$$\pi_{ij} = \begin{cases} \pi > \frac{1}{2}, & \text{for } i = t \text{ and } j \neq t \\ \frac{1}{2}, & \text{for } i, j \neq t \text{ and } i \neq j. \end{cases}$$

In this case the noncentrality parameter reduces to $\lambda_{Bal} = 4n(t-1)(\pi - \frac{1}{2})^2$.

Specific results for group divisible designs are also given in Andrews and David [2].

In all cases we can test H_0 approximately by referring Q to tables of percentage points of χ^2_{t-1}. For small tournaments the exact null distribution of Q can be generated by an exhaustive listing of all possible outcomes of the experiment (given the n_{ij}). This involves calculating the value of Q from each outcome and tabulating the probabilities with which Q takes on each of these values. There are indications that the χ^2 approximation provides a very conservative test (Andrews and David [2]).

2.4 Ranking Methods

Suppose now that the objects are ranked in 'blocks', where $k_j(\geq 2)$ of the $t(\geq 3)$ objects are ranked in block B_j, for $j = 1(1)b$. Such ranked data may be regarded as constrained paired-comparison data, since a ranking of k objects implies $\binom{k}{2}$ paired comparisons, with no circularities possible. This forced within block transitivity is the main distinction between rankings and paired comparisons. Nevertheless, we have found it useful to handle ranked data by replacing it by its constituent paired comparisons.

We confine ourselves here to a summary of the main results obtained in Andrews and David [2]. Let r_{ij} be the rank of object C_i in block B_j and let $r_i = \Sigma_j r_{ij}$, where the sum is over all blocks in which C_i appears ($r_{ij} = 1$ for C_i poorest). For a balanced incomplete block (BIB) design in which each object occurs in $m(\leq t)$ blocks of size k, our score s_i, calculated from the resulting paired-comparison table, is a linear function of r_i and also of the score proposed by Prentice [16] for ranked data. Correspondingly, the test statistic Q of Section 2.3, which reduces to

$$Q_{\text{BIB}} = \frac{4}{nt}\sum_{i=1}^{t}(r_i - m\frac{k+1}{2})^2,$$

where $n = m(k-1)/(t-1)$, is related to Prentice's test statistic C by

$$Q_{\text{BIB}} = \frac{k+1}{3}C_{\text{BIB}}.$$

This means that asymptotically Q_{BIB} is distributed as χ^2_{t-1} or $\frac{k+1}{3}\chi^2_{t-1}$ according as the data are originally in paired-comparison or ranked format.

For a BIB design the resulting paired-comparison table is completely balanced. In the absence of such balance our score s_i is not in general equivalent to Prentice's. However, our score takes into account the caliber of the competitors encountered by C_i. An illustration for a group divisible design is given in Andrews and David [2]. For a discussion *inter alia* of the merits of Prentice's score versus other proposals made for unbalanced ranked data, see Wittkowski [23, 24], references overlooked in our earlier paper.

2.5 Numerical Example

Consider the following data comparing several graders of student writing. Each of the students in the junior class at Wittenberg University is required to take a writing proficiency exam, and each student's paper is then read and marked by two members of a panel of graders. There was concern that some graders were considerably more lenient than others. To compare the graders, we examine the number of papers on which each grader gave a higher or lower mark than the other grader who read the paper. The graders are then the objects, and each paper provides a comparison of two graders. Let $\alpha_{ij} = 1$ and $\alpha_{ji} = 0$ if grader C_i gave the higher mark on a paper also marked by grader C_j. If C_i and C_j gave the same mark, we take $\alpha_{ij} = \alpha_{ji} = \frac{1}{2}$. Given below is a matrix of such 'comparisons' among a subset of seven of the graders.

$$\begin{bmatrix} - & 1.5 & 4.5 & 0.5 & * & 3.5 & 1.0 \\ 0.5 & - & 4.0 & 4.0 & * & 1.5 & * \\ 3.5 & 3.0 & - & 2.5 & 2.5 & 0.0 & 2.5 \\ 0.5 & 3.0 & 5.5 & - & * & 4.5 & 1.0 \\ * & * & 0.5 & * & - & 0.5 & 4.5 \\ 2.5 & 1.5 & 1.0 & 3.5 & 0.5 & - & 1.0 \\ 2.0 & * & 2.5 & 2.0 & 6.5 & 1.0 & - \end{bmatrix}$$

The dashes along the diagonal denote that no object was compared with itself, i.e., no grader gave both marks for a given paper. The asterisks denote pairs of graders for which there were no papers; note that there are 4 such pairs among the 21 distinct pairs of graders. The data are quite unbalanced, in that the number of papers marked by each pair varies greatly, ranging from 0 to 11.

GULLIKSEN'S METHOD

For the least squares approach we need data in the form of observed 'differences' $\{d_{ij}\}$ for all the pairs which have been compared. For simplicity, take as the observed difference between C_i and C_j the difference $d_{ij} = p_{ij} - p_{ji}$. Note that $d_{ij} = -d_{ji}$, as Kaiser and Serlin require.

To minimize the sum of squared discrepancies $\Sigma^*(d_{ij} - \theta_i - \theta_j)^2$, Gulliksen iteratively updates each score θ_i by adding the average of the discrepancies for that score:

$$\theta_i^{(k)} = \theta_i^{(k-1)} + \frac{1}{m+1} \sum_j^{(i)} [d_{ij} - (\theta_i - \theta_j)]$$

If we begin the procedure with a null initial score vector $\boldsymbol{\theta}^{(0)} = \mathbf{0}$, the final score vector, after six iterations, is

$$\hat{\boldsymbol{\theta}}' = (.106, -.034, -.151, .033, -.259, .090, .134).$$

COWDEN'S METHOD

We focus on the $\boldsymbol{u}^k$ of (1) and (2), the vector of 'win scores'. These scores, beginning with $\boldsymbol{u}^{(0)} = \boldsymbol{v}^{(0)} = \frac{1}{2}\mathbf{1}$, stabilize to three decimal places only after 11 iterations:

$$\boldsymbol{u}' = (.543, .528, .456, .538, .380, .502, .530).$$

Cowden suggests taking as u_i^{k+1} the mean of $u_i^{(k)}$ and $u_i^{(k-1)}$ "whenever it seems useful". But if this is done at every stage, the scores converge to three decimals only after 12 iterations, with slightly different final values:

$$\boldsymbol{u}' = (.542, .528, .456, .537, .382, .502, .531).$$

CHEBOTARIOV'S METHOD

We set the $d_{ij}^{(k)}$ of (3) equal to ± 1. It is evident from Chebotariov's calculations that he takes the number of rounds to be $m = \max n_{ij}$, which would be $n_{57} = 11$ for our data. Andrews [1] shows that the system of equations (4) has solution

$$\boldsymbol{x} = (\boldsymbol{I} + mt\epsilon)\left(\boldsymbol{I} + \epsilon\begin{bmatrix} N_1 & -n_{12} & \cdots & -n_{1t} \\ -n_{21} & N_2 & \cdots & -n_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ -n_{t1} & -n_{t2} & \cdots & N_t \end{bmatrix}\right)^{-1} \boldsymbol{u}$$

$$= (1+77\epsilon)\left(\boldsymbol{I} + \epsilon\begin{bmatrix} 20 & -2 & -8 & -1 & 0 & -6 & -3 \\ -2 & 19 & -7 & -7 & 0 & -3 & 0 \\ -8 & -7 & 32 & -8 & -3 & -1 & -5 \\ -1 & -7 & -8 & 27 & 0 & -8 & -3 \\ 0 & 0 & -3 & 0 & 15 & -1 & -11 \\ -6 & -3 & -1 & -8 & -1 & 21 & -2 \\ -3 & 0 & -5 & -3 & -11 & -2 & 24 \end{bmatrix}\right)^{-1} \begin{bmatrix} 2 \\ 1 \\ -4 \\ 2 \\ -4 \\ -1 \\ 4 \end{bmatrix}$$

Chebotariov's scores change considerably as ϵ varies:

$$\begin{array}{cccc} \epsilon: & 0.00 & 0.01 & 0.05 \\ & \begin{bmatrix} 2.00 \\ 1.00 \\ -4.00 \\ 2.00 \\ -4.00 \\ -1.00 \\ 4.00 \end{bmatrix} & \begin{bmatrix} 2.74 \\ 1.37 \\ -4.91 \\ 2.62 \\ -5.81 \\ -1.12 \\ 5.11 \end{bmatrix} & \begin{bmatrix} 4.21 \\ 2.25 \\ -5.90 \\ 3.79 \\ -9.66 \\ -0.92 \\ 6.24 \end{bmatrix} \end{array}$$

$$\begin{array}{cccc} 0.10 & 0.25 & 0.50 & 1.00 \\ \begin{bmatrix} 5.02 \\ 2.88 \\ -6.12 \\ 4.44 \\ -11.91 \\ -0.59 \\ 6.28 \end{bmatrix} & \begin{bmatrix} 5.95 \\ 3.74 \\ -6.19 \\ 5.23 \\ -14.60 \\ -0.05 \\ 5.92 \end{bmatrix} & \begin{bmatrix} 6.42 \\ 4.24 \\ -6.16 \\ 5.65 \\ -16.04 \\ 0.28 \\ 5.60 \end{bmatrix} & \begin{bmatrix} 6.72 \\ 4.57 \\ -6.13 \\ 5.92 \\ -16.93 \\ 0.50 \\ 5.36 \end{bmatrix} \end{array}$$

DAVID'S METHOD

To illustrate David's method, we choose $\alpha'_{ij} = \alpha_{ij}/\sqrt{n_{ij}}$. The various components $\boldsymbol{w}, \boldsymbol{\ell}, \boldsymbol{w}^{(2)}$, and $\boldsymbol{\ell}^{(2)}$ of his score vector $\boldsymbol{s}$ are then, respectively, the row-sums of $\boldsymbol{A}, \boldsymbol{A}', \boldsymbol{A}^2$, and $(\boldsymbol{A}')^2$:

$$\boldsymbol{A} = \left[\begin{array}{ccccccc} - & 1.06 & 1.59 & 0.50 & * & 1.43 & 0.58 \\ 0.35 & - & 1.51 & 1.51 & * & 0.87 & * \\ 1.24 & 1.13 & - & 0.88 & 1.44 & 0.00 & 1.12 \\ 0.50 & 1.13 & 1.94 & - & * & 1.59 & 0.58 \\ * & * & 0.29 & * & - & 0.50 & 1.36 \\ 1.02 & 0.87 & 1.00 & 1.24 & 0.50 & - & 0.71 \\ 1.15 & * & 1.12 & 1.15 & 1.96 & 0.71 & - \end{array}\right] \qquad \begin{array}{c} \boldsymbol{w} \\ \left[\begin{array}{c} 5.16 \\ 4.24 \\ 5.82 \\ 5.75 \\ 2.15 \\ 5.33 \\ 6.09 \end{array}\right] \end{array}$$

$$\begin{array}{c} \boldsymbol{w}^{(2)} \\ \left[\begin{array}{c} 27.76 \\ 23.92 \\ 26.18 \\ 30.70 \\ 12.61 \\ 27.25 \\ 27.07 \end{array}\right] \end{array} \qquad \begin{array}{c} \boldsymbol{s} \\ \left[\begin{array}{r} 5.10 \\ 0.59 \\ -9.93 \\ 4.79 \\ -10.95 \\ 4.33 \\ 6.08 \end{array}\right] \end{array}$$

$$\boldsymbol{\ell}' = [4.27 \quad 4.19 \quad 7.45 \quad 5.29 \quad 3.90 \quad 5.09 \quad 4.34]$$

$$\boldsymbol{\ell}^{(2)\prime} = [23.56 \quad 23.38 \quad 34.48 \quad 26.37 \quad 21.80 \quad 23.16 \quad 22.75]$$

In contrast to the two iterative methods, these scores are calculated in a few short steps.

COMPARISON OF THE METHODS

Since direct comparison of the various methods is rather awkward because of the different scales used, it is helpful to standardize each set of scores by subtracting the mean and dividing by the standard deviation of each set:

grader	Cowden		Gulliksen		Chebotariov		David	
C_1	0.77	1	0.80	2	0.76	1	0.69	2
C_2	0.54	4	-0.15	5	0.48	4	0.08	5
C_3	-0.68	6	-0.95	6	-0.79	6	-1.35	6
C_4	0.69	2	0.31	4	0.67	3	0.65	3
C_5	-1.96	7	-1.69	7	-1.88	7	-1.49	7
C_6	0.09	5	0.69	3	-0.01	5	0.59	4
C_7	0.56	3	0.99	1	0.76	2	0.83	1

(For Chebotariov's scores we have used $\epsilon = 0.25$, a value which he favors in his examples.) Given beside each set of scores are the objects' ranks induced by those scores. Agreement between the different methods is good, especially in finding the two harshest graders, C_5 and C_3.

Acknowledgements:

This work was supported by the U. S. Army Research Office. The authors are grateful to the referees for suggestions leading to an improved presentation.

2.6 References

[1] D. M. Andrews. Nonparametric Analysis of Unbalanced Paired-Comparison or Ranked Data. Ph.D. Thesis, Iowa State University, Ames. 1989.

[2] D. M. Andrews and H. A. David. Nonparametric Analysis of Unbalanced Paired-Comparison or Ranked Data. *J. Amer. Statist. Ass.*, **85**:1140-1146, 1990.

[3] P. Y. Chebotariov. Generalization of the Row Sum Method for Incomplete Paired Comparisons (Russian). *Avtomat. i Telemekh.* **50**, No. 8, 125-137. (English) *Automation and Remote Control* **50**, 1103-1113. 1989.

[4] D. J. Cowden. A Method of Evaluating Contestants. *The American Statistician*, **29**:82-84, 1974.

[5] E. L. Crow. Ranking Paired Contestants. *Comm. Statist. - Simula.,* **19**:749-769, 1990.

[6] H. E. Daniels. Round-Robin Tournament Scores. *Biometrika*, **56**:295-299, 1969.

[7] H. A. David. Ranking from Unbalanced Paired-Comparison Data. *Biometrika*, **74**:432-436, 1987.

[8] H. A. David. *The Method of Paired Comparisons.* Oxford University Press, New York, Second Edition, 1988 and Charles Griffin and Company, London, First edition, 1963.

[9] R. A. Groeneveld. Ranking Teams in a League with Two Divisions of t Teams. *Amer. Statist.,* **44**:277-281, 1990.

[10] H. Gulliksen. A Least Squares Solution for Paired Comparisons with Incomplete Data. *Psychometrika*, **21**:125-134, 1956.

[11] H. F. Kaiser and R. C. Serlin. Contributions to the Method of Paired Comparisons. *Applied Psychological Measurement*, **2**:421-430, 1978.

[12] M. G. Kendall. Further Contributions to the Theory of Paired Comparisons. *Biometrics*, **11**:43-62, 1955.

[13] J. W. Moon. *Topics on Tournaments.* Holt, Rinehart and Winston, New York. 1968

[14] J. W. Moon and N. J. Pullman. On Generalized Tournament Matrices. *SIAM Rev.*, **12**:384-399, 1970.

[15] S. Nishisato. *Analysis of Categorical Data: Dual Scaling and its Applications.* University of Toronto Press. Toronto, 1980.

[16] M. J. Prentice. On the Problem of m Incomplete Rankings. *Biometrika*, **66**:167-170, 1979.

[17] T. A. Prigarina, P. Y. Chebotariov and D. S. Schmerling. A Review of Some Papers on Paired Comparisons in the U.S.S.R. Unpublished manuscript. 1991.

[18] C. Ramanujacharyulu. Analysis of Preferential Experiments. *Psychometrika*, **29**:257-261, 1964.

[19] E. Seneta. *Non-Negative Matrices.* Wiley, New York, 1973.

[20] P. Slater. Inconsistencies in a Schedule of Paired Comparisons. *Biometrika*, **48**:303-312, 1961.

[21] M. Thompson. On any Given Sunday: Fair Competition Orderings with Maximum Likelihood Methods. *J. Amer. Statist. Ass.*, **70**:739-747, 1975.

[22] W. A. Thompson, Jr. and R. Remage, Jr. Rankings from Paired Comparisons. *Ann. Math. Statist.*, **35**: 1964.

[23] K. M. Wittkowski. Small Sample Properties of Rank Tests for Incomplete Unbalanced Designs. *Biometric Journal*, **30**:799-808, 1988a.

[24] K. M. Wittkowski. Friedman-Type Statistics and Consistent Multiple Comparisons for Unbalanced Designs with Missing Data. *J. Amer. Statist. Ass.*, **83**:1163-1170, 1988b.

3

On the Babington Smith Class of Models for Rankings

Harry Joe [1]
Joseph S. Verducci [2]

ABSTRACT In 1950, Babington Smith proposed a general family of probability models for rankings based on a paired comparisons idea. Mallows [9] studied several simple subclasses of the Babington Smith models, but the full class was considered computationaly intractible for practical application at that time. With modern computers, the models are simple to use. With this incentive, we investigate various properties of the Babington Smith models, including their characterization as maximum entropy models, the relationships among different parametrizations of the models, and the conditions under which various forms of stochastic transitivity, unimodality and consensus are obtained. The maximum entropy characterization suggests models that are nested within the Babington Smith models and models that are more general. Computational details for the models are briefly discussed. The models are illustrated with examples where words are ranked in accordance to their perceived degree of association with a target word.

3.1 Introduction

Consider the situation in which k items are to be independently ranked by a sample of judges. For example the judges might be graduate students in statistics, the items career options, and the criterion for ranking the order of personal preference.

A popular method for analyzing preferences is the method of paired comparisons (see David, [3], for a detailed introduction to this topic). The basic format of this method is to present each judge with all possible pairs $\{i, j\}$ of items from the set $\{1, ..., k\}$ of items, and, for each $i < j$, record $I_{ij} = 1$ if item i is preferred to item j and $I_{ij} = 0$ otherwise. For each judge J , let $P_J(i, j) = P(I_{ij} = 1|J)$ be the probability that judge J will prefer item i to item j. The model of paired comparisons assumes that, for judge

[1] Department of Statistics, University of British Columbia, Canada
[2] Department of Statistics, Ohio State University, Columbus, Ohio

J, the $\{I_{ij}\}$ are independent Bernoulli trials with parameters $P_J(i,j)$. Very often it is also assumed that the population of judges is homogeneous in the sense that the probabilities $P_J(i,j) = \alpha_{ij}$ are the same for all judges.

In making paired comparisons, a judge may or may not be consistent with a particular ranking of the items. Define a *ranking* of items to be a one–to–one mapping of $\{1,...,k\}$ to $\{1,...,k\}$, where $\pi(i) = a$ if item i is given rank a. In accordance with popular convention, we think of rank 1 as "best", etc. A set $\{I_{ij} : 1 \le i < j \le k\}$ of paired comparisons is *consistent with a ranking* π if $I_{ij} = 1$ whenever $\pi(i) < \pi(j)$.

Associated with any homogeneous paired comparison model $\{\alpha_{ij}\}$, Babington Smith [10] suggested the ranking model whose probabilities $P(\pi)$ are proportional to

$$\prod_{i<j}[\alpha_{ij}]^{I_{ij}(\pi)}[1-\alpha_{ij}]^{1-I_{ij}(\pi)}, \tag{1}$$

where $I_{ij}(\pi) = 1$ if $\pi(i) < \pi(j)$ and 0 otherwise. In words, the probability of any ranking π is the probability that a judge produces a set of paired comparisons consistent with the ranking π, conditional on the judge producing a set of paired comparisons consistent with some ranking. Under the transformation $\theta_{ij} = \alpha_{ij}/(1-\alpha_{ij})$, $i < j$, the model may be written in the form

$$P(\pi) = \prod_{i<j}(\theta_{ij})^{I_{ij}(\pi)}/C(\boldsymbol{\theta}), \tag{2}$$

where $0 < \theta_{ij} < \infty$ for each $i < j$, $\theta = (\theta_{ij}, i < j)$, and

$$C(\boldsymbol{\theta}) = \sum_{\nu}\prod_{i<j}(\theta_{ij})^{I_{ij}(\nu)},$$

is a normalizing constant, the sum being taken over all $k!$ possible rankings. The model thus has the form of an exponential family with canonical parameters $\log(\theta_{ij})$ and sufficient statistics $\{I_{ij}\}$. We will call this model, defined by either (1) or (2), the B-S model.

Let τ_{ij} denote the transposition of items i and j, so that for any ranking π, $\pi \circ \tau_{ij}(i) = \pi(j)$, $\pi \circ \tau_{ij}(j) = \pi(i)$, and $\pi \circ \tau_{ij}(m) = \pi(m)$ for $m \neq i$ or j. If π is such that $\pi(j) = \pi(i) + 1$, then model (2) implies that $\theta_{ij} = P(\pi)/P(\pi \circ \tau_{ij})$. Thus θ_{ij} may be interpreted as the odds that item i is preferred to item j when these items are ranked adjacently. Note that these odds do not depend on the ranks given to the other items. This property, called *independence of irrelevant alternatives* was studied by Luce [8].

The class of B-S models has two other appealing properties, *label invariance* and *reversibility.* That is, if $P(\pi)$ satisfies (2) and τ is any permutation of $\{1,\ldots,k\}$ representing a change of labels, then $P(\pi \circ \tau)$ satisfies (2) with θ_{ij} replaced by $\theta_{\tau(i),\tau(j)}$. Also, if γ is the permutation $\gamma(i) = (k+1) - i$ and $P(\pi)$ satisfies (2), then $P(\gamma \circ \pi)$ also has the form (2) with θ_{ij} replaced by $\theta_{ji} = 1/\theta_{ij}$. Thus reversing the meaning of the ranks, so that a rank of 1 indicates the "worst" item, also leads to a B-S model.

In addition to the above properties, the B-S model is a maximum entropy model. For $i \neq j$, let M_{ij} denote the marginal probability $P(\pi(i) < \pi(j))$. Then the model (2) maximizes the (Shannon) entropy $-\sum_\nu P(\nu)$ $\log[P(\nu)]$ among all distributions of rankings with fixed margins $\{M_{ij}\}$. This maximum entropy characterization suggests models that are submodels of (2), and also models that are more general. These are mentioned in Section 3.2, which includes computational details for the models. The natural derivation from paired comparisons and its appealing properties make the B-S model potentially very useful. However, because of previous computational difficulties, most past studies have focused on special simple forms of the B-S model.

In Section 3.3, we identify various subclasses of B-S models (equivalently, constraints on the parameters) that conform to notions of stochastic transitivity, unimodality and consensus. Some of this work was suggested by Critchlow, Fligner and Verducci [2]. Finally, examples and data analyses are presented in Section 3.4; in the examples, words are ranked in accordance to their perceived degree of association with a target word.

3.2 Alternative Parametrizations and Related Models

As is well known (see, for example, Brown, [1], p. 74) an exponential family may be parametrized by either the set of expectation parameters ($\{M_{ij}\}$ for B-S model) or the set of canonical parameters $\{\log(\theta_{ij})\}$. The expectation parameters are easily estimated from the data, but most properties of the model depend explicitly upon the canonical parameters. In this section, we make some connections between the two sets of parameters that are specific to the B-S model. Then we go on to define models that are nested in the B-S model and models that include the B-S model.

Although the B-S model is full in the sense that the canonical parameters $\{\log(\theta_{ij})\}$ are unconstrained, the marginal probabilities $\{M_{ij}\}$ are nevertheless constrained by the fact that all implied paired comparisons are consistent with rankings. The following theorem makes the constraints explicit.

Theorem 3.2.1. For any items i, j, m,

$$M_{im} + M_{mj} - 1 \leq M_{ij} \leq M_{im} + M_{mj}. \tag{3}$$

Proof. $M_{ij} = P[\pi(i) < \pi(j)] \geq P[\pi(i) < \pi(m) < \pi(j)] = M_{im} + M_{mj} - P[\{\pi(i) < \pi(m)\} \cup \{\pi(m) < \pi(j)\}] \geq M_{im} + M_{mj} - 1$. Similarly, $M_{ji} \geq M_{jm} + M_{mi} - 1$ implies that $1 - M_{ij} \geq (1 - M_{mj}) + (1 - M_{im}) - 1$ or $M_{ij} \leq M_{im} + M_{mj}$. ⊔

Of course the marginal probabilities are also constrained by $0 \leq M_{ij} \leq 1$, so that the inequality (3) is nontrivial whenever $M_{im} + M_{mj} \neq 1$. Let $\mathcal{M}$ be the set of $\{M_{ij}\}$ that are consistent with a probability distribution on rankings. It follows from standard exponential family theory that the maximal entropy distribution P on rankings, subject to the constraints that $P[\pi(i) < \pi(j)] = M_{ij}$ for each $i < j$, is given by (2) whenever $\{M_{ij}\} \in \text{interior}(\mathcal{M})$. In this case, also, there is a one–to–one correspondence between the marginal probabilities $\{M_{ij}\}$ and the conditional probabilities $\{\alpha_{ij}\}$. Although this relationship is fairly complicated, the following theorem, which is intuitively correct, is valid.

Theorem 3.2.2. For $i < j$, $\partial M_{ij}/\partial\theta_{ij} > 0$.

Proof. Let $C' = \partial C(\theta)/\partial\theta_{ij}$. Note that C' does not depend on θ_{ij} and $M_{ij} = \theta_{ij}C'/C(\theta)$. Hence,

$$\frac{\partial M_{ij}}{\partial\theta_{ij}} = \frac{C'}{C(\theta)} - \theta_{ij}\left(\frac{C'}{C(\theta)}\right)^2 = \theta_{ij}^{-1}M_{ij}(1 - M_{ij}) > 0.$$

⊔

A statistical property of model (2), which follows from exponential family theory is given below, but we consider it in the following more general context. For a data set of rankings with a sample of size n, let $n(\pi)$ be the (observed) frequency of π. The exponential family model (for a distribution of rankings) with the minimal sufficient statistic (dimension t),

$$(S_1, \ldots, S_t) = n^{-1}\left(\sum_{\pi} n(\pi)s_1(\pi), \ldots, \sum_{\pi} n(\pi)s_t(\pi)\right),$$

is

$$\exp\{-\gamma + \sum_{\ell=1}^{t} \lambda_\ell s_\ell(\pi)\}, \tag{4}$$

where $\gamma = \gamma(\lambda_1, \ldots, \lambda_t) = \log(\sum_\pi \exp\{\sum_\ell \lambda_\ell s_\ell(\pi)\})$. Let $C = C(\lambda_1, \ldots, \lambda_t) = \exp(\gamma)$.

The B-S model has the $t = k(k-1)/2$ dimensional minimal sufficient statistic $\{n^{-1}\sum_\pi n(\pi)I_{ij}(\pi), i < j\}$. The Mallows–Bradley–Terry (MBT) model (see Mallows, [9]) has the sufficient statistic, $\{n^{-1}\sum_\pi n(\pi)\pi(i),\ i = 1, \ldots, k\}$, which is the vector of average ranks. Since $I_{ji}(\pi) = 1 - I_{ij}(\pi)$ and $\pi(i) = k - \sum_{j \neq i} I_{ij}(\pi)$, the MBT model is a special case of (2). Since $\sum_i \pi(i) = k(k+1)/2$ for all rankings π, the dimension of the minimal sufficient statistic is $k-1$ for the MBT model, and the MBT model can be written as

$$\exp\{-\gamma + \sum_{i=1}^{k-1} \lambda_i[k - \pi(i)]\} \tag{5}$$

(cf. Critchlow, Fligner and Verducci, [2]), or the more symmetric form

$$\prod_{i=1}^{k} \theta_i^{k-\pi(i)}, \tag{6}$$

with the θ_i chosen so that $\sum_\pi \prod_{i=1}^k \theta_i^{k-\pi(i)} = 1$. In this latter parametrization, a better (lower) ranked item has a larger value of θ.

Models that are nested between (6) and (2) have a minimal sufficient statistic with components being either pairwise preference proportions or average rankings within a subset of $\{1,\ldots,k\}$. For example, let A be a subset of $\{1,\ldots,k\}$ with cardinality at least 3; then an exponential family model exists with minimal sufficient statistic

$$\left\{ n^{-1}\sum_\pi n(\pi) \sum_{j\neq i, j\in A} I_{ij}(\pi),\ i\in A;\quad n^{-1}\sum_\pi n(\pi) I_{ij}(\pi),\ i,j\notin A \right\}.$$

An example with data is given in Section 3.4.

If the model (2) does not provide a good fit, then one could try an exponential family model having a minimal sufficient statistic with some components of the form $n^{-1}\sum_\pi n(\pi)I[\pi(i) < \pi(m) < \pi(j)]$ and other components of the form $n^{-1}\sum_\pi n(\pi)I_{ij}(\pi)$.

The exponential family form (4) is the simplest computational form for all of the models mentioned here. The log–likelihood is $L = -n\gamma + n\sum_{\ell=1}^t \lambda_\ell S_\ell$. The likelihood equations are

$$\frac{\partial L}{\partial \lambda_v} = -n\frac{\partial \gamma}{\partial \lambda_v} + nS_v = 0, \quad v = 1,\ldots,t, \tag{7}$$

where $\partial\gamma/\partial\lambda_v = C^{-1}\sum_\pi s_v(\pi)\exp\{\sum_\ell \lambda_\ell s_\ell(\pi)\}$. The second derivatives are

$$\frac{\partial^2 L}{\partial\lambda_v\partial\lambda_w} = -n\frac{\partial^2\gamma}{\partial\lambda_v\partial\lambda_w}$$
$$= nC^{-2}\sum_\pi s_v(\pi)\exp\{\sum_\ell \lambda_\ell s_\ell(\pi)\}\cdot\sum_\pi s_w(\pi)\exp\{\sum_\ell \lambda_\ell s_\ell(\pi)\}$$
$$-nC^{-1}\sum_\pi s_v(\pi)s_w(\pi)\exp\{\sum_\ell \lambda_\ell s_\ell(\pi)\}. \tag{8}$$

Note that (7) are also moment or maximum entropy equations. Numerical solving of (7) is straightforward using (8) in Newton–Raphson iterations.

Once the maximum likelihood estimates $\hat\lambda_\ell$ are obtained, the parameters can be converted to a form of the model that is more interpretable. For example, consider the B-S model in form (2). The special case of the MBT model arises when $\theta_{ij} = \eta_i/\eta_j$ for some constants $\eta_1,\ldots,\eta_k$. A model nested between the B-S model and the MBT model arises when $\theta_{ij} = \eta_i/\eta_j$,

$i, j \in A$, for some constants η_i, $i \in A \subset \{1, \ldots, k\}$. The form (5) of the MBT model can be converted to the form (6) with $\theta_k = D = C^{-2/k(k-1)} = \exp\{-2\gamma/k(k-1)\}$, $\theta_i = D\alpha_i = D\exp\{\lambda_i\}$, $i = 1, \ldots, k-1$. With $\eta_i = \theta_i^{1/2}$, the MBT model can be written in the form $B^{-1}\prod_{i \neq j}(\eta_i/\eta_j)^{I_{ij}(\pi)}$ to compare with (2). That is, one way of determining the closeness of the B-S and MBT models for some data is by comparing $\hat{\theta}_{ij}$ with $(\hat{\theta}_i/\hat{\theta}_j)^{1/2}$, where $\hat{\theta}_{ij}$ are maximum likelihood estimates for (2) and $\hat{\theta}_i$ are maximum likelihood estimates for (6). An example of this is given in Section 3.4.

3.3 Stochastic Transitivity and Item Preference

In this section, we apply stochastic transitivity concepts from paired comparisons (see David, [3]) to the B-S model. Relations are obtained between transitivity conditions on the M_{ij} and the θ_{ij}. Various ideas of item preference and consensus are discussed and related to parametric restrictions in either the $\{M_{ij}\}$ or $\{\theta_{ij}\}$ parametrization of model (2). We start with some definitions.

Definitions. (Transitivity.) Let $\{p_{ij}, i \neq j\}$, be a set of "paired comparisons" probabilities. The set $\{p_{ij}\}$ is *weakly stochastically transitive* if $p_{ij} > 1/2$ and $p_{jm} > 1/2$ imply that $p_{im} > 1/2$. The set $\{p_{ij}\}$ is *strongly stochastically transitive* if $p_{ij} \geq 1/2$ and $p_{jm} \geq 1/2$ imply that $p_{im} \geq \max\{p_{ij}, p_{jm}\}$.

One application of the above definitions is to the $\{M_{ij}\}$. They could also be applied to the $\{\alpha_{ij}\}$ or the $\{\theta_{ij}\}$, with $\theta_{ii} = 1$ and $\theta_{ji} = 1/\theta_{ij}$ for $i < j$. Note that the θ_{ij}'s in (2) were defined only for $i < j$, but its interpretation as the odds of the event $\{\pi(i) < \pi(j)\}$ conditioned on the event $\{|\pi(i) - \pi(j)| = 1\}$ suggests this extended definition of θ_{ij}. With this definition, weak stochastic transitivity in the $\{\theta_{ij}\}$ holds if $\theta_{ij} > 1$ and $\theta_{jm} > 1$ imply that $\theta_{im} > 1$. Similarly, strong stochastic transitivity in the $\{\theta_{ij}\}$ holds if $\theta_{ij} \geq 1$ and $\theta_{jm} \geq 1$ imply that $\theta_{im} \geq \max\{\theta_{ij}, \theta_{jm}\}$.

A link between strong stochastic transitivity in $\{\alpha_{ij}\}$ and $\{M_{ij}\}$ is given in the Theorem 3.3.3 below. First the following results are needed.

The next lemma relates the θ parameters to the odds that item i is preferred to item j, conditional on each of the other items receiving a fixed rank. This lemma is used in Theorem 3.3.2 to relate the θ parameters to the α parameters within a special subfamily of the B-S model.

Lemma 3.2.2. Let π be any ranking such that $\pi(i) < \pi(j)$, then under model (2),

$$P(\pi)/P(\pi \circ \tau_{ij}) = \theta_{ij}/c_\pi \tag{9}$$

where

$$c_\pi = \prod_{m \neq i,j} (\theta_{jm}/\theta_{im})^{I_{imj}(\pi)} \tag{10}$$

with I_{imj} being the indicator of the event $\{\pi(i) < \pi(m) < \pi(j)\}$.

Proof. Without loss of generality, assume that $i = 1$ and $j = 2$, and let π be any ranking such that $\pi(1) < \pi(2)$. From (2) it follows that

$$P(\pi)/P(\pi \circ \tau_{12}) = \prod_{r=1}^{k-1} \prod_{m=r+1}^{k} (\theta_{rm})^{[I_{rm}(\pi) - I_{rm}(\pi \circ \tau_{12})]}$$

$$= \theta_{12} \prod_{r=1}^{2} \prod_{m=3}^{k} (\theta_{rm})^{[I_{rm}(\pi) - I_{rm}(\pi \circ \tau_{12})]}$$

since $\pi \circ \tau_{12}(r) = \pi(r)$ for $r = 3, \ldots, k$, $I_{12}(\pi) = 1$ and $I_{12}(\pi \circ \tau_{12}) = 0$. Also note that $I_{1m}(\pi)$ differs from $I_{1m}(\pi \circ \tau_{12})$ only when $\pi(1) < \pi(m) < \pi(2)$, in which case $I_{1m}(\pi) = 1$ while $I_{1m}(\pi \circ \tau_{12}) = 0$. A similar observation for $r = 2$ leads to

$$P(\pi)/P(\pi \circ \tau_{12}) = \theta_{12} \prod_{m=3}^{k} (\theta_{1m}/\theta_{2m})^{I_{1m2}(\pi)}$$

which has the form (9) for $i = 1$ and $j = 2$. ⊔

Definition. Let $\mathcal{T}$ consist of the family of probability functions of the form (2) where for each $i = 1, \ldots, k$, θ_{ij} is increasing (nondecreasing) in j.

Note that θ_{ij} increasing in j for all i is equivalent to the constraint that θ_{ij} is decreasing in i for all j. Intuitively, this constraint implies that items are ordered according to their numerical labels in that $i < j$ implies that the conditional odds that item i is preferred to item j are greater than or equal to 1, and that these odds increase as item i is compared with items having successively larger indices.

The next theorem shows that, within the family $\mathcal{T}$, conditioning on the event $\{|\pi(i) - \pi(j)| = 1\}$ attenuates the marginal probabilities of $\{\pi(i) < \pi(j)\}$.

Theorem 3.3.2. If $P \in \mathcal{T}$, then $1/2 \leq \alpha_{ij} \leq M_{ij}$ whenever $i < j$. If $1/2 < \alpha_{ij}$, then $\alpha_{ij} < M_{ij}$.

Proof. The increasing pattern of the θ_{ij} over j together with $\theta_{ii} = 1$ implies that $\theta_{ij} \geq 1$ for $i < j$. Thus $\alpha_{ij} = \theta_{ij}/(\theta_{ij} + 1) \geq 1/2$.

For $i < j$, let $S_{ij} = \{\pi : \pi(i) < \pi(j)\}$, $T_{ij} = \{\pi : |\pi(i) - \pi(j)| = 1\}$, $A = S_{ij} \cap T_{ij}$ and $B = S_{ij} \backslash A$. Then

$$M_{ij}/(1 - M_{ij}) = \sum_{S_{ij}} P(\pi) / \sum_{S_{ij}} P(\pi \circ \tau_{ij}) = \frac{\sum_A P(\pi) + \sum_B P(\pi)}{\theta_{ji} \sum_A P(\pi) + \theta_{ji} \sum_B c_\pi P(\pi)},$$

where c_π is given by (10). By assumption, θ_{ij} is increasing in j, and thus decreasing in i since $\theta_{ji} = 1/\theta_{ij}$. Hence $c_\pi \leq 1$ and equation (3) then implies that $M_{ij}/(1-M_{ij}) \geq \theta_{ij} = \alpha_{ij}/(1-\alpha_{ij})$ or $\alpha_{ij} \leq M_{ij}$. All the inequalities are strict if $\alpha_{ij} > 1/2$. $\sqcup$

Theorem 3.3.3. Strong stochastic transitivity of $\{\alpha_{ij}\}$ in the B-S model implies that $\{M_{ij}\}$ is also strongly stochastically transitive.

Proof. Without loss of generality, assume the natural ordering for the stochastic transitivity of $\{\alpha_{ij}\}$, that is, $\alpha_{ij} \geq 1/2$ if $i < j$ and $P \in \mathcal{T}$. Fix $i < m < j$. By Theorem 3.3.2, $M_{im} \geq 1/2$ and $M_{mj} \geq 1/2$. The conclusion follows once we show $M_{ij} \geq M_{im}$ and $M_{ij} \geq M_{mj}$. By symmetry, we prove only the former. Note that $M_{ij} = P[\pi(m) < \pi(i) < \pi(j)] + P[\pi(i) < \pi(m) < \pi(j)] + P[\pi(i) < \pi(j) < \pi(m)] \geq M_{im} = P[\pi(j) < \pi(i) < \pi(m)] + P[\pi(i) < \pi(m) < \pi(j)] + P[\pi(i) < \pi(j) < \pi(m)]$ if and only if $P[\pi(m) < \pi(i) < \pi(j)] \geq P[\pi(j) < \pi(i) < \pi(m)]$.

Let ν be any ranking such that $\nu(m) < \nu(j)$. By Lemma 3.2.2, $P(\nu)/P(\nu \circ \tau_{mj}) = \theta_{mj}/c_\nu$, and $c_\nu = \prod_{r \neq m,j} (\theta_{jr}/\theta_{mr})^{I_{mrj}(\pi)} \leq 1$ follows from the definition of $\mathcal{T}$. Since $\theta_{mj} \geq 1$, it follows that $P(\nu) \geq P(\nu \circ \tau_{mj})$. By summing over ν such that $\nu(m) < \nu(i) < \nu(j)$, $P(\pi(m) < \pi(i) < \pi(j)) \geq P(\pi(j) < \pi(i) < \pi(m))$. $\sqcup$

We now go on to concepts involving item preference and consensus. The idea of consensus should be distinguished from the more familiar concept of concordance. Kendall [7] proposed the well known 23 W statistic as an index of concordance or agreement among judges in a sample. This and other measures of concordance are related to the average correlation between randomly sampled pairs of judges (see Fligner and Verducci [5] for a review of concordance measures from this point of view). Unfortunately, a population may display a high degree of concordance even though distinct subpopulations tend to disagree.

On the other hand, the notion of consensus, defined formally below, implies that the probability function is unimodal in terms of a certain metric on the set of all rankings. Further homogeneity is also implied in that any subpopulation defined in terms of its ranking pattern on a subset of items, will also make the same majority choices as the whole population regarding the relative ranking of the other items. Lack of consensus suggests that the population may be better modeled as a mixture of more homogeneous subpopulations.

The following definitions follow David [3] and Fligner and Verducci [6].

Definitions. (Item preference). Item i is *weakly preferred* to item j, written $i \succ_w j$, if $P[\pi(i) < \pi(j)] > 1/2$. Item i is *preferred in expectation* to item j, written $i \succ_e j$, if $E[\pi(i)] < E[\pi(j)]$. Item i is *strongly preferred* to item j, written $i \succ_s j$, if for any ranking π such that $\pi(i) < \pi(j)$, $P(\pi) \geq P(\pi \circ \tau_{ij})$, with strict inequality for at least one ranking π.

It is easily verified that neither weak preference nor preference in expectation implies the other, but that strong preference implies both weak preference and preference in expectation. In terms of model (2), weak preference and preference in expectation are easily characterized in terms of the expectation parameters $\{M_{ij}\}$, whereas strong preference is more easily characterized by the canonical parameters. By definition $i \succ_w j$ if and only if $M_{ij} > 1/2$. Almost as directly, $i \succ_e j$ if and only if $\sum_{m \neq i} M_{im} > \sum_{m \neq j} M_{jm}$. If $\{M_{ij}\}$ is weakly stochastic transitive, the induced preference ordering of the items is the same as the ordering by weak preference but not necessary the same as ordering by preference in expectation. If $\{M_{ij}\}$ is strongly stochastic transitive (and no M_{ij} is equal to 1/2), the induced preference ordering of the items is the same as the ordering by weak preference and preference in expectation.

The strong preference ordering requires stronger conditions. The following theorem gives necessary and sufficient conditions on the parameters of the B-S model for item i to be strongly preferred to item j.

Theorem 3.3.4. For the B-S model, item i is strongly preferred to item j if and only if

$$\theta_{ij} \geq \prod_{m \in S} (\theta_{jm}/\theta_{im}) \tag{11}$$

for all subsets S of $\{1, \ldots, k\} \backslash \{i, j\}$, with strict inequality for some subset S. For S empty, the right hand side of (11) is interpreted as 1.

Proof. Let π be a ranking such that $\pi(i) < \pi(j)$, and express $P(\pi)/P(\pi \circ \tau_{ij}) = \theta_{ij}/c_\pi$ as in Lemma 3.2.2. For any subset S, there exists π such that $S = \{m : \pi(i) < \pi(m) < \pi(j)\}$; for this π, it follows that $P(\pi)/P(\pi \circ \tau_{ij}) \geq 1$ if and only if (11) holds. Strict inequality holds for some π if and only if (11) holds with strict inequality for some subset S. ⊔

The relationship of weak preference does not necessarily provide a linear ordering of the items $1, \ldots, k$, because weak preference may include circularities such as $i \succ_w j$, $j \succ_w m$, and $m \succ_w i$. On the other hand, preference in expectation is necessarily transitive in the sense that $i \succ_e j$ and $j \succ_e m$ imply that $i \succ_e m$. Finally, strong preference need not be transitive (see Fligner and Verducci, [6], for an example), but does preclude circularities, since it implies preference in expectation.

The following definitions lead to a geometrical characterization of weak stochastic transitivity of the $\{\alpha_{ij}\}$.

Definitions. (Unimodality). The *tau-distance* between any two rankings π and ν, is defined as the number of discordances

$$d(\pi, \nu) = \sum_{i<j} I([\pi(i) - \pi(j)][\nu(i) - \nu(j)] < 0)$$

between pairs of items, where $I(A) = 1$ or 0, is the indicator of event A. A probability function $P(\pi)$ on rankings is said to be *strongly unimodal* with mode π_0 if $d(\pi_0, \pi) < d(\pi_0, \nu)$ implies $P(\pi) \geq P(\nu)$ whenever $d(\pi, \nu) = 1$, with strict inequality in the case that $\pi = \pi_0$.

Theorem 3.3.5. For the B-S model, $P(\pi)$ is strongly unimodal with modal ranking $e = [12\cdots k]$ if and only if $\theta_{ij} \geq 1$ when $i < j$, with strict inequality when $j = i + 1$.

Proof. Let $i < j$ and choose any π such that $\pi(i) = \pi(j) - 1$. Then $d(\pi, \pi \circ \tau_{ij}) = 1$ and

$$P(\pi) = \theta_{ij} P(\pi \circ \tau_{ij}), \tag{12}$$

so that $\theta_{ij} \geq 1$ is necessary for strong unimodality. For $\pi = e$, $d(\pi, \pi \circ \tau_{ij}) = 1$ when $j = i + 1$ shows that $\theta_{i,i+1} > 1$ is also necessary.

Conversely, suppose that the conditions on the θ_{ij} are satisfied. First suppose that π and ν are two rankings with $d(\pi, \nu) = 1$ and $d(e, \pi) < d(e, \nu)$. These conditions imply that $\nu = \pi \circ \tau_{ij}$ for some items $i < j$ with $\pi(i) = \pi(j) - 1$. It then follows from (12) that $P(\pi) \geq P(\nu)$ whenever $\theta_{ij} \geq 1$. For $\pi = e$, $d(e, \nu) = 1$ implies that $\nu = \tau_{ij}$ for $j = i + 1$, whence (12) and $\theta_{i,i+1} > 1$ imply that $P(e) > P(\nu)$. ⊔

The following corollary, whose proof follows directly from Theorem 3.3.5, was proved through a somewhat different argument in Critchlow, Fligner and Verducci [2].

Corollary. For the B-S model, suppose that none of the M_{ij} are equal to 1/2. Then the set $\{\alpha_{ij}\}$ is weakly stochastically transitive if and only if P is strongly unimodal.

Thusfar we have exhibited the relationships among the four subfamilies B-S models defined by weak and strong stochastic transitivity of the $\{\alpha_{ij}\}$ and $\{M_{ij}\}$ parameters. We conclude this section by relating these subfamilies to the following notions of consensus from Fligner and Verducci [6].

Definitions. (Consensus). A population is said to have a *simple consensus* with consensus ranking ν if $i \succ_s j$ for $\nu(j) = \nu(i) + 1$. A population is said to have *complete consensus* with consensus ranking ν if $i \succ_s j$ for $\nu(i) < \nu(j)$.

Simple consensus does not imply complete consensus, even within the context of the B-S model. An example with $k = 4$ and $\nu = [1234]$ is for the B-S model with $\theta_{12} = 1.2$, $\theta_{13} = \theta_{14} = \theta_{23} = 1.1$, $\theta_{24} = \theta_{34} = 1.05$. It is easy to check that the inequalities in (11) are valid for $(i, j) = (1, 2)$, (2,3) and (3,4) but not for (2,4) because θ_{24} is less than θ_{41}/θ_{21}.

The following theorem on complete consensus, proved in a different development by Critchlow, Fligner and Verducci [2] follows directly from Theorem 3.3.4 and the definition of $\mathcal{T}$.

Theorem 3.3.6. Let $P \in \mathcal{T}$ be such that θ_{ij} is strictly increasing in j for each $i = 1, \ldots, k$. Then P has complete consensus with consensus ranking $e = [12\cdots k]$.

There are induced item preference orderings from the concepts in this section. The relationships are summarized in the following diagram. In the diagram, SST stands for strong stochastic transitivity and WST stands for weak stochastic transitivity. Enough of the θ_{ij} are assumed to be not equal to 1 in order that a complete ordering exists. A similar remark holds for the M_{ij}.

$$\begin{tikzcd}
\mathrm{SST}(\theta_{ij}) \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \mathrm{WST}(\theta_{ij}) \arrow[d, Leftrightarrow] \\
\text{complete concensus} \arrow[d, Rightarrow] & \text{strong unimodality} \\
\mathrm{SST}(M_{ij}) \arrow[r, Rightarrow] \arrow[d, Rightarrow] & \mathrm{WST}(M_{ij}) \\
\succ_e \text{ ordering} &
\end{tikzcd}$$

Note that for the MBT model, the identities $\theta_{ij}\theta_{jm} = \theta_{im}$ are satisfied and $\{\theta_{ij}\}$ is strongly stochastic transitive. It is not hard to show that there are no other implications in the above diagram. A strongly unimodal distribution which satisfies $\mathrm{SST}(M_{ij})$ can have a mode different from the ordering induced by the strong stochastic transitivity. An example with $k = 3$ is with $\theta_{12} = \theta_{13} = 1.01$ and $\theta_{23} = 2$; this is strongly unimodal with mode [123] but $M_{21} = .552$, $M_{13} = .559$, $M_{23} = .667$.

3.4 Examples and Data Analysis

The Graduate Record Examination Board have samples of college students, where the students were asked to rank five words according to strength of association with a target word. We have used two such samples for illustration here. The log–likelihoods and maximum likelihood estimates given below were obtained computationally using the exponential family form (4) and then transformed.

In the first example, the target word is "skunk", and the five choices were labeled (1) camel, (2) porcupine, (3) lion, (4) cat, (5) hound. In the rankings, 1 means the least associated with the target word. The sample size was $n = 124$. The observed and expected frequencies from three models are given in Table 1; only rankings with non–zero observed frequency are listed. The average ranks for the items are (1.52, 4.26, 2.61, 3.77, 2.84). The maximum likelihood estimates for the MBT model in the form (6) are $(\hat{\theta}_1, \ldots, \hat{\theta}_5) = (1.36, 0.24, 0.63, 0.34, 0.56)$ and the log–likelihood is -456.6. The log–likelihood for the B-S model is -439.5. The matrices of $\hat{M}_{ij}$ and $\hat{\theta}_{ij}$ are respectively

$$\begin{bmatrix} -- & .92 & .84 & .88 & .84 \\ .08 & -- & .15 & .35 & .16 \\ .16 & .85 & -- & .81 & .56 \\ .12 & .65 & .19 & -- & .27 \\ .16 & .84 & .44 & .73 & -- \end{bmatrix}, \quad \begin{bmatrix} 1.00 & 1.27 & 5.48 & 1.25 & 3.65 \\ 0.79 & 1.00 & 0.38 & 0.58 & 0.30 \\ 0.18 & 2.62 & 1.00 & 4.05 & 1.23 \\ 0.80 & 1.73 & 0.25 & 1.00 & 0.53 \\ 0.27 & 3.34 & 0.81 & 1.89 & 1.00 \end{bmatrix}$$

With 6 extra degrees of freedom and twice the difference in log– likelihood being 34.2, the B-S model fits the data much better than the MBT model. This is confirmed by looking at the expected frequencies for the 2 models (columns 3 and 4 of Table 1). The modal observed frequency is 22 for the ranking [15243]; the corresponding expected frequencies for the MBT model and B-S model are 11.2 and 17.4 respectively.

The observed preference proportion matrix $(\hat{M}_{ij})$ is strongly stochastic transitive with the induced ordering of items being 1,3,5,4,2 (this agrees with the ordering of the mode). The $(\hat{\theta}_{ij})$ matrix is weakly stochastic transitive and is not close to being strongly stochastic transitive. However the submatrix based on items 3,5,4,2 is almost strongly stochastic transitive and the model with sufficient statistic based on

$$(I_{13}, I_{15}, I_{14}, I_{12}, I_{35} + I_{34} + I_{32}, I_{53} + I_{54} + I_{52}, I_{43} + I_{45} + I_{42}) \qquad (13)$$

is suggested.

For this model with 7 parameters, the log–likelihood is -441.0, which compares well with the log–likelihood for the B-S model. The $\hat{\lambda}_\ell$'s (with index order according to (13)) are 1.69, 1.29, .306, .150, 1.37, 1.12, .370 and $C = 1.97 \times 10^5$. The expected frequencies for this model are in the last column of Table 1, and the modal expected frequency is 16.6. The comparison of the θ_{ij}'s for the 3 models (respectively MBT, B-S, (13)) can be made from the following matrices. In the matrices, the ordering is item 1, item 3, item 5, item 4, item 2.

$$\begin{bmatrix} 1.00 & 1.46 & 1.55 & 2.01 & 2.37 \\ 0.68 & 1.00 & 1.06 & 1.37 & 1.62 \\ 0.64 & 0.94 & 1.00 & 1.29 & 1.52 \\ 0.50 & 0.73 & 0.78 & 1.00 & 1.18 \\ 0.42 & 0.62 & 0.66 & 0.85 & 1.00 \end{bmatrix}, \quad \begin{bmatrix} 1.00 & 5.48 & 3.65 & 1.25 & 1.27 \\ 0.18 & 1.00 & 1.23 & 4.05 & 2.62 \\ 0.27 & 0.81 & 1.00 & 1.89 & 3.34 \\ 0.80 & 0.25 & 0.53 & 1.00 & 1.73 \\ 0.79 & 0.38 & 0.30 & 0.58 & 1.00 \end{bmatrix},$$

$$\begin{bmatrix} 1.00 & 5.44 & 3.63 & 1.36 & 1.16 \\ 0.18 & 1.00 & 1.28 & 2.74 & 3.95 \\ 0.28 & 0.78 & 1.00 & 2.12 & 3.08 \\ 0.74 & 0.37 & 0.47 & 1.00 & 1.45 \\ 0.86 & 0.25 & 0.32 & 0.69 & 1.00 \end{bmatrix}$$

We now go on to the second example. The target word is "song", and the five choices were labeled (1) score, (2) instrument, (3) solo, (4) benediction, (5) suit. In the rankings, again 1 means the least associated with the target word. The sample size was $n = 129$. The observed and expected frequencies for the B-S model are given in Table 2. The average ranks for the items are (2.80, 3.81, 4.64, 2.43, 1.32). The maximum likelihood estimates for the

MBT model in the form (5) are $(\hat{\theta}_1, \ldots, \hat{\theta}_5) = (0.48, 0.23, 0.097, 0.63, 1.93)$ and the log–likelihood is -380.7. The log–likelihood for the B-S model is -351.7. The matrices of $\hat{M}_{ij}$ and $\hat{\theta}_{ij}$ are given below; in the matrices, the order is item 5, item 4, item 1, item 2, item 3.

$$\begin{bmatrix} -- & .91 & .80 & .98 & .99 \\ .09 & -- & .59 & .91 & .98 \\ .20 & .41 & -- & .74 & .84 \\ .02 & .09 & .26 & -- & .83 \\ .01 & .02 & .16 & .18 & -- \end{bmatrix}, \quad \begin{bmatrix} 1.00 & 9.89 & 1.38 & 8.74 & 4.18 \\ 0.10 & 1.00 & 1.37 & 7.56 & 9.95 \\ 0.72 & 0.73 & 1.00 & 2.16 & 1.38 \\ 0.11 & 0.13 & 0.46 & 1.00 & 5.30 \\ 0.24 & 0.10 & 0.72 & 0.19 & 1.00 \end{bmatrix}$$

$(\hat{M}_{ij})$ is weakly stochastic transitive and almost strongly stochastic transitive, and $(\hat{\theta}_{ij})$ is weakly stochastic transitive. No submodel of (2) compared favorably here. The modal observed frequency is 34 for the ranking [35421]. The expected frequency for the B-S model is 28.3; the comparison of other expected frequencies to observed frequencies is quite good.

In both examples, the B-S model underestimates the size of the mode. However the comparison of the expected to observed marginal frequencies of triples, $P[\pi(i) < \pi(m) < \pi(j)]$, is quite good. The only other ranking model that fit the two data sets as well, in terms of log–likelihood, is the multistage ranking model of Fligner and Verducci [6] having the same number of parameters as the B-S model.

Acknowledgements:

This research has been supported by grants from the NSERC Canada and the United States NSF. We are grateful to Arthur Silverberg and the Educational Testing Service for providing the word association data.

Table 1. Observed and expected frequencies (under 3 models) for skunk association data.

ranking	observed freq.	expected freq.		
		MBT	B-S	(13)
15243	22	11.2	17.4	16.6
15342	15	9.9	14.2	12.9
14253	11	8.0	10.1	11.5
13452	10	2.7	3.1	2.3
15234	10	6.7	9.2	7.8
14352	5	7.1	8.2	8.9
13245	4	2.1	1.6	1.8
15423	4	3.2	1.9	2.2
25143	4	5.2	3.2	3.1
13254	3	3.5	3.0	3.7
15432	3	5.2	3.5	4.7
25341	3	4.1	3.9	3.6
14235	2	2.9	2.8	2.5
14325	2	1.5	0.7	0.9
35421	2	0.5	0.8	1.0
53214	2	0.0	0.0	0.0
12543	1	0.6	0.2	0.3
15324	1	3.6	2.3	2.9
21453	1	0.2	0.7	0.6
23154	1	1.6	0.5	0.7
23451	1	1.1	0.9	0.6
24351	1	2.9	2.2	2.5
25134	1	3.1	1.7	1.4
25431	1	2.2	1.0	1.3
32451	1	0.2	0.7	0.5
34125	1	0.3	0.4	0.3
34215	1	0.2	0.1	0.1
34251	1	1.4	0.4	0.5
34512	1	0.1	0.2	0.1
35124	1	0.8	1.4	1.1
35214	1	0.4	0.3	0.4
43251	1	0.2	0.3	0.4
45132	1	0.5	0.7	0.6
51324	1	0.0	0.0	0.0
52314	1	0.0	0.0	0.0
52431	1	0.0	0.0	0.0
53241	1	0.1	0.3	0.3
54312	1	0.0	0.0	0.1

Table 2. Observed and expected frequencies for song association data.

ranking	observed freq.	expected freq. (B-S)
34521	34	28.3
24531	21	20.7
14532	15	14.9
43521	9	13.1
35421	8	5.3
53421	8	9.5
54321	6	1.8
23541	4	2.7
14523	3	1.5
13542	2	2.0
15432	2	2.8
24513	2	2.1
25431	2	3.9
42531	2	1.7
53412	2	1.0
15423	1	0.3
32145	1	0.0
32451	1	0.1
32541	1	1.3
34512	1	2.9
43512	1	1.3
45321	1	3.9
52413	1	0.1
54231	1	0.2

3.5 References

[1] L. Brown. Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *IMS Lecture Notes - Monograph Series*, **9**, 1987.

[2] D. E. Critchlow and Fligner, M. A. and J. S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, **35**:294-318, 1991.

[3] H. A. David. *The Method of Paired Comparisons*. Griffin, London, second edition, 1988.

[4] M. A. Fligner and J. S. Verducci. Distance Based ranking Models. *J. Amer. Statist. Ass. B*, **83**:359-369, 1986.

[5] M. A. Fligner and J.S. Verducci. Aspects of Two Group Concordance. *Communications in Statistics: Theory and Methods*, **16**:1479-1503, 1987.

[6] M. A. Fligner and J. S. Verducci. Multi-stage ranking models. *J. Amer. Statist. Ass.* , **83**:892-901, 1988.

[7] M. G. Kendall. *Rank Correlation Methods.* Griffin, London, 1970.

[8] R. D. Luce. *Individual Choice Behavior.* Wiley, New York, 1959.

[9] C. L. Mallows. Non-null ranking models. *Biometrika*, **44**:114-130, 1957.

[10] B. Babington Smith. Discussion of Professor Ross's paper. *J. Roy. Statist. Soc. B*, **12**:53-56, 1950.

4

Latent Structure Models for Ranking Data

M. A. Croon and R. Luijkx [1]

ABSTRACT In this paper several latent structure models for analyzing data that consist of complete or incomplete rankings are discussed. First, attention is given to some latent class extensions of the Bradley-Terry-Luce model for ranking data. Next, various latent class models based on log-linear modeling of ranking data are described. Within this latter family of latent class models, a main distinction is made between models based on the assumption of quasi-independence within the latent classes, and models in which some form of association between the ranking positions is allowed to exist within the classes. All models are applied to a real data set from a large scale cross-national survey on political values.

4.1 Introduction

Latent Structure Models

Latent structure models are extensively used in the social and behavioral sciences, and their popularity in these circles is easily explained. One of the main problems with which empirical research in these sciences has to cope pertains to the imperfect and unreliable way in which theoretically important constructs are 'measured' or operationalized. Concepts such as 'intelligence' 'neuroticism', 'group cohesiveness',or 'political trust' simply elude direct measurement, and variation among respondents on such theoretical constructs can only be assessed by means of imperfect indicator variables. These indicator variables hopefully reflect variation on the underlying theoretical concept, but are probably also influenced by a host of other irrelevant disturbing factors. As a consequence, empirical investigators in the social and behavorial sciences have long been interested in methods by means of which the relation between underlying unobservable latent variables and observable manifest variables can be described and analyzed, and that is exactly what latent structure models do.

Depending upon the nature of the manifest and latent variables, many

[1] Faculty of Social Sciences, Tilburg University, Tilburg, The Netherlands

different forms of latent structure models may be formulated. By way of *factor analysis* (or by means of the related technique of *covariance structure analysis*) one may analyze the correlation or covariance structure among a large number of quantitative manifest variables in terms of a relatively small number of quantitative latent variables. *Latent trait models*, on the other hand, aim at the analysis of categorical (mostly dichotomous) responses to aptitude or attitude items in terms of underlying continuous latent traits. Finally, *latent class analysis* was developed to analyze the association between qualitative variables.

Although in social and behavioral research respondents are quite often asked to rank a given set of alternatives on a particular evaluation criterion, no special attention has yet been paid to the problem of developing latent structure models for ranking data. In this paper, we will describe several latent structure models for ranking data and illustrate the usefulness of these methods. The basic idea behind all models that we will discuss is that a heterogeneous population of respondents may be partitioned into a small number of homogeneous subpopulations, within each of which the choice or ranking processes are assumed to satisfy a relatively simple model. Seen in this way, these latent structure models are instances of *finite mixture models.*

Ranking Tasks: Some Notation and Terminology

Assume that a set of n stimuli is presented to the subjects who are instructed to select and rank the m alternatives which, in their view, score highest on the evaluation criterion defined by the investigator. Such a ranking task will be called a 'rank m out of n' task. If $m = n - 1$, we obtain complete rankings of the stimuli; for $m < n - 1$, the rankings are incomplete. In this paper we assume that ties are not allowed in the rankings. If we denote the alternatives by the first n integers, and arbitrary alternatives by either subscripted or unsubscripted symbols as i, j, k and l, the respondents' rankings can be represented by ordered m—tuples $(i_1, i_2, \cdots, i_m)$. In this m-tuple, i_1 represents the alternative that occupies the first position in the ranking, i_2 represents the alternative that occupies the second position in the ranking, etc. In general, i_r represents the alternative that occupies position r in the ranking, with $1 \leq r \leq m$.

The Data for Illustration

All the models described in this paper will be illustrated on data obtained from the cross-national survey *Political Action* (See [1]). In this survey respondents from five different western countries (West-Germany, The Netherlands, the United States, Great-Britain and Austria) were asked to select and rank their three most preferred alternatives from the following set of eight political goals:

1. Maintain a high rate of economic growth.

2. Make sure that this country has strong defense forces.
3. Give people more say in how things are decided at work and in their country.
4. Try to make our cities and countryside more beautiful.
5. Maintain a stable economy.
6. Fight against crime.
7. Move toward a friendlier, less impersonal society.
8. Move toward a society where ideas are more important than money.

In this paper only the U.S. data will be used.

The selection of these eight political goals was based on Inglehart's theory of value orientations in which a clear distinction is drawn between a materialistic and a post-materialistic value orientation (See [8]). The materialistic value orientation is characterized by a strong concern for social and economic stability, while the post-materialistic value orientation is mainly concerned with the more humane, ecological and spiritual aspects of social life. In this respect, it is clear that the political goals 1,2,5 and 6 tap the materialistic value orientation, whereas the remaining goals 3,4,7 and 8 tap the post-materialistic value orientation.

The ranking in which alternative i is in the first, alternative j in the second and alternative k in the third position will be denoted by the ordered triple (i, j, k). Its observed frequency will be denoted by f_{ijk}, and its theoretical probability by p_{ijk}.

For all models discussed in this paper specific FORTRAN computer programs were developed since none of the available standard packages for log-linear and latent class analysis seemed capable of dealing in an efficient way with the particular features shown by ranking data. As we shall see, the fact that particular patterns of structural zeros emerge if one summarizes ranking data in the form of a contingency table has to be taken into account in a log-linear and latent class analysis of ranking data. Upon request these program codes are available from the first author.

4.2 Latent Class Analyses Based on the Bradley-Terry-Luce Model

The BTL choice model

Although the Bradley-Terry-Luce model (in what follows, the BTL- model, for short) was first introduced by Bradley and Terry [4] in 1952 as a statistical model for analyzing choices between pairs of stimuli, its history seems to date back to at least 1929, when the set-theoretician Zermelo [16] arrived at basically the same model in an attempt to develop a mathematically sound way to rank chess masters on the basis of the results of round-robin tournaments. As a model for individual choice behavior, the BTL-model was thoroughly investigated by Luce [9] in his monograph *'Individual Choice Behavior'*. Luce showed how the BTL-model may be derived from an *Axiom of Choice.*

Let S denote the set of alternatives used in a choice experiment and let R be a subset of S: $R \subseteq S$. Let i be an arbitrary element of R, and hence of S. Let $p_R(i)$ and $p_S(i)$ denote the probabilities of selecting item i from either R or S, and let $P_S(R)$ represents the probability that one of the elements of R is selected when the entire set S of alternatives is presented. Then, Luce's choice axiom states that the choice probabilities satisfy the following condition:

$$p_S(i) \quad = \quad p_S(R).p_R(i)$$

Luce [9] showed that if a subject's choices satisfy this choice axiom, there exists a scale on which each alternative i has a (positive) scale value u_i such that:

$$p_R(i) \quad = \quad \frac{u_i}{\sum_{j \in R} u_j}$$

The scale values u_i are uniquely defined up to multiplication by a positive constant. In the case of a pairwise choice between alternatives i and j, we have $R = \{i, j\}$, and, hence, if p_{ij} denotes the probability of choosing i over j, we have:

$$p_{ij} \quad = \quad \frac{u_i}{u_i + u_j}$$

The BTL-model can be parametrized in another way. By defining

$$a_i = \ln \ u_i,$$

we obtain:

$$p_R(i) \quad = \quad \frac{\exp a_i}{\sum_{j \in R} \exp a_j}$$

with $-\infty < a_i < +\infty$. For pairwise choices, we have:

$$p_{ij} \quad = \quad \frac{\exp a_i}{\exp a_i + \exp a_j}$$

We will now discuss two different extensions of the BTL-model which have been proposed in the past for the analysis of rankings.

The BTL-model as a random utility model

The first adaptation starts from the well known fact that the BTL-model is compatible with a particular random utility model as defined in [2] or [10]. This point has been thoroughly investigated by Yellott [12, 13], but was already signaled by Bradley [3]. The basic assumptions of random utility models may be stated in the following way. Every time a stimulus is presented to a subject, it elicits a subjective impression of worth or value. The magnitude of this subjective impression may be represented by a real number. Instead of assuming that a stimulus always elicits the same subjective impression, one assumes that the magnitude of the subjective impression is a random variable. Let $\tilde{U}_i$ represent the random variable that corresponds to the fluctuating subjective impressions elicited by stimulus i. Then, the probability that alternative i will be chosen from set R is given by

$$p_R(i) \quad = \quad Prob(\tilde{U}_i = \max_{k \in R} \tilde{U}_k)$$

For pairwise choices we obtain

$$p_{ij} \quad = \quad Prob(\tilde{U}_i \geq \tilde{U}_j)$$

The BTL-model is compatible with the random utility model in which the random variables $\tilde{U}_i$ are independently distributed as extreme value distributions with constant scale parameters, but with possibly different location parameters. Without loss of generality, we may set the constant scale parameter equal to one, and obtain the following expression for the density function of the extreme value distribution for the random variable $\tilde{U}_i$:

$$f(u_i) \quad = \quad \exp\left[-(u_i - a_i) - \exp(u_i - a_i)\right]$$

for $-\infty < u_i < +\infty$, and in which a_i is the location parameter of the distribution.

Under this interpretation of the BTL-model, one easily derives expressions for the ranking probabilities in a ranking task. The probability $p_{i_1,i_2,\cdots,i_m}$ that in a 'm out of n' ranking task the incomplete ranking $(i_1, i_2, \cdots, i_m)$ is observed is given by:

$$p_{i_1,i_2,\cdots,i_m} \quad = \quad Prob(\tilde{U}_{i_1} \geq \tilde{U}_{i_2} \geq \cdots \geq \tilde{U}_{i_m} \geq \max_{k \notin \{i_1,i_2,\cdots,i_m\}} \tilde{U}_k)$$

Let $\mathcal{I} = \{1, 2, \cdots, m\}$ and define

$$\mathcal{J}_r \quad = \quad \mathcal{I} \backslash \{i_1, i_2, \cdots, i_{r-1}\}$$

for a given ordering $(i_1, i_2, \cdots, i_m)$. Note that $\mathcal{J}_1 = \mathcal{I}$.

If the random variables $\tilde{U}_i$ follow independent extreme value distributions, one may prove

$$p_{i_1,i_2,\cdots,i_m} = \prod_{r=1}^{m} \left(\frac{\exp a_{i_r}}{\sum_{j \in \mathcal{J}_r}^{n} \exp a_j} \right)$$

For a 'rank three out of n' task, the expressions for the ranking probabilities simplify to:

$$p_{ijk} = \frac{\exp a_i}{\exp a_i + \exp a_j + \exp a_k} \times \frac{\exp a_j}{\exp a_j + \exp a_k}$$

This expression shows that under this random utility BTL ranking model the probability of obtaining a particular ranking such as (i, j, k) is given by the product of the probability of selecting i from $\{i, j, k\}$ and the probability of selecting j from the set that remains after the first selection has been made, i.e. the probability of selecting j from $\{j, k\}$. A similar interpretation of ranking probabilities as products of successive selection probabilities also applies in the general case of a 'rank m out of n' task.

The Pendergrass-Bradley approach

Pendergrass and Bradley [11] proposed a different extension of the BTL-model to the analysis of rankings. In the case the subjects are required to rank three alternatives $\{i, j, k\}$, these authors assume that the probability of obtaining the complete ranking (i, j, k) is proportional to the product of the three paired comparison probabilities which are induced by the ranking:

$$p_{ijk} = C \cdot p_{ij} \cdot p_{ik} \cdot p_{jk}$$

The proportionality constant C is chosen so that the sum of all ranking probabilities p_{ijk} is equal to one.

If the paired comparison probabilities satisfy the BTL-model, one may derive

$$p_{ijk} = \frac{\exp(2a_i + a_j)}{\sum_{r,s \neq r} \exp(2a_r + a_s)}$$

By applying the basic principle of this approach, we obtain for the probability that the incomplete ranking $(i_1, \cdots, i_m)$ is observed in a 'rank m out of n' task the following expression:

$$p_{i_1,\cdots,i_m} = \frac{\exp\left(\sum_{r=1}^{m} (n - r) a_{i_r}\right)}{Q}$$

in which Q is the sum of terms like $\exp\left(\sum_{r=1}^{m} (n - r) a_{i_r}\right)$ over all possible incomplete rankings .

Latent class models for the analysis of ranking data based on the BTL model

For a discussion on how to obtain the maximum likelihood estimates of the scale parameters a_i under both models, and on how to test their statistical fit, we refer to [5]. Unfortunately, application of these methods to data from large surveys seldom results in an acceptable fit. The main reason for this consistent negative result probably lies in the fact that these models are unable to capture 'differences of opinion' in large populations, which are usually quite heterogeneous with respect to social and political attitudes.

In an attempt to extend the applicability of the BTL-model to the analysis of rankings in large samples from heterogeneous populations, Croon [5, 6] eveloped finite mixture models in which the BTL ranking models are coupled with the basic assumptions of latent class models. The point of departure of this approach is the assumption that the original heterogeneous population from which the respondents were sampled can be partitioned into a relatively small number of homogeneous subpopulations, the latent classes. Each respondent is assumed to belong to exactly one of these latent classes, but latent class membership is an unobserved variable. Assume that T latent classes are needed in a particular analysis and let t denote an arbitrary class. The scale values of alternative i in latent class t will be denoted by a_{it}. Let $\rho = (i_1, \cdots, i_m)$ be an arbitrary incomplete ranking. If we denote the probability of obtaining ranking ρ in latent class t by $p_{\rho,t}$, we obtain for the random utility ranking model:

$$p_{\rho,t} = \prod_{r=1}^{m} \left(\frac{\exp a_{i_r t}}{\sum_{j \in \mathcal{J}_r} \exp a_j} \right)$$

If π_t denotes the proportion of subjects belonging to latent class t, we derive

$$p_\rho = \sum_{t=1}^{T} p_{\rho,t} \cdot \pi_t$$

for the probability p_ρ of obtaining ranking ρ in a random sample from the entire population.

Similar expressions hold for the PB ranking models. For more information on these latent class models, and on the way in which the model parameters can be estimated by means of an E-M algorithm, we refer to [5].

An illustration

We give here the results of some analyses on the incomplete rankings of the eight political goals in the US sample (N=2090). These analyses were based on the random utility adaptation of the BTL-model. (We will not discuss the results of the analysis using the Pendergrass-Bradley approach, which gave very similar results.) The number of latent classes was systematically varied from 1 to 6. In the following table we give for each latent class number the log likelihood ratio statistic and the associated number of degrees of freedom. By means of the log likelihood ratio one tests the hypothesis that the model with a particular number of classes provides an acceptable description of the data against the general alternative that the set of ranking frequencies are multinomially distributed. This log likelihood ratio statistics is asymptotically distributed as a chi square distribution with the corresponding number of degrees of freedom. The general formula for computing the degrees of freedom is: $n(n-1)(n-2)-nT$, with n being the number of alternatives and T the number of latent classes.

t	*L*	*df*
1	1073.31	328
2	573.64	320
3	488.08	312
4	429.74	304
5	401.78	296
6	377.51	288

From this table we see that the value of the log likelihood statistic drastically decreases when the number of classes is increased, but, unfortunately, even the solution with six classes fails to provide a statistically acceptable fit to the data. Presumably, the latent class model based on this adaptation of the BTL model still remains a much too simple model to capture the diversity of political attitudes in the U.S. sample. Although we have certainly to reject the two-classes solution, it may be of some interest to take a closer look at it. If Inglehart's theory on value orientations is correct, one expects that one of the latent classes would represent the 'materialistic' respondents while the other would represent the 'post-materialists'. The following table gives the scale values of the eight political goals in the two classes.

i	*Class* 1	*Class* 2
1	−.05	−1.49
2	.77	−1.17
3	.00	.70
4	−1.08	−.31
5	1.33	.45
6	1.07	.15
7	−1.64	.56
8	−.40	1.12

In latent class 1, the materialistic alternatives 2, 5 and 6 score relatively high, while the post-materialistic items 4,7, and to a lesser extent also alternative 8, score low. The first latent class seems to represent the materialistic respondents. The interpretation of the second latent class as the subpopulation of post-materialistic respondents is probably also quite adequate since in this class the post-materialistic items 3, 7 and 8 score high, while the materialistic items 1, 2, and to a lesser extent alternative 6, score low. However, note that not all items conform to the expected pattern:

- In class 1 item 1 scores too low, whereas item 3 scores too high.
- In class 2 item 4 scores too low, whereas item 5 scores too high.

4.3 Latent Class Analyses Based on a Quasi-independence Model

Log-linear models for ranking probabilities.

In search for more flexible latent class models, a study of the log-linear analysis of ranking data was made. For more information on the log-linear analysis of 'rank 3 out of n' data, we refer to [7], but see also [12, 13] for similar ideas.

In the case of 'rank 3 out of n' data, the saturated log-linear model for the theoretical ranking probabilities $p_{ijk} > 0$ (with $i \neq j, i \neq k, j \neq k$) may be stated in the following way:

$$\ln p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

In this model, u is a normalizing constant; the terms u_1, u_2 and u_3 represent the main effects of the various alternatives corresponding to the first, second and third position in the ranking; the terms u_{12}, u_{13} and u_{23} represent the first-order interaction effects between the ranking positions; finally, the term u_{123} represents the second- order interaction between all ranking positions. The first- and second- order interaction terms are only defined for pairs and triples of distinct subscripts. Moreover, in order to obtain an identified log-linear model, some ANOVA-like restrictions have to be imposed on the main and interaction effects. The basic idea behind this log-linear model for 'rank 3 out of n' data is that the ranking frequencies can be inscribed in a $n \times n \times n$ contingency table whose three successive dimensions correspond to the three positions in the incomplete rankings. Since an alternative cannot occupy two or more different positions in the same ranking, only the $n(n-1)(n-2)$ cells that correspond to the possible rankings may contain a non-zero frequency. The remaining $n^3 - n(n-1)(n-2) = n(3n-2)$ cells necessarily contain structural zeros.

The quasi-independence log-linear model

The quasi-independence log-linear model is obtained by assuming that all first- and second-order interaction effects are zero. We then have

$$\ln p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

for any triple (i, j, k) of different subscripts. As identifying constraints we impose

$$\sum_{i=1}^{n} u_{1(i)} = \sum_{i=1}^{n} u_{2(i)} = \sum_{i=1}^{n} u_{3(i)} = 0$$

This model may also be written multiplicatively:

$$p_{ijk} = v \cdot v_{1(i)} \cdot v_{2(j)} \cdot v_{3(k)}$$

with, as identifying constraints,

$$\sum_{i=1}^{n} v_{1(i)} = \sum_{i=1}^{n} v_{2(i)} = \sum_{i=1}^{n} v_{3(i)} = 1$$

The concept of quasi-independence is an adaptation of the usual concept of independence to the case of contingency tables with structurally empty cells.

In the general case of a 'rank m out of n' task, we may write in terms of the multiplicative model

$$p_{i_1,\cdots,i_m} = v \cdot \prod_{r=1}^{m} v_{r(i_r)}$$

with

$$\sum_{i=1}^{n} v_{r(i)} = 1$$

for all $r = 1, \ldots, m$. The parameter v is a normalizing factor, which is needed to ensure that the sum of the ranking probabilities over all feasible rankings is equal to one.

It may be of some interest to note here that the Pendergrass-Bradley model for ranking probabilities is a submodel of the quasi-independence model introduced above. Under the Pendergrass-Bradley model, there exist scale values v_i such that we have $v_{r(i)} = v_i^{n-r}$ for all $r = 1, \cdots, m$. The random utility variant of the BTL-model for ranking data, on the other hand, is a submodel of the log-linear model in which the m-th position is independent of the configuration of the first $m - 1$ positions, i.e. of the model in which all interaction terms in which the m-th position is involved are equal to zero.

The latent class model based on the quasi-independence model

The quasi-independence model can easily be incorporated in a latent class model. Assume T latent classes are needed, and let an arbitrary class be denoted by t. The parameters v_r are assumed to be specific for each class; they will be denoted by $v_{r(i)t}$. As identifying constraints we impose for all r and all t:

$$\sum_{i=1}^{n} v_{r(i)t} = 1$$

Then, we may write for the probability of obtaining ranking $(i_1, \cdots, i_m)$ in latent class t:

$$p_{i_1,\cdots,i_m,t} = v_t \prod_{r=1}^{m} v_{r(i_r)t}$$

in which v_t is the normalizing factor for latent class t. If π_t represents the latent proportion of class t, we finally have:

$$p_{i_1,\cdots,i_m} = \sum_{t=1}^{T} p_{i_1,\cdots,i_m,t} \cdot \pi_t$$

The E.M. algorithm for estimating the quasi-independence latent class model

The maximum likelihood estimates of the model parameters can be obtained by means of an E.M-algorithm. We will restrict ourselves to the case of 'rank 3 out of n' data in our discussion of this algorithm.

The iterations of the E.M. algorithm consist of two steps:
an E(expectation)-step and a M(aximization)-step.

1. During the **E-step** the observed ranking frequencies f_{ijk} are distributed over the T classes in the following way:

   $$f_{ijkt} = f_{ijk} \times p_{t|ijk}$$

   in which the conditional probability $p_{t|ijk}$ is given by

   $$p_{t|ijk} = \frac{p_{ijkt} \cdot \pi_t}{\sum_t p_{ijkt} \cdot \pi_t}$$

   This conditional probability is computed on the basis of the provisory values of the model parameters.

2. During the **M-step**, the quasi-independence model is fitted, separately in each class, to the 'completed' set of ranking frequencies f_{ijkt}. This is done by using the Iterative Proportional Fitting Algorithm. Let e_{ijkt} denote the expected frequency corresponding to the

observed frequency f_{ijkt} under the quasi-independence model. These expected frequencies are obtained by means of the following iterative computing algorithm:

Step 1

$$e_{ijkt}^{(s)} = e_{ijkt}^{(s-1)} \times \frac{f_{i++t}}{e_{i++t}^{(s-1)}}$$

Step 2

$$e_{ijkt}^{(s+1)} = e_{ijkt}^{(s)} \times \frac{f_{+j+t}}{e_{+j+t}^{(s)}}$$

Step 3

$$e_{ijkt}^{(s+2)} = e_{ijkt}^{(s+1)} \times \frac{f_{++kt}}{e_{++kt}^{(s+1)}}$$

We use here, and also in what follows, the + subscript to denote summation over the corresponding subscript. So, for instance,

$$f_{i++t} = \sum_{j \neq i} \sum_{k \neq i,j} f_{ijkt}$$

That we have to use the Iterative Proportional Fitting Algorithm in fitting the quasi-independence model is due to the fact that this model does not allow for an analytic solution of the maximum likelihood optimization problem.

During each M-step, the latent proportions are also estimated again:

$$\pi_t = \frac{e_{+++t}}{N}$$

An example

The following table contains the global results of some latent class analyses based on the quasi- independence model. We have used once again the U.S. data.

t	L	df	p
1	920.36	314	0
2	385.70	292	0.0002
3	318.43	270	0.0228
4	269.86	248	0.1625

From a statistical point of view, only the solution with four classes is acceptable; the solutions with a smaller number of classes all result in a statistically unacceptable fit. In order to see in which respects these four classes differ among themselves, we report the first-choice parameters $v_{1(i)t}$ in the following table:

i	*Class* 1	*Class* 2	*Class* 3	*Class* 4
1	.113	.012	.007	.025
2	.131	.175	.023	.038
3	.063	.023	.005	.230
4	.009	.000	.005	.024
5	.404	.620	.068	.158
6	.241	.141	.878	.000
7	.008	.000	.006	.127
8	.030	.029	.009	.399
π_t	.318	.229	.163	.289

The second- and third choice parameters $v_{2(i)t}$ and $v_{3(i)t}$ showed a pattern similar to that of the first-choice parameters. These results indicate that under the quasi-independence model three slightly different 'materialistic' classes seem to exists in the U.S.A. The first three classes are all characterized by a strong preference of some 'materialistic' items, and by a resolute rejection of the 'post-materialistic' political goals. The differences between the three 'materialistic' classes are more difficult to interpret, and seem to be rather item-specific. Seventy-one percent of the American sample is estimated to belong to one of the materialistic classes. The fourth class probably represents the 'post-materialistic' subpopulation, although some of the alternatives do not conform to the pattern that could be expected here: In this class the alternative item 4, which is a very unpopular item in the U.S., scores too low, while the materialistic item 5, which is the most popular item in this sample, scores too high.

4.4 Models that Allow for Association Between Choices within the Classes

A General Model Allowing for Association between Choices

Latent class models based on a quasi-independence model do not always lead to satisfactory results. Often models of this kind only provide a statistically acceptable fit to the data if the number of latent classes is made large enough.

In a search for alternative latent class models, which possibly could explain the data in terms of a smaller number of latent classes, we first considered the log-linear model which includes all first-order, but not the second- and higher-order interaction effects. This first-order interaction model is in some sense the most simple extension of the quasi-independent model. In this section we restrict ourselves to a discussion of ranking data from a 'rank 3 out of n' task.

For the case of 'rank 3 out of n' data, the latent class model with first-order interactions can be written as

$$\ln p_{ijkt} = u_t + u_{1(i)t} + u_{2(j)t} + u_{3(k)t} + u_{12(ij)t} + u_{13(ik)t} + u_{23(jk)t}$$

In this model, which we refer to as the A_0-model, latent classes differ with respect to the main effects as well as with respect to the first-order interaction terms. It is interesting to note that, for $T = 1$, we simply obtain the hierarchical submodel of the saturated log-linear from which all second-order interaction terms are removed. Our limited experiences with this very general A_0-model, however, have been quite negative for $T \geq 2$.

We observed quite often that the final solutions under this model had many of their parameter estimates on the boundary of the parameter space. This was especially the case for the estimates of the first-order interaction terms. Some rather difficult identification problems are probably involved here.

Three submodels with Invariant First-Order Interaction Effects

Since the general A_0-model did not provide an acceptable alternative to the quasi-independence model considered earlier, we have investigated some submodels of it. In particular, we have considered models in which the first-order interaction terms are assumed to be the same in the various latent classes, which may still differ with respect to main effects. In these models the latent classes may differ with respect to the 'popularity' of the items, but the pattern of association between the choices (as described by first-order interaction terms) is assumed to be invariant over the different classes. We first consider the most general model of this kind, the A_1-model before discussing two interesting submodels of it.

Model A_1

For the most general model within this class, we may write for 'rank 3 out of n' data:

$$\ln p_{ijkt} = u_t + u_{1(i)t} + u_{2(j)t} + u_{3(k)t} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

In the following this model will be referred to as Model A_1. Note that for $T = 1$ this model too is equivalent to the 'no second-order interactions' submodel of the saturated log-linear model, and, hence, to Model A_0 with $T = 1$ as described above.

Next, we consider two submodels of A_1.

Model A_2

As a first interesting submodel of A_1 we will consider the model for which

$$u_{12(ij)} = u_{13(ij)} = u_{23(ij)} = u_{ij}$$

holds for all i, j. Under this model, which will be referred to as model A_2, one may write

$$\ln\ p_{ijkt} \quad = \quad u_t + u_{1(i)t} + u_{2(j)t} + u_{3(k)t} + u_{ij} + u_{ik} + u_{jk}$$

In this model only one set of invariant first-order interaction terms remains to be estimated.

Model A_3

A second interesting submodel of A_1 is the model in which the First by Second Choice, and the Second by Third Choice first-order interaction terms are included, but not the First by Third Choice interaction terms. Hence, for this model A_3, we may write:

$$\ln\ p_{ijkt} \quad = \quad u_t + u_{1(i)t} + u_{2(j)t} + u_{3(k)t} + u_{12(ij)} + u_{23(jk)}$$

In this model only interaction terms for pairs of consecutive positions in the ranking are included. Note that models A_2 and A_3 are both submodels of model A_1, but are themselves not hierarchically related to each other.

Estimating the Parameters by Means of an E.M. Algorithm. Let f_{ijk} be the observed frequency of ranking (i, j, k) and assume that T latent classes are needed for an analysis based either on model A_1, model A_2, or model A_3. Let N denote the sample size. The maximum likelihood estimates of the parameters of the three models can be obtained by means of an E.M. algorithm.

Each iteration of this algorthim consists of two steps:

- An **Expectation step** during which the frequencies f_{ijkt} with which ranking (ijk) occurs in latent class t is estimated again:

$$f_{ijkt} \quad = \quad f_{ijk} \times \frac{p_{ijkt} \cdot \pi_t}{p_{ijk}}$$

  with

$$p_{ijk} \quad = \quad \sum_t p_{ijkt} \cdot \pi_t$$

The probability p_{ijkt} of observing ranking (ijk) in latent class t is computed on the basis of the provisory values of the parameter estimates. The way in which these probabilities are computed depends on the model under consideration.

- A **Maximization Step** during which the maximum likelihood estimates of the model are determined again on the basis of the completed set of frequencies f_{ijkt}. The expression for the latent proportions π_t is extremely simple:

$$\pi_t = \frac{f_{+++t}}{N}$$

The estimation of the parameters of the log-linear model is more involved, since one has to rely on a subordinate iterative process, such as the Iterative Proportional Fitting Algorithm. More information on these estimation procedures are given in the next paragraph.

The Iterative Proportional Fitting Algorithm for Models A_1, A_2 and A_3 with Complete Data. We assume that the frequency f_{ijkt} with which ranking (ijk) occurs in class t is observed. The corresponding expected frequency will be denoted by e_{ijkt}.

For model A_1 the iterations of Iterative Proportional Fitting Algorithm consist of the following 6 steps:

1.

$$e_{ijkt}^{(s+1)} = e_{ijkt}^{(s)} \times \frac{f_{i++t}}{e_{i++t}^{(s)}}$$

2.

$$e_{ijkt}^{(s+2)} = e_{ijkt}^{(s+1)} \times \frac{f_{+j+t}}{e_{+j+t}^{(s+1)}}$$

3.

$$e_{ijkt}^{(s+3)} = e_{ijkt}^{(s+2)} \times \frac{f_{++kt}}{e_{++kt}^{(s+2)}}$$

4.

$$e_{ijkt}^{(s+4)} = e_{ijkt}^{(s+3)} \times \frac{f_{ij++}}{e_{ij++}^{(s+3)}}$$

5.

$$e_{ijkt}^{(s+5)} = e_{ijkt}^{(s+4)} \times \frac{f_{+jk+}}{e_{+jk+}^{(s+4)}}$$

6.

$$e_{ijkt}^{(s+6)} = e_{ijkt}^{(s+5)} \times \frac{f_{i+k+}}{e_{i+k+}^{(s+5)}}$$

For Model A_2, the iterations of the Iterative Proportional Fitting Algorithm consist of 4 steps, the first three being identical with the corresponding steps of the algorithm for the A_1 model. The fourth step itself consists of $n(n-1)$ substeps, each one corresponding to a pair (i, j) of distinct subscripts. During the substep that corresponds to the pair (i, j), the following computations take place for all $k = 1, \cdots, n$ (with $k \neq i$ and $k \neq j$) and for all t:

$$\begin{aligned} e_{ijkt}^{(new)} &= e_{ijkt}^{(old)} \times \frac{S_{ij}}{U_{ij}^{(old)}} \\ e_{ikjt}^{(new)} &= e_{ikjt}^{(old)} \times \frac{S_{ij}}{U_{ij}^{(old)}} \\ e_{kijt}^{(new)} &= e_{kijt}^{(old)} \times \frac{S_{ij}}{U_{ij}^{(old)}} \end{aligned}$$

with

$$S_{ij} = f_{ij++} + f_{i+j+} + f_{+ij+}$$

and

$$U_{ij}^{old} = e_{ij++}^{(old)} + e_{i+j+}^{(old)} + e_{+ij+}^{(old)}$$

For Model A_3, the iterations of the Iterative Proportional Fitting Algorithm consist of five steps, which are identical to the first five steps of the algorithm for fitting Model A_1.

In order to guarantee that the Iterative Proportional Fitting Algorithms converge to the maximum of the likelihood function, the starting values of the expected frequencies should satisfy the model under consideration. The easiest way out of this problem is to set all expected frequencies e_{ijkt} initially equal to 1.

After convergence of the Iterative Proportional Fitting Algorithm, the model parameters, such as $u_{1(i)t}, u_{2(i)t}, u_{3(i)t}, u_{12(ij)}, \cdots$, can be determined by solving appropriate systems of linear equations in these unknowns. This system of linear equations expresses the model parameters as functions of the natural logarithms of the expected frequencies e_{ijkt}.

Testing model fit. When the E.M. algorithm has converged, the hypothesis that the model under consideration applies to the data may be tested against the general multinomial hypothesis by means of a log likelihood ratio test. Let $\hat{p}_{ijk}$ be the estimate of the theoretical ranking probability under the particular model under consideration, and let $\hat{f}_{ijk} = N \cdot \hat{p}_{ijk}$ denote the corresponding expected frequency. Then, the log likelihood ratio statistic L is defined as:

$$L = 2 \times \sum_{i,j,k} f_{ijk} \ln \left(\frac{f_{ijk}}{\hat{f}_{ijk}} \right)$$

where the summation runs over all triples of distinct subscripts.

If the model under consideration is true, then the log likelihood statistic is asymptotically distributed as a chi square variate with degrees of freedom equal to the difference between the number of independent parameters under both models.

In the context of latent class analysis, model tests of this kind can be used to the test the hypothesis that the latent class model with a specified number T of classes is true against the general multinomial hypothesis. Let L_T denote the value of log likelihood statistic obtained by a latent class analysis with T classes. For Model A_1 the observed value of the statistic L_T should, for $n \geq 5$, be located under a chi square distribution with $(n^3 - 6n^2 + 11n - 3) - (3n - 2)T$ degrees of freedom; for Model A_2, the number of the degrees of freedom is given by $(n^3 - 4n^2 + 5n - 1) - (3n - 2)T$ if $n \geq 5$; for Model A_3, the number of degrees is $n^3 - 5n^2 + 8n - (3n - 2)T$.

Some Results

The Results of the A_2 Analyses on the U.S. Data. The U.S. ranking data were analyzed on the basis of model A_2 with $T = 1$ and $T = 2$. The global results are shown in the next table.

T	L	$d.f.$	p
1	325.523	273	.016
2	248.833	251	.527

Hence, we see that the solution with two latent classes provides an acceptable fit to the U.S. data. The next table contains the estimates of the main effects parameters in both classes.

	Class 1			*Class* 2		
i	$u_{1(i)1}$	$u_{2(i)1}$	$u_{3(i)1}$	$u_{1(i)2}$	$u_{2(i)2}$	$u_{3(i)2}$
1	.11	.29	.05	−1.25	−2.01	−1.33
2	.90	.69	−.09	−.60	−.42	−.10
3	−1.29	−.91	−.12	1.48	1.51	1.14
4	−1.95	−1.67	−1.43	−1.29	−.18	.30
5	2.23	1.49	.68	.59	.58	.33
6	2.37	2.57	2.20	−.09	−.73	−.78
7	−1.59	−1.15	−.48	−.21	.01	−.68
8	−.79	−1.31	−.82	1.38	1.24	1.13
$\bar{u}_{mat}$	1.40	1.26	.71	−.34	−.65	−.47
$\bar{u}_{pmat}$	−1.40	−1.26	−.71	.34	.65	.47

The estimates of the latent proportions were $\hat{\pi}_1 = .613$ and $\hat{\pi}_2 = .387$. The interpretation of these results is rather straightforward:

- The first latent class is a relatively pure 'materialistic' class in which the four materialistic alternatives are rated higher than the four post-materialistic ones. The clear opposition between the two groups of alternatives occurs at all three ranking positions, but it diminishes slightly when going from the first to the third position.

- When looking at the average scale values of the materialistic and post-materialistic alternatives in the second class, it should be clear that this class cannot be considered as a pure 'post-materialistic' class. A few rather striking exceptions make such an interpretation implausible: In this class, the post-materialistic items 4 and 7 score much too low, while the materialistic alternative 5 scores too high. It is probably safer to characterize this class as the class of persons who value the humane and spiritual aspects of life.

A Further Analysis of the First-Order Interaction Terms. Next, we turn to the interpretation of the interaction terms. Instead of giving the complete 8 × 8 matrix with estimated first-order interaction terms, we will report on the results of a bilinear decomposition analysis of these terms. Assume the first-order interaction terms u_{ij} are inscribed in a $n \times n$ matrix U. Since the terms u_{ij} are undefined for the case $i = j$, the main diagonal of this matrix is structurally empty. We say that the matrix U allows for a 'Bilinear Decomposition of Rank s' if there exist two $n \times s$ matrices X, the left factor matrix, and Y, the right factor matrix, such that

$$u_{ij} = \sum_{q=1}^{s} x_{iq} y_{jq}$$

holds for all $i, j = 1, \cdots, n$ with $j \neq i$. In practice, we are interested in the bilinear decomposition of the lowest rank which still provides an acceptable fit to the incomplete matrix. To this end, we determine, for successive values of s, the decomposition of U which minimizes the following least squares loss function:

$$\phi = \sum_{i,j \neq i} \left(u_{ij} - \sum_{q=1}^{s} x_{iq} y_{jq} \right)^2$$

For more information on the bilinear decomposition model and on the technical details of the estimation procedure, we refer to [7].

In the present example, the rank 1 decomposition left 50.3 % percent of the variance of the interaction terms unexplained. For the rank 2 decomposition, this figure decreased to 23.2 %. The next table gives the result of the latter decomposition.

i	x_{i1}	x_{i2}	y_{i1}	y_{i2}
1	−.60	−.32	.23	−.56
2	−.34	−.43	−.50	−.53
3	.67	−.66	−.47	.07
4	.39	.18	−.56	.67
5	−.23	−.19	.07	−.42
6	−.65	.46	.65	−.03
7	.27	.88	.87	.57
8	.73	.07	−.37	.48

From the information in these coordinate matrices, one may conclude that, to a large extent, the pattern of the first-order interaction terms is dominated or determined by the contrast between the two types of alternatives. An interesting feature of this bilinear decomposition is that the contrast between materialistic and post-materialistic alternatives shows itself most distinctively in the first component of the left factor matrix X, and in the second component of the right factor matrix Y. It is not clear why different components from the left and right factor matrix should be involved in this way.

A Comparison with the A_3 analyses. The U.S. data were also analyzed by means of the A_3-model. The next table gives some global results.

T	L	*df*	*df*
1	375.673	230	0
2	242.987	208	.0485
3	191.608	186	.3736

Since models A_2 and A_3 are not related to each other in a hierarchical way, it is difficult to compare the relative fits of both models to the same data. However, it is probably safe to conclude that the two-class solution of the A_2 analysis represents the data better than the two-class solution of the A_3 analysis. This is remarkable since fewer parameters are estimated under the A_2-model than under the A_3. This result seems to indicate that all three kinds of first-order interaction terms (First by Second Choice, Second by Third Choice, and First by Third Choice) are needed in a comprehensive latent class model of this type. Removing one set of these interaction terms has more detrimental effects than setting corresponding terms in the three sets equal to each other.

Acknowledgements:

The second author's contribution was made possible by grant #500-278-009 from the Social Sciences Section of the Dutch National Science Foundation (NWO).

4.5 References

[1] Barnes, S.H. and M. Kaase. *Political Action. Mass Participations in Five Western Countries.* London: Sage. 1979.

[2] Block, H.D. and J. Marschak. Random Orderings and Stochastic Theories of Response. In I. Olkin, S.G. Ghurye, W. Hoeffding, W. Madow and H. Mann (Eds.), *Contributions to Probability and Statistics* pp. 97-132, 1960. Stanford: Stanford University Press.

[3] Bradley, R.A. Another Interpretation of a Model for Paired Comparisons. *Psychometrika*, **30**:315-318, 1963.

[4] Bradley, R.A. and M.E. Terry. Rank Analysis of Incomplete Block Designs I. *Biometrika*, **39**:324-345, 1952.

[5] Croon, M.A. Latent Class Models for the Analysis of Rankings.In: G. de Soete, H. Feger and K.C. Klauer (Eds.), *New Developments in Psychological Choice Modelling* pp. 99-121, 1989a. Amsterdam: Elsevier Science Publishers.

[6] Croon, M.A. The Analysis of Partial Rankings by Log-Linear and Latent Class Models. In: R. Coppi and S. Bolasco (Eds.), *Multiway Data Analysis* pp. 497-506, 1989b. Amsterdam: Elsevier Science Publishers.

[7] Croon, M.A. Log-Linear Analysis of Partial Preference Rankings. *Methodika* 1991. (still to appear).

[8] Inglehart, R. *The Silent Revolution.* Princeton: Princeton University Press. 1977.

[9] Luce, R.D. *Individual Choice Behavior.* New York: Wiley. 1959.

[10] Luce, R.D. and P. Suppes. Preference, Utility and Subjective Probability. In R.D. Luce, R.R. Bush and E. Galanter (Eds.), *Handbook of Mathematical Psychology*, **3**:97-132, 1965. New York: Wiley.

[11] Pendergrass, R.N. and R.A. Bradley. Ranking in Triple Comparisons. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. Madow and H. Mann (Eds.), *Contributions to Probability and Statistics* pp. 133-151, 1960. Stanford: Stanford University Press.

[12] Plackett, R.L. Random Permutations. *Journal of the Royal Statistical Society, Series B*, **30**:517-534, 1968.

[13] Plackett, R.L. The Analysis of Permutations. *Applied Statistics*, **24**:193-202, 1975.

[14] Yellott, J.I. The relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment and the Double Exponential Distribution. *Journal of Mathematical Psychology*, **15**:109-144, 1977.

[15] Yellott, J.I. Generalized Thurstone Models for Ranking: Equivalence and Reversibility. *Journal of Mathematical Psychology*, **22**:48-69, 1980.

[16] Zermelo, E. Die Berechnung der Turnierergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, **29**:436-460, 1929.

5

Modelling and Analysing Paired Ranking Data

Paul D. Feigin[1]

ABSTRACT Two models for paired rankings are presented. They describe two different ways in which a *post-ranking* is related to its *pre-ranking* for each case (or subject). These models are compared to regression models in order to help motivate their forms. Analysis of paired ranking data is considered in the light of testing for that type of departure from a null model that corresponds to either of the proposed models. The procedure suggested uses a *bootstrap* method to ascertain the strength of the departure from the null model, and helps one to decide which departure is more strongly indicated. Some analyses of simulated test data sets are described, as well as the analysis of data due to Rogers [7] which is also analysed in Critchlow and Verducci [3].

Key Words and Phrases: Rankings, permutation metrics, bootstrap, hypothesis testing.

5.1 Introduction

When analysing ranking data which results from a particular experimental design, one approach is to try to mimic the corresponding analysis for ordinary Euclidean data. In order to do so, one seeks appropriate mappings from $\Omega = \Omega_k$, the set of permutations of k letters, into a Euclidean space $\mathcal{R}^s$, and then does the relevant MANOVA analysis in this range space. This approach is described in the paper of Feigin and Alvo [6], and is also applied to the case of paired rankings in Feigin [5]

In the special case of a matched pairs design, an alternative approach is to consider paired ranking models that describe how a *post-ranking* (ν) may be generated given the *pre-ranking* (π). Such models are often based on *distances* (see Critchlow, Fligner and Verducci [2] for a discussion of models based on distances). We will discuss two such paired ranking models in the sequel.

In another paper, Critchlow and Verducci [3] suggest how to test whether the post-ranking has been drawn towards a pre-determined *idealised* rank-

[1] TECHNION — Israel Institute of Technology, Haifa, Israel

ing (say, λ). One of the models we consider deals with this type of departure from a null model. A method is suggested for both estimating this ranking λ as well as testing if its effect is significant.

On applying the methods to the same data set as used by Critchlow and Verducci [3], the results indicate that an alternative departure from the null model may provide a better explanation of the data. This alternative departure corresponds to the second type of model presented in this paper. The different alternative models correspond to different mechanisms for the systematic deviation of post-rankings from pre-rankings.

The models are presented in Section 5.2, and an estimation procedure for each is discussed in Section 5.3, as well as a bootstrap method for evaluating the strength (significance) of departure from the null model. In Section 5.4 we show the results of applying the procedures to some simulated test data sets in which data were generated according to the models of Section 5.2. In Section 5.5 we apply the procedure and discuss the meaning of the results for the Rogers [7] data set reported in Critchlow and Verducci [3].

5.2 Two Models

Model I

Consider the problem of ranking four colours in the order they would be preferred for a particular object, say a car. The choice may be made based on looking at colour cards in the office (call this the π ranking), or by looking at four appropriately painted cars in the showroom (call this the ν ranking). A person may choose a different ranking under the two conditions because the environment (lighting, background) may actually change the way the four colours are ordered on the dimensions that influence the choice of colours.

In terms of permutations, the environment is actually changing the labels of the four colours. For example, in the office's yellow light, the yellow colour looks very bright; while in the showroom's natural light, the red looks very bright and the yellow looks duller. If brightness was an issue, then the yellow and red colours would change roles. A judge who preferred bright colours would rank red high in the post situation, whereas someone who preferred duller colours would choose yellow in this situation. The opposite would hold in the pre situation.

In mathematical terms, if there were no random variation,

$$\nu = \pi \circ \eta \quad ; \tag{1}$$

where the notation $\pi \circ \eta$ denotes (permutation) group multiplication and η denotes the permutation which changes the labels.

In other contexts, we can often think of the post-ranking as being influenced by new (or other) "light" being shed on the items being ranked, as compared to the pre-ranking situation. This "light" could correspond to the course that the students studied between the pre- and post-rankings, for the example to be discussed in Section 5.5.

In order to formulate the model, we need a distance $d(\mu, \omega)$ between two members μ and ω of Ω. Such distances are discussed in Diaconis [4] and also in Critchlow, Fligner and Verducci [2] and Critchlow [1]. We do not need to specify which distance is to be used at this stage.

The model for the post-ranking ν given the pre-ranking π is

$$\mathcal{P}(\nu|\pi) = C(\theta)\exp\{-\theta d(\nu, \pi \circ \eta)\} \quad ; \tag{2}$$

where η is some fixed (*label-changing*) permutation, $\theta \geq 0$, and $C(\theta)$ is a normalising constant.

The *null model* corresponds to the case when $\eta = e$, where e denotes the identity permutation. In this situation the post-ranking is centered about the pre-ranking.

In the corresponding situation of Euclidean data (x, y), the model corresponds to saying that y is centered about $x + a$, the null model pertaining when $a = 0$. This latter case corresponds to the situation in which the ordinary paired t-test is used — where one uses the normal model for the difference $y - x$. Explicitly,

$$y = x + a + \varepsilon \quad ; \tag{3}$$

where ε is a N(0,σ^2) random error; or

$$f(y|x) = K(\sigma)\exp\{-\Delta(y, x + a)/(2\sigma^2)\} \quad ; \tag{4}$$

where in this case $\Delta(s, t) = (s - t)^2$. We will call this the *shift* model.

In the next section we will pursue this analogy further when we look at the estimation of η and the testing of whether $\eta = e$.

Model II

Another approach to paired rankings treats the items as having fixed properties, but the judges who do the rankings have had their preference functions changed between the pre and post situations. In this context, it is natural to consider the post ranking ν as having moved closer to some *idealised* or given ranking λ, and away from the pre-ranking π.

This situation can be modelled in many different ways, but it is easiest to do so if we again use distance functions. The model we consider has the form:

$$\mathcal{P}(\nu|\pi) = C(\theta, \phi; \lambda, \pi)\exp\{-\theta d(\nu, \pi) - \phi d(\nu, \lambda)\} \quad ; \tag{5}$$

where $\theta, \phi \geq 0$. The value of ϕ gives the strength of the attraction of ν to λ; whereas the value of θ gives the strength of the attraction of ν to the pre- ranking π. It is both their absolute and relative values that seem to be important in being able to detect deviations from the null model — see Section 5.4.

The null model corresponds to the case when $\phi = 0$, whereupon we return to the *same* null model described for Model I (2). It is in this sense that the two models correspond to different deviations from the same null model. In the data analysis that we discuss in the sequel we will try to discern which departure is appropriate for a given data set.

In ordinary Euclidean space a corresponding model to Model II would have y centered about some convex combination of a value b and the given x. We might use the regression formulation:

$$y = \alpha b + (1-\alpha)x + \varepsilon \quad ; \tag{6}$$

where $0 \leq \alpha < 1$ and ε is some random error with zero mean.

Another way of writing this equation is

$$(y-b) = (1-\alpha)(x-b) + \varepsilon \quad ; \tag{7}$$

so that if $\alpha > 0$ then, *on average*, y should be closer to b than x is. In other words $\Delta(y,b) < \Delta(x,b)$ on average. This fact, when applied to the distance d on $\Omega \times \Omega$, forms the basis of our estimation and testing procedure.

In order to compute the constants $C(\theta,\phi;\lambda,\pi)$ for given θ and ϕ it appears that we have to do the computation for each pair of permutations (λ and π) separately. However, this is often not necessary, as we explain below.

We define the notion of *right invariance* of a distance d as follows:

$$d(\mu,\omega) = d(\mu\circ\eta, \omega\circ\eta) \text{ for all } \mu,\omega,\eta\in\Omega \quad . \tag{8}$$

Lemma 5.2.1 *If d is right invariant then the constant $C(\theta,\phi;\lambda,\pi)$ of (5) can be written $C(\theta,\phi;\lambda\circ\pi^{-1})$.*

Proof: The proof follows straightforwardly from the right invariance relation:

$$\begin{aligned} C(\theta,\phi;\lambda,\pi)^{-1} &= \sum_{\nu\in\Omega} \exp\{-\theta d(\nu,\pi) - \phi d(\nu,\lambda)\} & (9)\\ &= \sum_{\nu\in\Omega} \exp\{-\theta d(\nu\circ\pi^{-1},e) - \phi d(\nu\circ\pi^{-1},\lambda\circ\pi^{-1})\} & (10)\\ &= \sum_{\eta\in\Omega} \exp\{-\theta d(\eta,e) - \phi d(\eta,\lambda\circ\pi^{-1})\} & (11) \end{aligned}$$

We note that the Kendall distance, as well as many others used in analysis of ranking data, is right invariant — see Critchlow [1] for example.

Another property that would reduce the difficulty in computing C depends on the geometry of Ω induced by d. Take permutations π,λ fixed. Define the set of permutations

$$S(d_1,d_2;\pi,\lambda) = \{\omega : d(\omega,\pi) = d_1; d(\omega,\lambda) = d_2\} \quad . \tag{12}$$

Call a distance measure *doubly balanced* if the cardinalities of the sets S satisfy:

$$\sharp S(d_1,d_2;\pi,\lambda)=\sharp S(d_1,d_2;\pi',\lambda') \text{ whenever } d(\pi,\lambda)=d(\pi',\lambda') \quad . \qquad (13)$$

From the definition of Model II (5), it is clear that the following is true.

Lemma 5.2.2 *If the distance d is* doubly balanced *then the constant C of the definition (5) can be written:*

$$C(\theta,\phi;\lambda,\pi)=C(\theta,\phi;d(\lambda,\pi)) \qquad (14)$$

We can to use these lemmas to help simulate pairs of permutations according to the non-null model of (5).

The question remains as to which distances $d(\cdot,\cdot)$ are doubly balanced. From the "symmetry" of the symmetric group and its graphic representation as a polytope in $\mathcal{R}^{k-1}$, this property would depend on the relationship between the distance measure and the structure of the polytope. Despite a first impression to the contrary, it seems that this property will be quite rare.

By looking at the polytope diagram of Thompson [8], we see that for $k=4$ some of the faces are hexagonal and some are square. If we use the Kendall distance measure, then we are computing the minimum number of edges traversed moving between two orderings. Choosing two orderings (π^{-1} and λ^{-1}) that are at Kendall distance $d(\pi,\lambda)=2$, the size of the set $S(1,1;\pi,\lambda)$ would depend on whether π^{-1} and λ^{-1} are at opposite corners of a square face, or at distance two on a hexagonal face! In other words, the Kendall distance is *not* doubly balanced.

Distinguishing Between the Models

One of the natural questions is what are the chances of distinguishing between the (non-null) models. The answer depends on what π's we are given. Again the regression analogy will help explain. Consider the model (6), rewritten slightly differently as:

$$y=x+\alpha(b-x)+\varepsilon \quad ; \qquad (15)$$

and compare it to the *shift* equation (3). We notice that if the x's do not vary much, we have little chance of deciding between the two alternatives.

In terms of the paired rankings situation: if the pre-rankings (π's) are concentrated in one part of the space Ω, then it will be difficult to distinguish between the two models. In this case, the relabelling η will have a similar effect to that of attracting the ν's towards some λ in the vicinity of the $\pi\circ\eta$'s !

5.3 Estimation and Hypothesis Testing

In our current analysis of the two models we will treat the θ parameter as a nuisance parameter. Our goal is to estimate the η or λ permutations, and to test whether the data support a departure from the null model (that is, if $\eta = e$ or if $\phi = 0$ for the two models I and II, respectively). In the parallel Euclidean models, it is as if we are treating σ^2 as a nuisance parameter and testing whether $a = 0$ in the shift model, or whether $\alpha = 0$ in the regression model (6).

The approach will also be to treat the π's as given, or fixed, as in the case of the usual regression analysis with fixed x's.

In the following, we assume that we are given n paired rankings in the form $\{(\pi_i, \nu_i); i = 1, \ldots, n\}$. This information may also be given by presenting the pairs of inverse (or *bracket* form) rankings : $\{(\pi_i^{-1}, \nu_i^{-1}); i = 1, \ldots, n\}$. Here $\pi^{-1} = <\pi^{-1}(1), \ldots, \pi^{-1}(k)>$ for any ranking π, and $\pi^{-1}(r)$ denotes the item that received rank r (out of the k ranks) under the ranking π.

Model I

From the (conditional) likelihood for Model I, we see that the maximum likelihood estimate $\hat{\eta}$ of the permutation η satisfies

$$\sum_{i=1}^{n} d(\nu_i, \pi_i \circ \hat{\eta}) = \min_{\eta \in \Omega} \sum_{i=1}^{n} d(\nu_i, \pi_i \circ \eta) \quad . \tag{16}$$

The estimate $\hat{\eta}$ is related to the notion of *ρ-median* for the empirical distribution of the paired rankings; see Diaconis [4] (p. 108).

The value of the minimum sum in (16) should have some information on how significantly the true η differs from e. However, the expected value of this sum depends on the nuisance parameter θ in the model.

We proceed by pursuing the paired t-test paradigm a little further. Writing the differences as $\{v_i = y_i - x_i\}$, the test statistic for testing for zero difference can be written in its F-statistic form as:

$$\begin{aligned} W &= \frac{n\bar{v}^2}{s_v^2} && (17) \\ &= \frac{n(n-1)\bar{v}^2}{\sum \Delta(v_i, \bar{v})} && (18) \\ &= (n-1)\left[\frac{\sum \Delta(v_i, 0)}{\sum \Delta(v_i, \bar{v})} - 1\right] && (19) \end{aligned}$$

where, as earlier, we define the distance $\Delta(s,t) = (s-t)^2$. We note that $\bar{v}$ is the minimiser (over a) of $\sum \Delta(v_i, a)$. The null hypothesis of no shift in the paired t-test is rejected when W is large; and the critical value is determined by the theoretic $F_{1,n-1}$ distribution, which is approximately the same as the χ_1^2 distribution for n large.

Based on the expression (19) we propose using the following test statistic for testing whether η is significantly different from e. Compute

$$T(e) = \frac{\sum d(\nu_i, \pi_i \circ \hat{\eta})}{\sum d(\nu_i, \pi_i \circ e)} \tag{20}$$

where we have explicitly put the (identity permutation) argument e in the denominator although it does not need to appear in this particular expression. Small values of $T(e)$ indicate large deviations from the null hypothesis.

The actual significance level is difficult to ascertain. One ad-hoc approach is to compare $n(1/T(e) - 1)$ with the χ_1^2 distribution; pursuing the analogy with the scalar paired t-test.

Another approach is based on *bootstrapping* the statistic $T(e)$. Intuitively, we are seeking to determine how *stable* $\hat{\eta}$ is under resampling the original data set of paired rankings; and how the deviation of $\hat{\eta}$ from e (as measured by $T(e)$ compared to 1) compares to that of the $\hat{\eta}^*$ from their theoretical value $\hat{\eta}$ according to the bootstrap distribution. If $\hat{\eta}$ was just a chance minimiser, then under resampling other minimisers will appear, and the ratio $T(e)$ should be like any of the resampled or bootstrapped $T^*(\hat{\eta})$'s . A bootstrap *p-value* will tell us how extreme $T(e)$ really is. In the situation of ranking data, because of the discrete nature of the space Ω, values of $T(e) = 1$ can of course occur. This value will correspond to $W = 0$ in the context of equation (19); that is, a zero deviation from the null model *in the direction of Model I deviations.*

Formally, we generate bootstrap samples $\{(\pi_i^*, \nu_i^*); i = 1, \ldots, n\}$ and compute

$$T^*(\hat{\eta}) = \frac{\sum d(\nu_i^*, \pi_i^* \circ \hat{\eta}^*)}{\sum d(\nu_i^*, \pi_i^* \circ \hat{\eta})} \tag{21}$$

where $\hat{\eta}^*$ denotes the maximum likelihood estimate of η for the bootstrap sample. If we produce B bootstrap samples then we will have statistics $\{T_b^*(\hat{\eta}); b = 1, \ldots, B\}$ and we define the bootstrap *p-value* by

$$p^* = [\sharp\{b : T(e) \geq T_b^*(\hat{\eta})\}]/B \quad . \tag{22}$$

Model II

Estimating λ of Model II by maximum likelihood is more problematic and so the following alternative heuristic is employed.

Under the null model, for any λ, the value of $d(\nu, \lambda)$ should be approximately equal to $d(\pi, \lambda)$. In other words, on average, ν should be no closer or further from λ than π is. Nothing can be said if $\pi = \lambda$. As Critchlow and Verducci [3] point out in their work, the way the distance $d(\nu, \lambda)$ relates to $d(\pi, \lambda)$ does also depend on where π is situated with respect to λ — our assumption that the ratio should approximate one in the null case will really only be valid if the the collection of π_i's are fairly well spread out. For the time being we treat this limitation as one treats the limitation of regression methods when the spread of the independent variable is small in the sample. We propose the following procedure.

Define

$$R(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{d(\nu_i, \lambda)}{d(\pi_i, \lambda)} \mathrm{I}(\pi \neq \lambda) \quad ; \tag{23}$$

and suppose that $\hat{\lambda}$ is the minimiser of $R(\lambda)$. If $R(\hat{\lambda})$ is much smaller than 1, then it would indicate that there is a departure from the null model in the direction of $\hat{\lambda}$ in the sense of Model II. The test statistic we propose is of the form

$$U(q) = \frac{R(\hat{\lambda})}{q} \quad . \tag{24}$$

where we would naturally like to substitute an expected value of $R(\lambda)$ (under the null hypothesis) for q. The problem, as Critchlow (1990, private communication) has pointed out, is that this expected value depends on the nuisance parameter θ in the null model. We are currently investigating a method based on estimating θ first, and estimating the expected value of $R(\lambda)$ subsequently. Meanwhile, we propose using $q = 1$, which seems to lead to a conservative test statistic.

In order to ascertain how significant the departure of U(1) is from its nominal value 1, we again propose a bootstrap approach. Intuitively, we wish to see if the departure of $R(\hat{\lambda})$ from its (approximate) theoretical value 1 is significantly different from the departures of $R^*(\hat{\lambda}^*)$ from their theoretical (bootstrap) value $R(\hat{\lambda})$.

Formally, we compute

$$U^* = U^*(R(\hat{\lambda})) = \frac{R^*(\hat{\lambda}^*)}{R(\hat{\lambda})} \tag{25}$$

for each bootstrap sample and compute a bootstrap *p-value* as defined in (22) with T replaced by U.

Computational Issues

Computing the statistics $T(e)$ or $U(q)$ involves minimising a sum (over n paired rankings) with respect to the set of permutations. This calculation can be done approximately if a good guess for the minimiser is available and then a local minimiser is sought in the neighbourhood of the guess (see the "PINOUGHT" routine of Critchlow [1]). We chose to seek the global minimum by an exhaustive search procedure — the practical feasibility of this choice becomes unrealistic for $k > 7$.

The computational load is of course much more serious given the bootstrap analysis which requires finding the minimiser for each bootstrap sample. For this reason bootstrap replicates were limited to $B = 100$. The bootstrap samples were generated in the program by multinomial sampling with the aid of IMSL routine "RNUND" .

We computed the Model I and II statistics based on three permutation metrics: Kendall's τ distance (I); Spearman's rank correlation (S); and the Footrule (D). (The notation corresponds to Diaconis' [4] (p. 112) .) Hence, there are six candidate statistics: the T and the U statistics for each metric I, S, and D .

5.4 Analysis of Simulated Data Sets

In order to gain some experience with the methods proposed in the previous section, we generated data sets according to the two models described in Section 5.2. Each of the six test statistics was computed for each data set created, and the corresponding bootstrap *p-value* was computed.

The pre-rankings were generated at random (subroutine "RNPERM" of IMSL) and for each pre-ranking the corresponding post-ranking was generated according to the appropriate model. The distance d used in the generation of the paired rankings was the Kendall metric I. Rankings of $k = 5$ items were generated, and the sample sizes were $n = 50$ for each test data set. As mentioned earlier, in the analysis of each test data set, bootstrap samples of size $B = 100$ were chosen. Such an analysis runs for about six minutes of CPU time on a CONVEX computer. However, paying greater attention to vectorization should reduce the time further.

Note that the rankings are expressed in their *inverse* or *bracket* forms in the tables. In order to obtain some feeling for the non-uniformity of the distribution of ν given π we also note that for $\theta = 0.1$ the value of $e^{-0.1}$ is 0.90, so that for models I and II (with ϕ fixed), for each unit of distance further away from π, the probability of obtaining such a ν drops by a ratio of 0.9. For the Kendall metric for $k = 5$, the range of values is $0 \leq I \leq 10$.

In Tables 1 and 2 we have analysed data generated for a null model with $\theta = 0.1$ and $\theta = 0.2$, respectively. The true η under Model I is the identity $< 12345 >$; and the true λ under Model II is of course arbitrary, or undefined. We see that the *p-values* are all large, and that we would not suspect deviations from the null model if we were given these data sets. For the case of larger values of θ (say $\theta = 0.3$), we have found that the estimated value of η is nearly always the identity ranking.

In Tables 3, 4, and 5 we have analysed data sets generated from Model I, with $\theta = 0.1$, $\theta = 0.2$ and $\theta = 0.3$, respectively. In this simulation $\eta = < 25134 >$. We see that the T test does not succeed for the first case. In this case, the post-rankings ν are not very tightly concentrated about the shifted pre-ranking $\pi \circ \eta$, and so the detection of the effect of η

is difficult. For the cases $\theta = 0.2$ and $\theta = 0.3$ the T test has much better success, and in the latter case has little trouble in correctly estimating η, and in detecting its significant effect. For the three cases, the U tests show no significant deviations from the null model as is to be expected given that they are sensitive to different forms of deviation from the null model.

Tables 6, 7, 8, and 9 are the results of analysing test data sets generated from Model II with various values of θ and ϕ, for a given $\lambda = < 25134 >$. (There is no connection between the role of η in Model I and λ in Model II — we just arbitrarily chose the same permutation for generating the non-null test data sets.)

There seems to be something a little strange about the behaviour of the test based on U. For $\theta = 0.1$ it performs reasonably, as it does for the case $\theta = 0.3, \phi = 0.3$. However, for the case $\theta = 0.2, \phi = 0.3$ it fails (and does so on several repetitions of the simulation experiment). This anomaly could be connected with some of the limitations and reservations about the efficacy of the procedure which were mentioned in the previous section. Nevertheless, there is some room for optimism based on the tabulated results.

As anticipated, the T test does not detect deviation in the Model II direction. In this sense, notwithstanding the problems concerning U, if the T test does come up significant it seems that we can be quite confident that something of a Model I departure is being picked up.

It is clear that these test data sets do not provide a comprehensive justification for using the proposed test statistics. In particular, the U test for Model II type departures, being derived in a heuristic manner, needs a more extensive investigation of its properties. Nevertheless, these data sets do indicate that the procedures are potentially very useful for detecting two very different departures from null models for paired rankings.

TABLE 1:
Test Data – Null Model
$(k = 5, \theta = 0.1, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall Spearman Footrule	0.947 0.968 0.972	$< 21543 >$ $< 21543 >$ $< 12543 >$	0.56 0.59 0.61
U	Kendall Spearman Footrule	1.000 0.940 0.972	$< 45213 >$ $< 34512 >$ $< 34512 >$	1.00 0.62 0.94

TABLE 2:
Test Data – Null Model
$(k = 5, \theta = 0.2, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	0.936	$< 13425 >$	0.33
	Spearman	0.928	$< 13425 >$	0.14
	Footrule	0.914	$< 14325 >$	0.16
U	Kendall	1.007	$< 23145 >$	0.86
	Spearman	0.967	$< 25413 >$	0.86
	Footrule	0.966	$< 21435 >$	0.86

TABLE 3:
Test Data – Model I
$(k = 5, \eta = < 25134 >, \theta = 0, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	0.889	$< 25341 >$	0.36
	Spearman	0.934	$< 25341 >$	0.48
	Footrule	0.899	$< 25341 >$	0.18
U	Kendall	1.025	$< 25143 >$	1.00
	Spearman	0.953	$< 53124 >$	0.74
	Footrule	0.985	$< 52314 >$	0.99

TABLE 4:
Test Data – Model I
($k = 5, \eta = < 25134 >, \theta = 0.2, n = 50, B = 100$)

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	0.796	$< 21435 >$	0.01
	Spearman	0.873	$< 25134 >$	0.09
	Footrule	0.862	$< 21435 >$	0.09
U	Kendall	1.033	$< 23145 >$	1.00
	Spearman	0.979	$< 15243 >$	0.97
	Footrule	0.993	$< 13254 >$	0.99

TABLE 5:
Test Data – Model I
($k = 5, \eta = < 25134 >, \theta = 0.3, n = 50, B = 100$)

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	0.785	$< 25134 >$	0.00
	Spearman	0.834	$< 25134 >$	0.00
	Footrule	0.792	$< 25134 >$	0.00
U	Kendall	0.956	$< 32415 >$	0.86
	Spearman	0.927	$< 32415 >$	0.45
	Footrule	0.954	$< 43512 >$	0.88

TABLE 6:
Test Data – Model II
$(k = 5, \lambda = <25134>, \theta = 0.1, \phi = 0.1, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	1.000	$<12345>$	1.00
	Spearman	1.000	$<12345>$	1.00
	Footrule	0.989	$<15342>$	0.82
U	Kendall	0.831	$<25134>$	0.07
	Spearman	0.821	$<25134>$	0.00
	Footrule	0.823	$<25134>$	0.01

TABLE 7:
Test Data – Model II
$(k = 5, \lambda = <25134>, \theta = 0.1, \phi = 0.3, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	0.939	$<12453>$	0.43
	Spearman	0.951	$<12453>$	0.30
	Footrule	0.973	$<12453>$	0.39
U	Kendall	0.835	$<21534>$	0.06
	Spearman	0.856	$<21534>$	0.04
	Footrule	0.843	$<21534>$	0.02

TABLE 8:
Test Data – Model II
$(k = 5, \lambda = < 25134 >, \theta = 0.2, \phi = 0.3, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	1.000	$< 12345 >$	1.00
	Spearman	1.000	$< 12345 >$	1.00
	Footrule	1.000	$< 12345 >$	1.00
U	Kendall	0.951	$< 25413 >$	0.85
	Spearman	0.917	$< 25413 >$	0.47
	Footrule	0.925	$< 25431 >$	0.64

TABLE 9:
Test Data – Model II
$(k = 5, \lambda = < 25134 >, \theta = 0.3, \phi = 0.3, n = 50, B = 100)$

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	1.000	$< 12345 >$	1.00
	Spearman	1.000	$< 12345 >$	1.00
	Footrule	1.000	$< 12345 >$	1.00
U	Kendall	0.739	$< 25143 >$	0.01
	Spearman	0.760	$< 25143 >$	0.00
	Footrule	0.757	$< 25143 >$	0.00

5.5 Analysis of Rogers Data

The data set due to Rogers [7] was reported in Critchlow and Verducci [3] and they discuss the meaning of the items being ranked. In brief, the rankings represent the ordering of four criticism styles (A,C,P,T) which are translated in Table 10 to styles (1,2,3,4) respectively. Thirty-eight students were asked to rank four passages, each of which corresponded to one of the styles, before and after a literature course. The data is given in Critchlow and Verducci [3].

The analysis performed by Critchlow and Verducci [3] was centered on testing whether post-rankings were being attracted to a particular *given* ranking ($< 2134 > = < \text{PCAT} >$) which was that of the course instructor. In terms of our models, their procedure tests for alternatives of the Model II type.

We applied our method of analysis to the same data, and the results are presented in Table 10.

TABLE 10:
Rogers Data
(k = 4, n = 38, B = 100)

Statistic	Metric	Statistic Value	Optimal Perm (η or λ)	Bootstrap p-value
T	Kendall	0.647	$< 4123 >$	0.00
	Spearman	0.709	$< 4123 >$	0.00
	Footrule	0.653	$< 4123 >$	0.00
U	Kendall	0.817	$< 2134 >$	0.13
	Spearman	0.796	$< 2314 >$	0.05
	Footrule	0.824	$< 2143 >$	0.21

The results show that there is considerably stronger evidence that the deviation from the null model is towards Model I. The evidence of a deviation from Model II is not as strong, although the bootstrap *p-values* for U may be conservative as discussed previously.

If we accept the conclusion of a deviation towards Model I, then we are led to a different understanding of the effect of the course on the students' ranking. Rather than diverting the students' preferences to a particular ranking, the effect of the course is to change the perception of the four styles in the eyes of the students.

The optimal $\hat{\eta}$ is $< 4123 >$ which corresponds to (2341) in the direct ranking notation $(\eta(1), \ldots, \eta(4))$ where $\eta(i)$ is the rank assigned to item i. In terms of the four styles, they are transformed as follows :

$$\begin{array}{cccc} \text{A} & \text{C} & \text{P} & \text{T} \\ \text{C} & \text{P} & \text{T} & \text{A} \end{array} \qquad (26)$$

After the course, students are treating C as they did A before the course, and so on.

If we wish to consider the Model II test statistics U, then we see that the estimated λ is $< 2134 > = <$ CAPT $>$. This ranking is at Kendall distance two from the idealised ranking $<$ PCAT $>$ tested for by Critchlow and Verducci [3].

This discussion leads one to suggest that the tools presented here can help discriminate between different psychological interpretations of the way in which a teacher educates his students. Is the teacher causing his students to think like he does (Model II); or is he enlightening the subject matter, causing the students to *label* the issues differently (Model I)? In the latter case, the students are consistently ranking according to their pre-course preference function, but the items have been re-ordered with respect to that function.

Acknowledgements:

I thank Doug Critchlow and Joe Verducci for showing me their preprint, and for their constructive criticism that led to improvements in the Model II analysis. I also thank two anonymous referees for their comments and useful suggestions.

5.6 References

[1] Critchlow, D.E. *Metric Methods for Analyzing Partially Ranked Data*, Lecture Notes in Statistics, **34**, 1985. Springer-Verlag, Berlin.

[2] Critchlow, D.E., Fligner, M.A. and Verducci, J.S. Probability models on rankings, *J. Math. Psych.*, **35**:294-318, 1991.

[3] Critchlow, D.E. and Verducci, J.S. Detecting a trend in paired rankings. To appear in *J. Roy. Statist. Soc., Ser. C*, **41**, 1992 .

[4] Diaconis, P. *Group Representations in Probability and Statistics* , Lecture Notes – Monograph Series, Vol. 11, 1988. Institute of Mathematical Statistics, Hayward, CA .

[5] Feigin, P.D. Analysis of paired rankings, *Preprint*, 1987.

[6] Feigin, P.D. and Alvo, M. Intergroup diversity and concordance for ranking data: an approach via metrics for permutations, *Ann. Statist.*, **14**:691–707, 1986.

[7] Rogers, T. Students as literary critics: A case study of the interpretive theories, processes and communities of ninth grade students. To appear in *J. Reading Behavior.*

[8] Thompson, G.L. Graphical techniques for ranked data, *This volume - discussion paper*, 1992.

6

Maximum Likelihood Estimation in Mallows's Model Using Partially Ranked Data

Laurel A. Beckett [1]

ABSTRACT Consider a sample from a population in which each individual is characterized by a ranking on k items, but only partial information about the ranking is available for the individuals in the sample. The problem is to estimate the population distribution of rankings, given the partially ranked data. This paper proposes use of an EM algorithm to obtain maximum likelihood estimates of the parameters in Mallows's model for the distribution of rankings. Medical applications are discussed where the items are manifestations of a disease or a developmental process, the ranking is the sequence in which they first appear over time, and the partial ranking results from observation of the subjects cross-sectionally or at a few specified times. The methods are illustrated for a longitudinal study of a community population aged 65 years and older, where the signs are self-reporting of impairment in different physical activities.

6.1 Introduction

This paper addresses the general situation in which each individual in a population has a specified sequence or ranking of k items, such as the order in which signs of a disease appear over time. Knowing the population distribution of rankings is important in clinical settings. For example, in AIDS research, the distribution of sequences in which the T lymphocytes' response to different antigens is lost could be used to stage asymptomatic patients or to monitor response to therapy. The sequence in which skills are acquired is of interest in developmental psychology in children. Conversely, in gerontology, physical or cognitive functions may be lost successively with age or disease.

If a large sample of people could be monitored carefully over a long enough time period for all the signs to appear, the population distribution of rankings could be estimated directly. Obtaining complete data on the

[1] Harvard University, Cambridge, MA

sequence of events is often difficult, however. Some signs may have appeared before the start of the study, and the order of appearance may not be known. Death, loss to follow-up, or termination of the study may prevent an individual from being followed until all signs have appeared. Finally, observations may be available only at a limited number of time points or even at a single time point, as in cross-sectional studies. The exact sequence of appearance of signs in an individual will then be unobservable. Such studies give rise to partial rankings of the items.

This paper extends the results of Smith and Evans [13] for cross-sectional data (one occasion of measurement for each individual) to the general case with observations at V epochs or occasions. Section 6.2 reviews a parametric model for ranking distributions proposed by Mallows [11] and gives notation for the partial ranking setting. Section 6.3 presents a method for obtaining maximum likelihood estimates and states some of their properties. In Section 6.4, the methods are illustrated for items measuring physical function in a sample of persons aged 65 years and older.

6.2 Notation

Let the complete sequence of k items for an individual be denoted by $\mathbf{R} = R_1, \ldots, R_k$, where R_j is the rank of the j^{th} item (in the clinical setting, the order of appearance of the j^{th} sign). The distribution of the vector $\mathbf{R}$ on S_k, the set of all $k!$ permutations of the integers $1, \ldots, k$ is given by $f(\mathbf{r}) = pr[\mathbf{R} = \mathbf{r}]$.

For many clinical examples, such as the signs or symptoms of a disease, the course of the disease is likely to follow similar, if not identical, patterns in different patients. Researchers in aging and in child development have suggested that there is a natural hierarchy of physical functions, so the sequence in which skills are acquired by children or lost in illness or aging is most likely to correspond to the natural ordering [10]. Deviations may occur but extreme changes of the ordering are unlikely. These theoretical considerations suggest a population distribution where most individuals have sequences which are near each other in some metric on S_k.

Mallows [11] proposed a parametric model for f embodying this idea, having the form

$$f(\mathbf{r}) = C(k, \lambda)e^{-\lambda d(\mathbf{r}, r_0)} \tag{1}$$

where r_0 is a location parameter, the most likely permutation, on S_k. The function $d(\mathbf{r}, r_0)$ is a distance which measures how far the individual's sequence is from the most likely sequence, and λ is a scale parameter, so that e^{λ} is the proportional reduction in the likelihood when the distance from r_0 is increased by a single unit. The constant of summation $C(k, \lambda)$ is given by

$$C(k,\lambda) = \left[\sum_{\mathbf{r}\in S_k} e^{-\lambda d(\mathbf{r},r_0)}\right]^{-1}. \tag{2}$$

Various choices are possible for the metric d. Diaconis [5] discussed several choices. The example here uses Kendall's metric, which defines $d(\mathbf{r}, \mathbf{s})$ to be the minimum number of transpositions of adjacent items to go from $\mathbf{r}$ to $\mathbf{s}$, but the methods and results can be generalized to other metrics.

Feigin and Cohen [6] gave an algorithm for maximum likelihood estimation when the complete ranking is available for an individual. Fligner and Verducci [7] have discussed generalizations of Mallows's model, and Critchlow [2] considered extensions to partially ranked data. In other work, Fligner and Verducci [7] have used the idea of a metric on ranks to test the hypothesis of uniformity over all rankings against an alternative reflecting a trend.

In the clinical setting with observations at V epochs or occasions, only partial rankings are available. For each individual there is an unobservable complete ranking $\mathbf{R}$ and a vector $\mathbf{T}$ where the entry T_v denotes how many signs have appeared by the v^{th} epoch of observation. The times of observation are assumed to be independent of the rankings. The observable data for individual i at epoch v are T_v and an unordered list of the signs which have appeared by that time; these can be summarized by a vector $\mathbf{X}$ which records for each sign j when it was first observed:

$$X_j = \begin{cases} 0 & \text{if } R_j \leq T_1, \\ v & \text{if } T_v < R_j \leq T_{v+1}, \\ V & \text{if } T_V < R_j. \end{cases} \tag{3}$$

Thus smaller values of X_j denote those signs observed earlier, which correspond to those signs with lower ranks for that subject.

The data $\mathbf{X}$ will generally provide information that the original ranking $\mathbf{R}$ could only be one of a small subset of the permutations S_k. For example, suppose that $k = 4$ and $V = 2$, with one symptom observed at the first time and two more at the second time. The representation $X_1 = 0, X_2 = X_3 = 1$, $X_4 = 2$ indicates that symptom 1 was already present at the first time of observation, symptoms 2 and 3 appeared between the first and second times of observation, and symptom 4 had not yet appeared at the second time of observation. This vector $\mathbf{X}$ could have arisen either from the ranking (1,2,3,4) or from (1,3,2,4). Thus a given pattern of censoring, $\mathbf{T}$, partitions S_k into equivalence classes of rankings, each of which would give rise to the same partial ranking when the censoring pattern $\mathbf{T}$ prevails.

Critchlow [2] noted that it is useful to represent the equivalence classes determined by the pattern of observation times, $\mathbf{T}$, as right cosets. Let S_T denote the subgroup of S_k formed of all permutations which give partial rankings equivalent to that obtained from $(1, 2, \ldots, k)$ under the observation pattern $\mathbf{T}$. For each equivalence class of rankings under the observation pattern **T**, there corresponds a right coset of S_T. Each partial ranking **X** can be identified with the set of all full rankings which induce it under this observation pattern, and this set is a right coset of S_T, namely, $S_T\mathbf{r}$, where **r** is any permutation which induces this partial ranking. We write this set as $S_T(\mathbf{X})$. The set of all such partial rankings for a given **T** can be identified with the set of all corresponding right cosets, denoted S_k/S_T.

The next section discusses the use of the X_j to estimate the two parameters of the Mallows model, the location parameter r_0 and the scale parameter λ.

6.3 Maximum Likelihood Estimation Using the EM Algorithm

When the rankings **R** are known for all the individuals in the sample, the maximum likelihood estimates can be obtained by a simple iterative process, as described in Feigin and Cohen [6]. The likelihood becomes much more complex in the partial ranking setting. Critchlow [2] modified the Mallows model to permit metrics defining the distance between partial rankings. This approach is appealing when the goal is to describe the distribution of partial rankings under a fixed design for incomplete information, **T**. The methods are computationally tractable as well. In the clinical setting, however, the goal is to estimate the population distribution of the complete sequences, and the times of observation **T** are random variables. Thus the approach used in Smith and Evans [13] for a single occasion of observation ($V = 1$) was to adapt the EM approach proposed by Dempster, Laird, and Rubin [3] and estimate the true frequencies of the full rankings in the sample, then maximize the resulting likelihood. In the present paper, this technique is extended to multiple occasions of measurement.

The steps in the EM algorithm are:

1. Initial step. Obtain initial estimates of the location parameter r_0 and the scale parameter λ.

2. E-step. Use the current parameter estimates to estimate the expected value of the sufficient statistics for the complete ranking data.

3. M-step. Use the estimated sufficient statistics to obtain maximum likelihood estimates of the two parameters.

The E-step and M-step are iterated until convergence is obtained.

The initial estimates are determined by the marginal frequencies of the times of first occurrence of the k items. The mean times of first occurrence are calculated by

$$\bar{X}_j = \sum_{i=1}^{n} X_j(i) \tag{4}$$

where i indexes the subjects in a sample of size n. The initial rank of item j, $\hat{R}(j)$, is one more than the number of items j' with $\bar{X}_{j'} < \bar{X}_j$. Thus the most frequently occurring symptom is estimated to have rank 1, and the least frequent to have rank k. A reasonable initial estimate for λ is to average $(k - \bar{X}_{R_j})/(k - \bar{X}_{R_{j+1}})$ over j and take the logarithm of the average. This estimate is based on the property of Mallows's model that there is a constant reduction in the likelihood, e^{λ}, for every unit increase in the distance from the most likely permutation. In practice, convergence does not seem particularly sensitive to choice of the initial value for λ.

Under the assumption that $\mathbf{R}$ and $\mathbf{T}$ are independent, the (unobservable) counts $n(\mathbf{r})$, the numbers of individuals with sequence $\mathbf{r}$, are sufficient for the scale and location parameters. The E step of the EM algorithm requires estimation of these numbers using the current estimates of r_0 and λ and the observed counts with each partial ranking, $m(\mathbf{X})$. For each person in the sample, the possible permutations which could have given rise to the observed data under the sampling pattern T are identified by using right cosets, noting that $\mathbf{X}$ and $\mathbf{r}$ are compatible if $\mathbf{r}$ belongs to the right coset determined by $\mathbf{X}$ under T, $S_T(\mathbf{X})$. A necessary and sufficient condition for $\mathbf{r} \in S_T(\mathbf{X})$ is:

$$(R_j - R_{j'})(X_j - X_{j'}) \geq 0 \qquad \text{for all } j, j'. \tag{5}$$

This condition guarantees that items ranked lower in $\mathbf{R}$ occur at earlier occasions of observation in $\mathbf{X}$.

The likelihood of each possible permutation $\mathbf{r}$ in $S_T(\mathbf{X})$ is found using equation (1), and the relative likelihood over all of permutations in $S_T(\mathbf{X})$ is given by

$$g_X(\mathbf{r}) = \frac{e^{-\lambda d(\mathbf{r},r_0)}}{\sum_{\mathbf{s}\in S_T(\mathbf{X})} e^{-\lambda d(\mathbf{s},r_0)}}. \tag{6}$$

The observed number of individuals with each partial ranking, $m(\mathbf{X})$, is then distributed over the full rankings in $S_T(\mathbf{X})$ to give the E-step estimates of the number observed with each full ranking, $n(\mathbf{R})$. Since $\mathbf{r}$ may give rise to different partial rankings under varying observation patterns, it is necessary to sum over all $\mathbf{X}$, giving

$$\hat{n}(\mathbf{r}) = \sum_{\mathbf{X}} m(\mathbf{X}) g_X(\mathbf{r}) I_{[\mathbf{r}\in S_T(\mathbf{X})]}. \tag{7}$$

The M step then finds maximum likelihood estimates of the parameters r_0 and λ, using the current estimates $\hat{n}(\mathbf{r})$ of the complete data on permutations for each individual in the sample. This part of the algorithm has been described in Smith and Evans [13], following Critchlow [2]. Maximizing the likelihood iteratively over all possible estimates of r_0 is usually unnecessary in practice, since the likelihood declines rapidly away from the initial guess, regardless of λ. A modified method has been implemented which estimates

λ for the current estimates of the complete data and for the initial estimate of r_0, then checks the rankings in the neighborhood of the initial estimate to see if re-estimating λ improves the likelihood. The E step and M step are iterated until convergence is obtained.

Theorem 1. *The initial estimates $\hat{R}(j)$ for r_0, based on equation (4), are consistent estimates of the true parameters, provided that $pr[T_v - T_{v-1} < k] > 0$ for at least one observation epoch v.*

PROOF: By the Strong Law of Large Numbers, the sample mean times of first occurrences $\bar{X}_j$ defined in equation (4) converge almost surely to their population means μ_j. Without loss of generality, let $r_0 = (1, \ldots, k)$. It now suffices to show that $\mu_j < \mu_{j'}$ if $j < j'$.

Fix a pair of items with $j < j'$. Partition the sample space of all possible rankings, S_k, into $(k-2)!$ subsets corresponding to the possible choices of ranks for the other $k-2$ items. Each cell in the partition then contains two rankings, one with $R_j < R_{j'}$ and one with $R_j > R_{j'}$. The probability of the second sequence, with items j and j' in the "wrong" order, is always less than the probability of the first sequence by at least a factor of $e^{-\lambda}$ since it requires at least one additional pairwise adjacent switch from r_0.

The difference between the mean times of first occurrence of items j and j' is given by

$$\begin{aligned}
\mu_{j'} - \mu_j &= \sum_{v<v'} (v'-v)\big(pr[X_{j'} = v', X_j = v] - pr[X_j = v', X_{j'} = v]\big) \\
&= \sum_{v'<v} (v'-v) \sum_{A(v',v)} \big(pr[R_j < R_{j'}] - pr[R_j > R_{j'}]\big) \\
&\geq \sum_{v'<v} (v'-v) \sum_{A(v',v)} pr[R_j < R_{j'}](1 - e^{-\lambda}) \\
&= (1 - e^{-\lambda}) \sum_{v<v'} (v'-v) pr[X_{j'} = v', X_j = v] \\
&> (1 - e^{-\lambda}) pr[X_j < X_{j'}]. \qquad (8)
\end{aligned}$$

where $A(v', v)$ denotes the cells in the partition which give items j and j' the times of occurrence v and v'. Note that the last line of equation (8) contains the probability that items j and j' appear at different times and in the correct order; this probability depends on the observation pattern T, on the positions of j and j' in the most likely ranking r_0, and on the parameter λ. The probability will be positive if any symptoms ever occur at different times. ▪

Remarks

a. The probability in the last line of equation (8) that items appear at different times and in the correct order depends on the observation pattern $\mathbf{T}$, on the positions of j and j' in the most common ranking r_0, and on the parameter λ. The probability will be smaller for items whose true ranks are adjacent, for small values of λ, and when sampling times are rarely expected to occur between items. For example, it may be difficult to distinguish between the last two items if few people in the study are observed at a time when all but one symptom has occurred.

b. Theorem 1 is stated for Kendall's metric, but the proof is valid for any metric for which the distance from the origin increases when two items are transposed out of their "correct" order relative to each other. This class includes Spearman's metric. Fligner and Verducci [7] discussed more general families of models with this property.

c. Equation (8) can be used to give a lower bound on the probability that the initial estimates $\hat{R}_j$ are correct, since this requires that $\mu_{j'} > \mu_j$ for all j', j with $R_{j'} > R_j$.

Theorem 2. *An upper bound on the probability that the initial guess at the most likely ranking is wrong is given by:*

$$pr[\hat{r}_0 \neq r_0] \leq \sum_{j<j'} E[e^{-c(j,j')N(j,j')}] \tag{9}$$

where $N(j, j')$ is the number of individuals for whom items j and j' occur at different times, and $c(j, j')$ is a positive constant depending on λ.

PROOF: As in Theorem 1, assume that the most likely ranking is $(1, \ldots, k)$. The initial guess at the most likely ranking will be wrong if any of the pairs of mean first occurrence times are reversed from the correct order, so that

$$\begin{aligned} pr[\hat{r}_0 \neq r_0] &\leq \sum_{j<j'} pr[\bar{X}_{j'} < \bar{X}_j] \\ &= \sum_{j<j'} \mathrm{E}_N pr[\bar{X}_{j'} < \bar{X}_j | N(j,j')] \\ &< \sum_{j<j'} 2\mathrm{E}_N[e^{-N(j,j')\epsilon^2/(2pq+2\epsilon/3)}] \end{aligned} \tag{10}$$

where the last inequality follows from a result in Uspensky [14], with p defined to be the probability that item j' occurs before item j if they are not tied, $q = 1 - p$, and $\epsilon = .5 - p$. The exponential term in the last line of equation (10) is decreasing in p, and, for Kendall's metric, $p < e^{-\lambda(j'-j)} < .5$, so that the probability of equation (10) is bounded by

$$pr[\hat{r}_0 \neq r_0] < 2 \sum_{j<j'} \mathrm{E}_N[exp(\frac{-N(j,j')(1/4 - e^{-\lambda(j'-j)} - e^{-2\lambda(j'-j)}}{1/3 + 4/3e^{-\lambda(j'-j)} - 2e^{-2\lambda(j'-j)}})]. \tag{11}$$

Provided that the distribution of observation patterns $\mathbf{T}$ guarantees some untied observations, this probability converges to 1 rapidly as the sample size increases. ■

Conditional on the true ranking r_0 being known, the likelihood $f(\mathbf{R}|\lambda)$ is a regular exponential family with regard to λ. Dempster, Laird, and Rubin [3] showed that repeated application of the E and M steps leads to the maximum likelihood estimate based on the incomplete data in the regular exponential family case. Thus, if the correct estimate of the true ranking r_0 has been obtained, the EM procedure will lead to the maximum likelihood estimate of the scale parameter, λ. Theorem 2 ensures that the probability of the correct estimate being chosen is close to 1 for large samples. In the low-probability case that the correct estimate of r_0 is not chosen, the properties of the EM algorithm for estimating λ are not known.

Testing goodness of fit of the Mallows model is complicated by the possibility of a large number of cells (k^{V+1} possible values of the vector $\mathbf{X}$). As in the cross-sectional case presented in Smith and Evans [13], where $V = 1$, the expected number of observations in cell $\mathbf{X}$ can be calculated, conditional on the pattern of times of observation, $\mathbf{T}$, as

$$\mathrm{E}[m(\mathbf{X})] = \Big(\sum_{X' with\ same T} m(\mathbf{X}') \Big) \sum_{\mathbf{r} \in S_T(\mathbf{X})} \hat{g}_{\mathbf{X}}(\mathbf{r}) \tag{12}$$

Asymptotically, if r_0 were known, the chi-square statistic $\sum_{\mathbf{X}}(m(\mathbf{X}) - \mathrm{E}m(\mathbf{X}))^2/\mathrm{E}m(\mathbf{X})$ has a chi-square distribution with $k(V+1) - 2$ degrees of freedom (one degree lost for fixing the total number for each pattern T, one for estimating λ.) In practice, if n is large enoughto justify a goodness of fit test, the probability of having the correct value for the most likely sequence r_0 is near 1, and cells can be combined if the expected values are too small. The goodness of fit test is illustrated for the data set in the next section.

More detailed diagnostics can be carried out if the Mallows model does not appear to fit well. Patterns of deviation may be identified by a residual analysis, similar to the spectral analysis proposed in Diaconis [4, 5]. This analysis is particularly straightforward for Kendall's metric. For a fixed censoring pattern, $\mathbf{T}$, each of the possible realizations of $\mathbf{X}$ has the same number of rankings in the corresponding right coset, $S_{\mathbf{T}}(\mathbf{X})$. It can be shown by symmetry arguments that the probability under Mallows's model with Kendall's metric of the coset is proportional to the minimum distance from r_0 to a compatible ranking in the coset. Since the probability decreases exponentially with distance, we would expect the logarithm of the frequencies to decrease linearly with distance. Moreover, the slope should be the same for all values of $\mathbf{T}$, although the intercepts may differ.

Thus the residual analysis begins by finding the shortest distance from a ranking in $S_{\mathbf{T}}(\mathbf{X})$ to r_0, say $d(\mathbf{X})$. Then graphical displays and analysis of covariance can be used to check whether the pairs $(d(\mathbf{X}), \log m(\mathbf{X}))$ follow parallel linear relationships for the values of **T**. A curvilinear relationship would suggest that Kendall's distance may be the wrong metric. Failure of parallelism would suggest that **T** and **R** are not independent, since the values of r_0 and λ may depend on how many ties have occurred and whether they are early or late in the observation period. Other deviations from parallel lines might identify specific symptoms as failing to fit the model.

6.4 Example

The methods proposed in Section 6.3 are now illustrated with an example involving 3809 non-institutionalized persons aged 65 years and over in East Boston who were interviewed annually about common medical and social problems of older people. Boston is one center of the US National Institute on Aging Established Populations for Epidemiologic Studies of Elderly (Cornoni-Huntley et al., 1986) [1]. Smith and Evans [13] looked at cross-sectional results from three different sets of questions about physical function in the initial year of the study, separately for males and females. The discussion in the present paper focuses on a single set of questions, the Rosow and Breslau items [12], for which the cross-sectional analyses suggested that the Mallows model was a good fit and that the parameters were close for males and females.

In-home interviews were conducted in year 1 and year 4 of the study. At each of these times the respondent was asked whether he or she could climb a flight of stairs, walk half a mile, or do heavy work around the house without aid. Table 1 shows the self-reported ability to perform each of the three activities in year 1 and year 4. Data from males and females have been pooled. Respondents were excluded from the present analysis if they were missing any items at one or both interviews or if the answers were provided by a proxy rather than self report. There were 2653 subjects with complete responses at both year 1 and year 4 interviews. The most commonly reported symptom of impaired physical function among the Rosow-Breslau items was being unable to do heavy work around the house, followed by inability to walk half a mile and inability to climb a flight of stairs. This is in agreement with the cross-sectional data reported by Smith and Evans [13].

The present analysis coded the j^{th} item by $X_j = 0$ if the respondent reported being unable to perform the activity at the year 1 interview, $X_j = 1$ if the respondent could perform the activity at year 1 but not at year 4, and $X_j = 2$ if the respondent could perform the activity at both years.

Table 1.
Change in Self-Reported Physical Function: Year 1 to Year 4.
East Boston Older Persons

Number of Subjects Able to Perform Activity Combinations Unaided

Yr. 1					Yr. 4				
	- - -	- -H	-S-	-SW	W- -	W-H	WS-	WSH	Missing
- - -	129	1	46	3	7	1	23	7	125
- - H	5	1	2	0	0	0	0	2	4
-S-	43	0	101	22	1	0	46	19	83
-SH	21	3	21	10	0	0	4	31	18
W- -	7	0	2	1	3	0	21	6	10
W-H	3	0	2	0	1	0	3	3	0
WS-	8	0	60	12	8	1	224	159	78
WSH	5	5	65	54	4	7	227	1148	115
Missing	7	0	9	2	1	1	11	6	21

W = walk half a mile or 8 city blocks
S = climb a flight of stairs
H = heavy work around the house

Of the 2653 who had complete self report at both interviews, 1148 never reported any impairment and 129 were impaired in all activities at both interviews. Potential ambiguities in the coding scheme might arise if impairment were reported in one activity at year 1 but not at year 4 and in another at year 4 but not at year 1. Such cases were rare (23 out of 2653, or less than 1%) and were handled by using only the time of first report of impairment and ignoring recoveries. Thus the 2 people who reported that they could do housework but could not climb stairs or walk half a mile at year 1 and that they could climb stairs but not walk half a mile or do housework at year 4 were coded with $X = 0$ for walking half a mile, $X = 1$ for climbing stairs, and $X = 2$ for doing heavy work around the house. The estimates of the parameters λ and r_0 are based on two years of data from the $1374 = 2653 - (1148 + 129)$ people with some impairment at one or more year and no ambiguity and on the first year of data from the 23 with self-reports which could not be coded.

Table 2 shows the partial orderings created from the data in Table 1 by using this coding procedure. Thus, for example, the 7 people coded with $X = 0$ for walking and climbing stairs and $X = 1$ for doing housework would be consistent with either the sequence walk, stairs, housework or the sequence stairs, walk, housework. The most common sequences are those consistent with the ordering in which ability to do housework was lost first, followed by ability to walk half a mile and to climb stairs.

Table 2.
Partial Rankings from East Boston Physical Function Data Time of First Appearance of Inability to Perform Activity

Time of First Impairment WSH	No. of Subjs.	Time of First Impairment WSH	No. of Subjs.	Time of First Impairment WSH	No. of Subjs.
000	217	001	7	002	3
010	44	011	21	012	3
020	186	021	25	022	41
100	10	101	5	102	0
110	8	111	5	112	5
120	72	121	65	122	54
200	30	201	4	202	3
210	9	211	4	212	7
220	383	221	227	222	1148

W = walk half a mile or 8 city blocks
S = climb a flight of stairs
H = heavy work around the house

0 = impairment reported at first interview
1 = impairment reported at second but not first interview
2 = impairment reported at neither interview

The parameters of Mallows's model were fitted using the EM algorithm as described in the preceding section, and a Fortran program on an IBM PC-compatible 386 personal computer. The maximum likelihood estimate of the location parameter or most likely sequence was housework, walking, stairs, and the scale parameter estimate was $\hat{\lambda} = 1.70$. This is consistent with the separate analyses of year 1 data for males and females reported in Smith and Evans [13], where the same sequence was found for both genders, with $\hat{\lambda} = 1.90$ for males (based on $n = 477$ with some impairment) and 1.70 for females ($n = 892$). Fitting the Mallows model to the year 1 and year 4 marginals of the pooled data given in Table 1 for the 2653 with no missing data gave similar values, with $\hat{\lambda} = 1.72$ for year 1 ($n = 1049$) and $\hat{\lambda} = 1.91$ for year 4 ($n = 1016$). The value $\hat{\lambda} = 1.70$ can be interpreted to mean that every reversal of an adjacent pair of symptoms from the most common ordering, heavy work – walk half mile – stairs, reduces the likelihood of the new ordering by $e^{1.70}$ or about 72%. Fitting the Mallows model not only confirms the obvious ordering, but also provides a quantitative summary of how strongly this sequence predominates in the population.

The next question is whether the Mallows model gives an adequate fit to the data. There are 27 cells in Table 2, of which 3 (those with X values (0,0,0), (1,1,1), and (2,2,2)) are not informative. The remaining 24 cells can be pooled according to the right coset to which they correspond and grouped by the partition into which the right cosets divide the space of all permutations, using equation (5). For example, the censoring patterns with one impairment in one item reported first and the other two later but not ordered can all be treated as a group. Conditional on these censoring patterns, the expected number can be found by applying Mallows's model with the maximum likelihood estimates to the total number with the censoring pattern.

Table 3 shows the 24 informative cells grouped according to three possible censoring patterns: one symptom each before year 1, at year 4, and never seen; one symptom at year 1 or 4 and the other two symptoms both at year 4 or never; and two symptoms both at year 1 or both at year 4 and the third later or never. For each grouping, the possible right cosets are shown and the corresponding partial orderings are given. The observed frequencies are shown and compared with the expected frequencies under Mallows's model with the maximum likelihood estimates. There was good agreement between the observed and expected frequencies; none of the three chi-square statistics was significant.

Table 3.
Adequacy of Fit of Mallow's Model To East Boston Physical Function Data

Time of First Impairment (X(W),X(S),X(H))	Compatible Rankings	Observed Frequency	Expected Frequency
(0,1,2)	WSH	3	2
(0,2,1)	WHS	25	16
(1,0,2)	SWH	0	0
(1,2,0)	HWS	72	78
(2,0,1)	SHW	4	3
(2,1,0)	HSW	9	14
(0,0,1),(0,0,2),(1,1,2)	WSH,SWH	15	12
(0,1,0),(0,2,0),(1,2,1)	HWS,WHS	295	294
(1,0,0),(2,0,0),(2,1,1)	HSW,SHW	44	50
(0,1,1),(0,2,2),(1,2,2)	WSH,WHS	116	127
(1,0,1),(2,0,2),(2,1,2)	SHW,SWH	15	18
(1,1,0),(2,2,0),(2,2,1)	HWS,HSW	618	607

W = walk half a mile or 8 city blocks
S = climb a flight of stairs
H = heavy work around the house

0 = impairment reported at first interview
1 = impairment reported at second but not first interview
2 = impairment reported at neither interview

Figure 1 presents a graphical display of the log of the observed frequencies of the 24 informative cells versus the minimum distance from the estimated most likely ranking. The cells have been grouped more finely than in Table 3, by exact censoring pattern, and one-half was added to frequencies before taking logs since some cells were empty. The plots generally appear consistent with the model, showing parallel lines, although some sample sizes are small. Thus, Mallows's model appears to provide an adequate fit to the partially ordered data given by the longitudinal data on physical function.

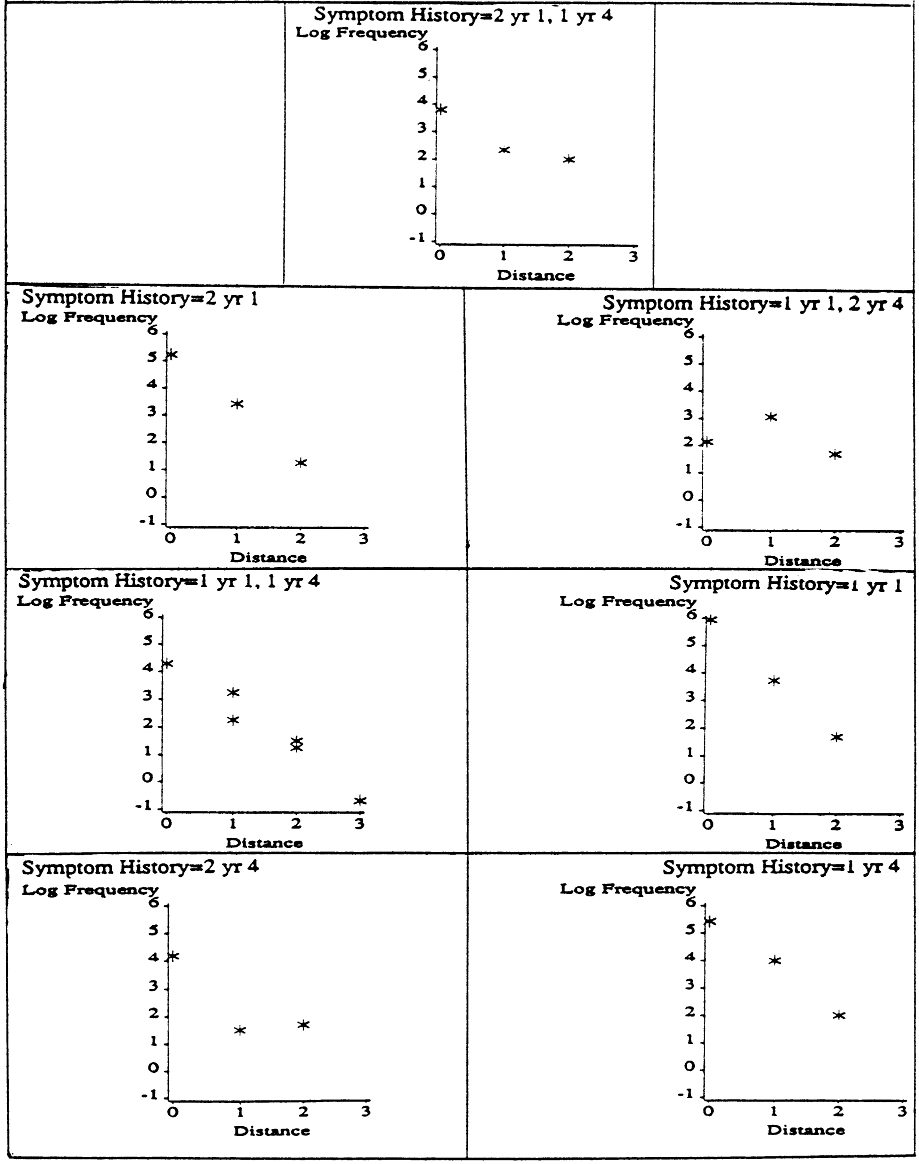
6.5 Discussion

The methods here permit estimation of a population distribution of sequences from incomplete longitudinal data, that is, when the order in which symptoms occur is known only up to a partial ranking. The family of models proposed by Mallows allows for a most likely sequence, with other sequences less likely according to how far they are away from the most likely sequence in some metric. This class of models is a helpful way to summarize the natural history of disease in a population, where some typical or most common sequence of events may characterize many but not all members of a population. The methods proposed here allow the use either of cross-sectional data or, more generally, data collected at a limited number of time points. The parameters can be estimated by maximum likelihood using the EM algorithm. Goodness of fit can be tested using the expected number of observations under Mallows's model, conditional on the censoring patterns observed.

Two key assumptions are made. First, each individual is assumed to acquire or lose symptoms in a characteristic fixed sequence. The possibility that one symptom may appear, then disappear while another appears, is not considered within this model. Second, the pattern of censoring, that is, how many new symptoms have appeared by each occasion of observation, is assumed to be independent of the individual's sequence. The residual analyses described here can be used to test this assumption. If the most likely ranking or the change in likelihood for other rankings differs according to how rapidly the symptoms are occurring, the plots will fail to have parallel lines.

There are many potential applications of these methods in studying natural history of disease. In AIDS research, for example, it would be desirable to know the order in which positive response to skin tests of various antigens is lost as immune function declines; this would permit more refined

FIGURE 1
Model Validation Using Graphical Displays



Symptom History=2 yr 1, 1 yr 4

Symptom History=2 yr 1

Symptom History=1 yr 1, 2 yr 4

Symptom History=1 yr 1, 1 yr 4

Symptom History=1 yr 1

Symptom History=2 yr 4

Symptom History=1 yr 4

staging of disease, monitoring loss of immune function by less invasive procedures, and possibly early evaluation of the efficacy of therapy. Further research problems include the development of efficient computational methods for large numbers of symptoms and large numbers of times of observations and the comparison of estimates for subgroups of people within the population. Inconsistent longitudinal sequences, which have been ignored here, might be handled by modifying the model or by treating them as measurement errors. Finally, a very interesting problem is subset selection. It is desirable to be able to identify useful groups of symptoms. For example, a set of symptoms could be selected which had the largest scale parameter λ; this would give a population distribution highly concentrated around a single sequence or permutation, so that most people would have a very similar natural history.

Acknowledgements:

Supported in part by a grant (AG02107) and a contract (AG92117) from the National Institute on Aging and by the Statistical and Data Analysis Center of the AIDS Clinical Trials Group under a contract from the National Institute of Allergy and Infectious Disease (AI95030).

6.6 References

[1] Cornoni-Huntley, J., Brock, D. B., Ostfeld, A. M., Taylor, J. O. and Wallace, R. B. (eds), *Established Populations for Epidemiologic Studies of the Elderly Data Book*, US Government Printing Office, Washington D.C., 1986.

[2] Critchlow, D., Metric methods for analyzing partially ranked data, *Springer Lecture Notes in Statistics* **34**, 1985.

[3] Dempster, A. P., Laird, N.M. and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Statist. Soc. B* **39**, 1–38, 1977.

[4] Diaconis, P., *Group Representations in Probability in Statistics*, Institute of Mathematical Statistics, Hayward, 1988.

[5] Diaconis, P., A generalization of spectral analysis with application to ranked data, *Ann. Math. Statist.* **17**, 949–979, 1989.

[6] Feigin, P. D. and Cohen, A., On a model for concordance between judges, *J. R. Statist. Soc. B* **40**, 203–213, 1978.

[7] Fligner, M. A. and Verducci, J. S., Distance based ranking models, *J. R. Statist. Soc. B* **48**, 359–369, 1986.

[8] Fligner, M. A. and Verducci, J. S., A nonparametric test for judge's bias in an athletic competition, *Appl. Statist.* **37**, 101–110, 1988.

[9] Fligner, M. A. and Verducci, J. S., Posterior probabilities for a consensus ordering. *Psychometrika* **55,** 53–63, 1990.

[10] Katz, S., Ford, A.B., Moskowitz, R.W., et al. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. *J. Am. Med. Assoc.* **53**, 914–923, 1963.

[11] Mallows, C. Non null ranking models I. *Biometrika* **49**, 114–130, 1957.

[12] Rosow, F. and Breslau, N., A Guttman health scale for the aged. *J. Gerontol.*, **21**, 556-559, 1966.

[13] Smith, L.A. and Evans, D.A., Estimating a population distribution of sequences of *k* items from cross-sectional data, *Appl. Statist.*, **39**, 1990.

[14] Uspensky, J.V., *Introduction to Mathematical Probability*, New York: McGraw Hill, p. 205, 1937.

7

Extensions of Mallows' ϕ Model

Lyinn Chung[1]
John I. Marden[2]

ABSTRACT Mallows' ϕ model is a one-parameter exponential family model for vectors of ranks. Fligner and Verducci have extended this model to multistage ranking situations. In this paper we introduce a class of models based on so-called *orthogonal contrasts* of the objects to be ranked, which we use to analyze three sets of data. The first set, from the GRE, consists of 98 students' ranking of five words according to their association with the word *idea*. The second is the American Psychological Association's 1980 presidential election data. The final set illustrates an approach to rank-based analysis-of-variance.

7.1 Introduction

The ranking literature contains a series of papers that develops theory and practice for a particular set of nonnull ranking models based on Kendall's τ metric. Mann [23] first presented the basic model in order to exhibit an alternative distribution against which Kendall's τ test is most powerful. Mallows [22] derived this model from the point of view of ranking data, i.e., data arising from a number of judges ranking a number of objects. He also incorporated the Spearman's ρ metric into the model. (For a discussion of these and other metrics, see Critchlow [10] and Diaconis [14].) The model based solely on the τ metric was termed the "ϕ model," a name that has become familiar in the field. Later, Shulman [26] arrived at the ϕ model using a slightly different argument. Analysis and use of the ϕ model can also be found in Feigin and Cohen [16], Cohen and Mallows [8, 9], Cohen [6], Fligner and Verducci [18, 19], Chung and Marden [5] and Marden [24]. Barton, David and Mallows [2] proposed a similar model when there are only two distinct types of objects (or ranks).

The basic ϕ model has only one parameter, hence would not be expected to fit especially well for large data sets, large either in the number of

[1] Clinical Statistics, Abbott Laboratories, Abbott Park, Illinois
[2] Department of Statistics, University of Illinois, Champaign, Illinois

judges or objects. Fligner and Verducci [19] greatly expanded the possibilities by extending the model to multistage ranking models. At the first stage, the judge decides which object to rank #1. At the second stage, the judge decides which of the remaining to rank #2, ..., until the ranking is complete. In their ϕ component model, the stages are independent with a single parameter for each stage. Setting these parameters equal yields the ϕ model.

Chung and Marden [5] were interested in applying such models to ranked data obtained by calculating the rank statistics from independent samples; data for which popular nonparametric procedures such as Kendall's τ, Mann-Whitney and Wilcoxon, Jonckheere-Terpstra, Kruskal-Wallis, and Friedman tests are indicated. Since observed data is often rife with ties, the ϕ model was extended to allow for ties. Also, in order to analyze analysis-of-variance problems, a larger class of possible stages, which we call *orthogonal contrasts*, were introduced.

Marden [24] suggested some models for ranking situations that used contingency table models with the factors being orthogonal contrasts of the objects to be ranked. Goldberg's [20] data on the ranking of 10 professions on their perceived social prestige was analyzed, continuing work of Feigin and Cohen [16], Cohen and Mallows [8, 9], Cohen [6], Fligner and Verducci [18, 19], and van Blokland [31].

In this paper we apply the above notions to three data sets, hoping to show the usefulness of these models in a variety of settings. The first consists of 98 students' ranking of five words according to their association with the word *idea*. Fligner and Verducci [18] apply a ϕ component model. We look at some other ϕ models based on sets of orthogonal contrasts, and find some very satisfactory models. See Section 7.4. In Section 7.5 the data consist of the rankings of five candidates for president in the 1980 American Psychological Association elections. Diaconis [15] presents an extensive analysis of this data. Unlike our first example, there is enough data (5738 full rankings) to use a complete contingency table analysis. We decompose the rankings into a four-way table by using a set of orthogonal contrasts, thereby arranging the data in a format convenient for analysis. Using this structure, we compare the results of voters who give full rankings to those who only specify their top one, two or three favorite candidates.

The final example, in Section 7.6, exhibits a possible approach to analysis of variance (ANOVA) based on ranks. There is a vast literature on this topic, but not much on testing for two- and higher-way interactions. One approach that has gained wide currency is to perform the usual normal-theory calculations on the ranks rather than the raw data. When testing for main effects in location-family models, this method works well. However, Thompson [30] shows that in a two-way ANOVA, if both main effects are present but there is no interaction, the usual statistic for testing that there is no interaction has an asymptotic mean of infinity, suggesting that the true level approaches 1. We take a different tack, placing a parametric model, the ϕ model with orthogonal contrasts, on the ranks themselves. Main effects and interactions are then defined in terms of the model's parameters. Although the model we use does not arise from any location-family linear model, the hope is that the model describes the true distribution well enough, and that it presents a convenient framework for an ANOVA.

Section 7.7 provides some further discussion of contrasts, in particular how to choose them, and why orthogonality is important.

7.2 The General Model

We assume we have a set $\mathbf{O}_m$ of m objects, denoted $\mathbf{O}_m = \{o_1, o_2, \cdots, o_m\}$, and the data consist of a number of independent rankings of the objects. Mathematically, we are interested in models on $\mathbf{P}_m \equiv \{$*permutations of the first m integers*$\}$. A vector $y \in \mathbf{P}_m$ could be an *ordering* of the objects, in which case $y_i = j$ means object o_j is ranked i^{th}, or it could be a *ranking* of the objects, in which case $y_i = j$ means that object o_i is ranked j^{th}. For exposition it does not matter, but the two possibilities do in general give rise to distinct models. We will present our orthogonal contrast models in terms of rankings, so that for example $y = (3, 2, 4, 5, 1)$ means object o_1 is ranked third, o_2 is ranked second, ..., and o_5 is ranked first. See Remark 1 for models on orderings.

Our models are based on sets of *orthogonal contrasts* of the objects. A *contrast* is a comparison of groups of objects. Consider an example in Böckenholt [3] in which there are eight objects $\equiv$ soft drinks:

$$\begin{aligned}\mathbf{O}_m = \{&\text{Coke, Pepsi, 7-up, Sprite, Diet Coke,}\\ &\text{Diet Pepsi, Diet 7-up, Diet Sprite}\}. \end{aligned} \tag{1}$$

One might wish to compare the colas to the non-colas, or the diet drinks to the non-diet drinks. Technically, a contrast C is defined by $C \equiv (I_1, \cdots, I_K)$ where $I_1, \cdots, I_K$ are disjoint nonempty subsets of $\mathbf{O}_m$. Then some possible contrasts are

C_1 = ({Coke, Pepsi, 7-up, Sprite}, {Diet Coke, Diet Pepsi, Diet 7-up, Diet Sprite};

C_2 = ({Coke, Pepsi, Diet Coke, Diet Pepsi}, {7-up, Sprite, Diet 7-up, Diet Sprite});

C_3 = ({Coke}, {Pepsi}, {Diet Coke}, {Diet Pepsi});
and

C_4 = ({Coke, Pepsi}, {7-up, Sprite}).

Thus C_1 compares the non-diet to the diet drinks, C_2 compares the colas to the non-colas, C_3 compares the four colas to each other, and C_4 compares the non-diet colas to the non-diet non-colas.

The *value* a particular judge has for a particular contrast is defined to be the set of ranks of the objects in the subsets $I_1, \cdots, I_K$ relative to the union $C^U \equiv \cup_{k=1}^{K} I_k$. The relative ranks of the objects *within* groups I_k are irrelevant. The formal definition follows.

Definition 1 For $y \in o\mathbf{P}_m$ and $C = (I_1, I_2, \cdots, I_K)$, a contrast of objects in $\mathbf{O}_m$, the *value* of the contrast at the ranking y is

$$C(y) = (\{\bar{y}_i | o_i \in \mathbf{I}_1\}, \{\bar{y}_i | o_i \in \mathbf{I}_2\}, \cdots, \{\bar{y}_i | o_i \in \mathbf{I}_K\})$$

where $\bar{y}_i$ is the rank of y_i relative to the ranks $\{y_i | o_i \in C^U\}$. □

Continuing the Soft Drink example, suppose $y = (3, 8, 5, 7, 6, 2, 4, 1)$. The values of the four contrasts follow.

Values for y=38576241

Contrast	C_1	C_2
Value	$(\{3,5,7,8\},\{1,2,4,6\})$	$(\{2,3,6,8\},\{1,4,5,7\})$
Contrast	C_3	C_4
Value	$(\{2\},\{4\},\{3\},\{1\})$	$(\{1,4\},\{2,3\})$

The values for C_1 and C_2 are found by just grouping the ranks for the diet and non-diet drinks, or cola and non-cola drinks, respectively. The value for C_3 is found by first finding the ranks for the colas, 3862, then reranking these relative to each other: 2431. These numbers are the $\bar{y}_i$'s. C_4 is similar but groups the non-diet colas and non-colas.

Knowing the value of a contrast only gives partial information about the entire ranking y. However, if one know the values for enough of the contrasts, the entire ranking can be reconstructed. In particular, it is enough to know just the pairwise contrasts $(\{o_i\},\{o_j\})$ for $i < j$. The minimal number needed is $m - 1$. Special sets of contrasts, orthogonal contrasts, are especially efficient. Two contrasts are orthogonal if the comparisons they represent are not confounded, as below.

Definition 2 Two contrasts, C as above and $D = \{J_1, \cdots, J_L\}$, are said to be *orthogonal* if either

$$(i)\ C^U \cap D^U = \emptyset;$$

$$(ii)\ C^U \subset J_l \text{ for some } l;$$

or

$$(iii)\ D^U \subset I_k \text{ for some } k.$$

A set of contrasts, $(C_1, ...C_T)$, is orthogonal if each pair is orthogonal. □

The orthogonality designation is justified by noting that if $Y \sim$ Uniform $(\mathbf{P}_m)$, then $C_1(Y), \cdots, C_T(Y)$ are independent, each distributed uniformly over its space. The idea is that if two contrasts involve completely distinct sets of objects, or if one contrast makes comparisons solely within one of the groups in the other contrast, then the contrasts are orthogonal.

In the example, C_1 and C_4 are orthogonal since the first compares non-diets to diets, while the second is a comparison within the non-diets. Likewise, C_2 and C_3 are orthogonal since the former compares colas to non-colas, and the latter compares types of colas. None of the other pairs are orthogonal. For example, C_1 and C_2 both involve comparisons of Coke and Pepsi to Diet 7-up and Diet Sprite.

An orthogonal contrast model depends on a set of T orthogonal contrasts, $(C_1, C_2, \cdots, C_T)$. If $T = m - 1$ then the set of possible values of this vector is in one-to-one correspondence with the set of rankings $\mathbf{P}_m$. Given a set of orthogonal contrasts, we consider the T-way contingency table for elements of $\mathbf{P}_m$ with the factors being the contrasts and the levels being the possible values of the contrast. Thus if $Y \sim$ Uniform$(\mathbf{P}_m)$, then the contingency table exhibits total independence of the factors, as well as uniform marginal distribution for each of the contrasts. Any other log-linear model can also be considered. With a sufficient number of replications of Y, the usual hierarchical models can be used to analyze the data. See the APA voting example in Section 7.4 with $m = 5$ and 5738 complete observations.

When the number of observations is not substantially larger than $m!$, the resulting contingency table will be too sparse to fit all the factorial models, so one may have to collapse some categories or make do with only lower-order models, or both. In many examples, one expects to observe a trend in the y_i's given by the order of the groups in a contrast C. It may then be sensible to collapse the categories of C by using the Jonckheere-Terpstra statistic, i.e.,

$$d(C(y)) = \sum_{\substack{i \in I_k \\ k < l}} \sum_{j \in I_l} I(\{y_i > y_j\}), \tag{2}$$

where $I(A)$ is the indicator function of A. Now $\underline{d}(y) \equiv (d(C_1(y)), \cdots, d(C_T(y)))$ is assumed to be a sufficient statistic. One can also combine contrasts in the following way. If for contrasts C and D as above, $C^U = J_l$ for some l, we can define the combined contrast of length $K + L - 1$ by

$$C^* D \equiv \ (J_1, ..., J_{l-1}, I_1, ..., I_K, J_{l+1}, ..., J_L). \tag{3}$$

If we now use (2) to reduce the model, we have $d(C^* D(y)) = d(C(y)) + d(D(y))$.

A special case of the model that assumes independence of the contrasts is the ϕ model, the exponential family model with $\underline{d}$ as the natural sufficient statistic. That is, for parameter $\underline{\theta} \in R^T$ and set of orthogonal contrasts

$\underline{C} \equiv (C_1, \cdots, C_T)$, the density of Y is

$$f(y; \underline{\theta}, \underline{C}) = e^{\underline{\theta}' \underline{d}(y) - \sum \Psi(\theta_i; C_i)} \tag{4}$$

with respect to Uniform($\mathbf{P}_m$). Since the contrasts are independent under the dominating measure, and the density factors, the contrasts are indeed independent under (4). From Chung and Marden [5], we have that for real θ and contrast C,

$$\Psi(\theta; C) = \Psi_{\#C^U}(\theta) - \sum_{k=1}^{K} \Psi_{\#I_k}(\theta), \tag{5}$$

where

$$\Psi_q(\theta) = \sum_{i=1}^{q} ln \left[\frac{1 - e^{i\theta}}{i(1 - e^{\theta})} \right].$$

Note that if we can combine two contrasts as in (3), then setting their parameters equal in (4) will automatically combine them in the natural sufficient statistic.

Mallows' ϕ model with modal ordering $(o_1, o_2, ..., o_m)$ is (4) with $C = (\{0_1\}, \{0_2\}, \cdots, \{0_m\})$, and the model in Barton, David and Mallows [2] is (4) with $C = (I_1, I_2)$. The general model with $T = 1$ contrast is given in Critchlow [10]. See Remark 3.

Remark 1 When $y \in \mathbf{P}_m$ represents the ordering of the objects, the contrasts are defined on the ranks rather than the objects. Thus the sets I_k are subsets of the ranks $\{1, 2, \cdots, m\}$. The values of the contrasts are then given in terms of the indices of the objects. Fligner and Verducci [18, 19] introduced such models with the contrasts

$$C_k = (\{k\}, \{k+1, \cdots, m\}), k = 1, \cdots, m-1.$$

Here, C_1 asks which object is ranked first, C_2 asks which is ranked first among the remaining $m-1$ objects, ..., and C_{m-1} asks which of the last two objects is ranked higher. Fligner and Verducci present the corresponding ϕ model, which they call the ϕ- component model, as well as the "Free" model, which posits the contrasts independent but otherwise unrestricted, and some other models with ordering constraints.

Remark 2 This section is historically backwards. Mallows' ϕ model is the starting point, followed by Barton, David and Mallows [2] and Critchlow [10], all of whom consider just one contrast. Fligner and Verducci [18, 19] were the first to use orthogonal contrasts. Chung and Marden [5] extended the ϕ-component model to arbitrary sets of contrasts, and Marden [24] in analyzing the Goldberg [20] data considered Free versions of those models, as well as models in which the contrasts were not necessarily independent.

7.3 Ties, Partial Rankings

The previous section assumes that the data y consist of complete rankings, but it is common for there to be ties, partial rankings or incomplete rankings. Ties may arise in a number of ways. In order-statistic models, one starts with a set of independent observations and then ranks them. If the underlying distributions are not continuous, there are likely to be ties, and in practice even with continuous distributions one often sees ties due to roundoff error. In ranking models, it may be that judges are asked to rank only their top 3 choices, or sort the objects into groups of 5, etc., or it may be that each judge decides how fine to make the ranking. Silverberg [27] models situations in which the judges rank the top q of their choices. Diaconis ([14], Chapters 5 and 9) approaches partial rankings group-theoretically, and in Diaconis [15] analyzes the APA voting data. Critchlow [10] presents a general method, with examples, for extending metrics on full rankings to metrics on partial rankings. Smith [24] has an example on medical data that consist of the subset of symptoms a person exhibits at a given point in time, yielding a ranking of "1" to the symptoms that have appeared, and a "2" to those that have not. See Sections 7.5 and 7.6 for other examples.

There are many ways in which information about rankings can be missing. The discussion here is geared toward situations wherein the ranking of the objects may be only a partial ordering, but it is complete in that for any $i \neq j$, we have either $y_i < y_j$, $y_i = y_j$, or $y_i > y_j$. In contrast, there may be m objects, but the judge only sees $m' < m$ of them, hence the unseen objects can not be compared to each other nor to the seen ones. Or one may only know that the judge prefers "1" to "2" and "3" to "4". The basic approach below can be followed for these more complicated structures, but the notation will become more cumbersome, and in particular the nice results for the ϕ model will not necessarily hold.

The models we present for cases in which y may have ties are motivated by censored-data models. That is, we assume that there is a latent random vector $W \in \mathbf{P}_m$ generated by one of the models in Section 7.2, but there is a variable, independent of both W and its distribution, that "ties" certain values in the vector W to produce the observed Y. Note that we have not yet specified the convention used to represent a vector with ties. Midranks are the most common, but any systematic approach will yield the same result. Chung and Marden [5] present an explicit representation of the independent variable and the tying mechanism, in which the vector y consists of the integers from 1 to S, where $S \equiv S(y)$ is the number of distinct elements in y. We will adopt this convention.

To illustrate, return to the soft drink example of Section 7.2. A person who likes non-diet colas best, then 7-up, then Sprite, then diet colas, then Diet 7-up, and lastly Diet Sprite, but who does not make a distinction between Coke and Pepsi, or Diet Coke and Diet Pepsi, would have $y = (1, 1, 2, 3, 4, 4, 5, 6)$ according to our notation, with $S(y) = 6$. (Midrank notation would yield (1.5,1.5,3,4,5.5,5.5,7,8).) One whose only distinction is between diet and non-diet drinks, preferring the former to the latter, has $y = (2,2,2,2,1,1,1,1)$.

Let the *pattern of ties* for y be defined by $T(y) \equiv (\#T_1(y), \ldots, \#T_S(y))$, where $T_s(y) = \{\ i \mid y_i = s\}$. For the above two examples of y, $T(y) =$ (2,1,1,2,1,1) and (4,4), respectively. Then for any distribution P on $W \in \mathbf{P}_m$, the corresponding probability for y is, conditioning on the ancillary $T(y) = t$,

$$P[Y = y \mid T(Y) = t] \;=\; P[W \;\; is\ consistent\ with \;\; y]. \tag{6}$$

Thus the conditional likelihood of any vector y of tied rankings can be found by simply adding the likelihoods of all complete rankings that are consistent with y. There are 4 complete rankings consistent with $y = (1,1,2,3,4,4,5,6)$ given by the possible relative preferences of Coke and Pepsi, and of Diet Coke and Diet Pepsi: (1,2,3,4,5,6,7,8), (2,1,3,4,5,6,7,8), (1,2,3,4,6,5,7,8) and (2,1,3,4,6,5,7,8). There are $4! \times 4!$ complete rankings consistent with $y = (2,2,2,2,1,1,1,1)$ given by the possible rankings within the non-diet drinks and within the diet drinks.

Whether this model is an appropriate representation for a particular situation has to be decided on a case-by-case basis. The least innocuous assumption is that the pattern of ties T is independent of the order W of the latent ranks. In the medical example above, it is reasonable as long as the observation time point is independent of the order of onset of the symptoms. In ranking situations, one is assuming that the judge could give a full ranking, but (due to time constraints, etc.) does not. This assumption seems fine if each judge is asked to give the same pattern t. However, if different judges give different patterns, it may be inappropriate to consider the judges' distributions of W to be homogeneous over patterns. See Section 7.5.

For an arbitrary model on $\mathbf{P}_m$, the presence of ties can cause identifiability and computational problems, but these are slight for the ϕ model. Chung and Marden [5] show that for given $\underline{C}$, $\underline{\theta}$ and pattern of ties t, Y has conditional density

$$f(y; \underline{\theta}, \underline{C}, t) \;=\; e^{\underline{\theta}' \underline{d}(y) - \sum \Psi(\theta_i; C_i, y)}, \tag{7}$$

where $d(C(y))$ is defined again by (2) even when y has ties, and

$$\Psi(\theta; C, y) \;=\Psi^U_{\#C}(\theta) \;-\; \sum_{k=1}^{K} \Psi_{\#I_k}(\theta) \;-\; \sum_{s=1}^{S} \Psi_{\#T_s}(\theta) \;+\; \sum_{k=1}^{K}\sum_{s=1}^{S} \Psi_{I_k \cap T_s}(\theta).$$

The dominating measure in this case is Uniform($\{y|T(y) = t\}$). Note that the density (7) factors just as the density (4) without ties. The $d(C_i(Y))$'s may not be independent, however, since the dominating measure is not generally a product measure.

In the examples in Sections 7.4 and 7.6, we use maximum likelihood estimators and likelihood ratio tests for inference about $\underline{\theta}$. Given independent observations y_1, ..., y_n with correspondings patterns of ties t_1, ..., t_n, we assume the data is generated according to (7). The maximum likelihood estimator for $\underline{\theta}$ is found by solving the equations

$$\sum_{j=1}^{n} \Psi'(\hat{\theta}_i; C_i, y_j) = \sum_{j=1}^{n} d(C_i(y_j)) \tag{8}$$

for $i = 1, ..., T$. Using Chung and Marden [5], we have that asymptotically, the $\hat{\theta}_i$'s are approximately independent normals, where the asymptotic mean and variance of $\hat{\theta}_i$ are θ_i and

$$v_i(\theta_i; C_i, y_j = \left[\sum_{j=1}^{n} \Psi''(\hat{\theta}_i; C_i, y_j)\right]^{-1}, \tag{9}$$

respectively.

The parameter $\underline{\theta}$ is not necessarily easy to interpret. Consequently, we also use the parameter τ based on Kendall's τ. When there are no ties in y, the sample Kendall's τ for contrast C is (10)

$$k(y; C) = 1 - 2 \cdot \frac{d(C(y))}{M(C)}, \quad M(C) \equiv \max\{d(C(y)) \mid y \in \mathbf{P}_m\}, \tag{10}$$

which takes values between +1 and -1, and has an interpretation similar to that of a correlation coefficient. When a contrast has $K = 2$ groups, then $M(C) = \#I_1 \times \#I_2$. The corresponding parameter is defined by $\tau(\theta) \equiv E_\theta[k(Y; C)]$. We want this parameter to be defined independently of the pattern of ties, hence in general we set $\tau(\theta) \equiv E_\theta[k(W; C)]$. If y has ties, we estimate $\tau(\theta)$ by first finding $\hat{\theta}$ as in (8), and then calculating

$$\tau(\hat{\theta}) \equiv E_{\hat{\theta}}[k(W; C)] = 1 - 2 \cdot \frac{\Psi\prime(\hat{\theta}; C)}{M(C)} \tag{11}$$

by (5). We use the Δ-method to find the asymptotic standard error of this statistic:

$$s.e.(\tau(\hat{\theta})) = \frac{2}{M(C)} \Psi''(\hat{\theta}; C)\sqrt{v(\hat{\theta}; C, \{y_j\})}. \tag{12}$$

Remark 3. In Section 7.2 we mentioned that Critchlow [10] contains the model (4) with $T = 1$. His model can actually be given as a special case of (7) in two distinct ways. Suppose the data arises from a judge sorting m objects into the top k_1 objects, next k_2, ..., and last k_K. If y is the ordering of the objects, then the corresponding *phi* model is (4) with the one contrast

$$C = (\{1, \cdots, k_1\}, \{k_1{+}1, \cdots, k_1{+}k_2\}, \cdots, \{k_1{+}\cdots k_{K-1}{+}1, \cdots, k_1{+}\cdots k_K\}).$$

On the other hand, if y is the vector of (tied) ranks, then the model is (7) with the one contrast being $C = (o_1, o_2, ..., o_m)$ and pattern of ties $t = (k_1, k_2, ..., k_K)$.

7.4 Example: Word Association

Fligner and Verducci [18] present data of 98 students' responses to a question on the Graduate Record Examination. The students were to rank five words according to their association with the word "idea," where the five words were 1) thought, 2) play, 3) theory, 4) dream, 5) attention. We will abbreviate these words to *Tt, P, Ty, D* and *A* at times. For a given student, $y_i = j$ means word i was ranked j, the higher the number j the stronger the association of the word i with "idea." The modal ranking was (5,1,4,3,2). Fligner and Verducci fit ϕ-component models. Table 1 contains a number of orthogonal contrast models, along with their Λ's for testing $\underline{\theta} = 0$, where $\Lambda = 2 \cdot ln(LRS)$, LRS being the Likelihood Ratio Statistic. We leave the brackets off singleton sets. The R^2 is the ratio of Λ to the maximum possible for this data set, the maximum being 529.43. The sets presented are a sample of the many tried. Although the search among the sets proceeded in an *ad hoc* manner, guidance can be found from the mean ranks of the words, which are (4.8,1.3,3.6,3.3,1.9). Thus one is immediately lead one to link words "play" and "attention," and words "theory" and "dream."

Table 1: Some ϕ-models and their R^2's

Contrast ↓	I = Mallows	II	III
C^1	(P,A,D,Ty,Tt)	(P,{Tt,Ty,D,A})	(Tt,{P,Ty,D,A})
C^2		(A,{Tt,Ty,D})	(Ty,{P,D,A})
C^3		(D,{Tt,Ty})	(D,{P,A})
C^4		(Ty,Tt)	(P,A)
Λ	435.80	454.50	456.20
R^2	82.3%	85.8%	86.2%

Contrast ↓	IV	V
C^1	(Tt,{P,Ty,D, A})	({Tt,Ty,D}, {P,A})
C^2	({P,A},{Ty,D})	(Tt,{Ty,D})
C^3	(Ty,D)	(Ty,D)
C^4	(P,A)	(P,A)
Λ	462.60	462.72
R^2	87.4%	87.4%

Models II and III are analogous to the ϕ-component models of Fligner and Verducci, defined on the objects rather than ranks. See Remark 1. One can see that all the models, even the one-parameter Mallows' ϕ model, pick up a substantial portion of the variation in the data. The best is model V as far as the Λ goes, although there is very little difference among the models

II through V. We tried many other models, none of which were as good as V.

The set of contrasts in V make some substantive sense. Contrast 1 compares (thought, theory, dream) to (play, attention), which divides the words into those fairly close to "idea" and those very far. The next two contrasts separate the 3 close words, and the last contrast compares the two others.

Before looking closely at the estimates for this model, we look at the fits of the individual contrasts to the ϕ model. Contrasts 3 and 4 fit exactly since they are binomials. Table 2 gives the distributions for Contrasts 1 and 2. The "G" statistic is the likelihood-ratio-like chi-squared statistic $G = 2 \cdot \sum Observed \cdot ln(Observed/Expected)$.

Table 2: Marginal fits for Model V

Contrast 1: {Tt, Ty, D} vs. {P, A} — G = 12.40

$d(C_1(y)) \rightarrow$	0	1	2	3	4	5	6
Observed	80	13	4	0	0	0	1
Expected	78.19	13.86	4.91	0.87	0.15	0.01	0.00

Contrast 2: Tt vs. {Ty, D} — G = 2.90

$d(C_2(y)) \rightarrow$	0	1	2
Observed	4	9	85
Expected	2.02	12.96	83.02

The ϕ model seems to fit well, except for the one observation in category "6" for Contrast 1. This person's ranks suggest that "play" and "attention" are both closer to "idea" than any of "thought," "theory" and "dream," which leads one to believe that this person misunderstood the instructions and ranked the words in reverse order. In fact, this person's ranks were (1,4,2,3,5), almost the exact reverse of the modal ranking. Fligner and Verducci [18] note this and one other observation as suspect. If we leave out this observation, and recalculate the ϕ model, we find $\Lambda = 485.93$, $R^2 = 91.5\%$, and the G's for testing the fits of Contrasts 1 and 2 are, respectively, 2.35 and 1.80. The model now seems to fit well, and we will continue without the outlier. Table 3 contains the calculations for the fitted model. The τ_i's and their standard errors are from (11) and (12).

Table 3: Statistics for ϕ Model V

Standard errors in parentheses

Contrast↓	$\hat{\theta}_i$	$\hat{\tau}_i$	Λ
1. {Tt,Ty,D} vs. {P,A}	1.91 (0.19)	-0.93 (0.018)	331.20
2. Tt vs. {Ty,D}	1.97 (0.25)	-0.85 (0.041)	125.42
3. Ty vs. D	0.57 (0.21)	-0.28 (0.098)	7.62
4. P vs. A	-1.00 (0.23)	0.46 (0.090)	21.70
Total			485.93

Table 3 shows that the τ's for Contrasts 1 and 2 are highly negative, meaning there is a very strong trend for people to associate "thought," "theory" and "dream" more closely with "idea" than "play" or "attention," and to associate "thought" more closely than "theory" or "dream." There is a mild tendency to prefer "theory" to "dream," and a slightly stronger tendency to prefer "attention" to "play."

We use contingency table techniques to investigate whether independence of the contrasts seems viable. In order to avoid too sparse of a table, Contrasts 1 and 2 were collapsed before making any tests, so that Contrast 1 has just the categories 0 and 1-6, and Contrast 2 has categories 0-1 and 2. Thus the data is now reduced to a 2^4 contingency table. It is still a fairly sparse table: Of the 16 cells, there are four 0's, two 1's, and three 2's. We start by testing pairwise independence of the contrasts, using the usual G-test of independence. The values of G for the 6 tests range from 0.045 to 3.386, revealing no evidence of dependence. Trying some other models, we have that the G for testing mutual independence of the four contrasts is 19.79 on 11 degrees of freedom (d.f.), that for testing there is no 3- or 4-way interaction is 11.38 on 5 d.f., and that for testing no 4-way interaction is .38 on 1 d.f. With the sparseness of the table, it is hard to know how seriously to take these latter results. Using 1000 simulations, we estimate the p-value for the test of mutual independence to be .058, which leads us to compare the observed and expected tables. We see that about 11 of the 19.79 G-points are due to the extreme cells (0,0-1,0,0) and (1-6,2,1,1), for which the observed counts are both 0 and the expected counts are 2.86 and 2.62, respectively. Thus there may be something slightly wrong here with the model, but overall the ϕ model fits quite well, and explains clearly the observed features of the data.

7.5 Example: APA Voting

Diaconis [14, 15] analyzes the data for the 1980 American Psychological Associations election. Five people (whom we will call A, B, C, D and E) were running for president, and voters were asked to rank their choices from 1=favorite to 5=least favorite. The Hare [21] system was used to determine the victor, who turned out to be Candidate C. Of the 15,449 ballots received, 5738 consisted of complete rankings. The others gave only their top 1, 2 or 3 candidates. Diaconis [15] used spectral analysis to analyze these data. A brief summary of his results are that there are two camps, people who like candidates A and C, and those who like candidates D and E. Candidate C received the most first-place votes but also a large number of last place votes. That there were two camps was not surprising as the APA is a strong mixture of academicians and clinicians. Diaconis also analyzes the voters who ranked only $q = 1$, 2 or 3 candidates. He finds that the $q = 1$ people have approximately the same profile as the people who gave full rankings, the $q = 2$ people liked candidate A better and D and E worse, while the $q = 3$ voters preferred candidates D and E. McCullagh [25] fit inversion models to the $q = 3$ and full data, as well as a latent class model with two classes to the model. The first class ranked the candidates in the order ED, BA, C, while the second class ranked them C, A, B, DE.

Stern [29] investigated several voting schemes in addition to the Hare system, and fit several models, including latent class models, to the data. The data analysis confirmed the previous results, but it is particularly interesting that some voting schemes elected Candidate C, and others Candidate A.

Our objective is to use orthogonal contrasts to arrange the data into a four-way contingency table. The ϕ models are not especially appropriate here since they detect monotone trends, while the voting tends to be quadratic in that some people get a high number of first and fifth place votes, while others get high numbers of second, third or fourth place votes. Thus instead of evaluating a set of contrasts on the Λ as in Section 7.4, we decompose the $G(= 1717.51)$ for testing uniformity versus the saturated model into the deviance due to the marginals of the contrasts and the two-, three- and four-way interactions. A good set of contrasts is one for which the resulting table is easy to understand, i.e., one for which lower order interactions have relatively more weight. From Diaconis' results, we know the contrasts should involve comparing AC with DE, and comparing A to C and D to E. Table 4 shows the calculations for several models for the set of complete rankings.

Model II is best in terms of having the most deviance explained by the marginals of the contrasts, while Model IV explains the most with just the marginals and two-way interactions. None of the models, except perhaps V, has much four-way interaction. We will concentrate on Model II, a good part of the reason being that it separates the noncontroversial candidate 2 from the others early, leaving the more interesting comparisons between and within the groups AC and DE. Table 18 shows the data arranged in the table for this model.

The question arises how seriously to take such putative chi-squared statistics. The data does not represent a random sample, some major deviations being a selection effect as well as people tending to vote in clumps. These chi-squared statistics can legitimately be thought of as results of randomization tests, or as the Kullback-Leibler distance between the observed distribution and the expected under the null model. (Actually, the distance is $ln(LRS)/5738$.) We will take the general view that if the statistic is not much larger than the degrees of freedom, then the effect is ignorable. Otherwise, we will look more closely to see whether something important and interesting is happening, or if there is just a wrinkle in the observed distribution.

Table 4: Models for the APA voting data: Decompositions of Λ

Degrees of freedom in parentheses

Model→	I	II	III
C_1	(C,ABDE)	(B,ACDE)	(ABC,DE)
C_2	(AB,DE)	(AC,DE)	(B,AC)
C_3	(A,B)	(A,C)	(A,C)
C_4	(D,E)	(D,E)	(D,E)
Effects↓			
Marginals	611.19 (11)	941.34 (11)	590.47 (13)
2-way interactions	914.32 (39)	569.77 (39)	956.99 (41)
3-way interactions	178.34 (49)	186.31 (49)	155.53 (47)
4-way interactions	13.66 (20)	20.10 (20)	14.53 (18)

Model→	IV	V
C_1	(AC,BDE)	(C,ABDE)
C_2	(B,DE)	(A,BDE)
C_3	(A,C)	(B,DE)
C_4	(D,E)	(D,E)
Effects↓		
Marginals	831.73 (13)	533.00 (10)
2-way interactions	762.61 (41)	1001.76 (35)
3-way interactions	98.01 (47)	142.43 (50)
4-way interactions	25.17 (18)	40.32 (24)

In Section 7.5, the complete rankings are analyzed thoroughly using straightforward contingency table analysis. In Section 7.5 we follow McCullagh [25] in fitting a latent class model to these data. Section 7.5 contains analysis of the partially ranked data, the main effort being to ascertain the difference between the partial and full rankings.

Complete Rankings - Contingency Table Analysis

Taking the contrasts in II, we tried to find a parsimonious log-linear model to fit the table. The model that fits the two three-way interactions Contrasts 1-2-3 and Contrasts 1-2-4 has a goodness of fit statistic of 32.78 on 29 degrees of freedom. Note this is the model positing that Contrasts 3 and 4 are conditionally independent given Contrasts 1 and 2. Presumably, all the information in the data can be found by scrutinizing the two 3-way tables.

We start with the marginals. Contrast 1 gives the rank of Candidate B. The percentages are (14,19,25,25,18), so that 14% of the voters ranked B first, 19% second, etc. This candidate is not particularly liked nor hated. Contrast 2 is the most interesting, as it pits candidates A and C against D and E. Let ij be the event that A and C are ranked i^{th} and j^{th}, in some order, among the four A, C, D and E. (That is, we are abbreviating value ({1,2},{3,4}) of the contrast to ”12.”) The marginal percentages for the values in the order (12,13,14,23,24,34) are (28,12,12,14,12,22). Half the voters ranked either AC or DE together as first and second, with AC favored more often. The other four possibilities have uniformly small percentages. This distribution reveals the same second-order effect that Diaconis finds. Finally, the pairwise Contrasts 3 and 4 show that A is barely preferred to C by a 50.5% majority, and 52% prefer D to E.

There is significant lack of independence between the contrasts (Table 4), so we investigate next the 2-way tables. Table 5 presents the results of the pairwise tests of independence.

Table 5: Testing pairwise independence of the contrasts

Contrasts →	1&2	1&3	1&4	2&3	2&4	3&4
G	207.29	15.77	37.96	285.36	27.65	4.95
d.f.	20	4	4	5	5	1

Clearly, the interactions of Contrasts 1 and 2 and Contrasts 2 and 3 contain the bulk of the pairwise dependence. For the former pair, the dependence lies primarily in the difference between the people who vote 12 for Contrast 2 and the others. The G for that collapsed table is 170.93 on 4 d.f., while the G for testing independence leaving out the 12 category is 36.37 on 16 d.f. Table 6 shows the distributions for the collapsed table.

Table 6: Distribution of Contrast 1 given Contrast 2

Entries are percentages of row totals

2. AC vs. DE ↓; 1. Rank of B→	1	2	3	4	5
12	8	14	35	24	19
13,14,23,24,34	15	21	21	25	18

The most striking feature is that those with Contrast 2 being 12 have a greater portion of their Contrast 1 values at 3, and fewer at 1 and 2, than the others. We interpret this to mean the people who are much in favor of Candidates AC insist on ranking Candidate B third rather than first or second, suggesting that B is closer to the DE camp.

The other large effect comes from Contrasts 2 and 3. Table 7 shows that the AC supporters prefer C to A by a 2:1 margin, while the DE supporters prefer A to C by a 3:2 margin.

Table 7: Distributions of Contrasts 3 and 4 given Contrasts 1 and 2

Entries are percentages within Contrast 1 or 2 categories

C_1: Rank of B →	1	2	3	4	5
C_3 - A preferred to C	57	48	50	51	48
C_4 - D preferred to E	51	54	49	43	44

C_2: AC vs. DE →	12	13	14	23	24	34
C_3 - A preferred to C	35	43	54	55	64	61
C_4 - D preferred to E	46	41	50	49	49	52

The Contrasts 1 and 3 have mild dependence given by the fact that 57% of the people who rank B first prefer A to C, while the people who rank B from second to fifth have percentages from 48 to 51 in favor of A. As before, people who do not put AC at the top tend to prefer A to C. For Contrasts 1 and 4, we see that people who like B prefer D to E, while people who rank B low prefer E to D. Contrasts 2 and 4 show that the AC people tend to prefer E to D, while the DE people prefer D to E. The last pair, Contrasts 3 and 4, show a weak dependence which, when conditioning on Contrasts 1 and 2, actually disappears. For this pair, we see that of the people who prefer A to C, 49% prefer D to E, and of those who prefer C to A, 46% prefer D to E.

It is interesting to see whether we can now place the candidates on a scale as in an unfolding model. Leaving out B, the order CAED works well since the AC supporters prefer C to A and E to D, while the DE supporters reverse these preferences. Placing B in the middle to obtain CABED is slightly misleading: B supporters prefer A to C, but also D to E.

Turning to the 3-way tables, consider Contrasts 1, 2 and 3. Testing the no-3-way interaction model for this collapsed table yields $G = 95.55$ on 20 d.f. Table 8 contains the observed percentage of people who rank A above C for the 30 Contrast 1 × Contrast 2 possibilities, as well as the expected percentage under the no-3-way interaction model.

Table 8: Contrast 3 by Contrasts 1 and 2

Entries are the percentage who favor A over C, Observed and Expected

1. Rank of B↓;						
2. AC vs. DE →	12	13	14	23	24	34
1	52,39	57,47	66,58	60,59	59,67	50,65
2	27,32	38,40	45,51	59,61	68,60	56,58
3	35,37	50,45	51,57	62,57	69,66	62,63
4	31,35	38,43	58,54	44,55	66,64	71,61
5	39,34	40,41	52,53	47,53	54,62	61,60

The biggest discrepancies between the observed and expected percentages occur in the first and fourth rows. When B is ranked first, the people who like AC rank A above C more often than expected, and the people who like DE do so less often. That is, when B is the favorite, the stark differences in feelings about A and C are moderated in the AC and DE groups. In the fourth row, last column, we see that 71% prefer A to C, while we expect only 61%. These people rank D and E first and second, in some order, and B fourth. They strongly dislike C, ranking A third and C last 71% of the time. The fourth row, fourth column contains those that rank D and E first and fifth, B fourth, hence A and C second and third. They like C better than A, 11 percentage points better than expected. The rest of the differences were less than 10 points.

Table 9: Contrast 4 by Contrasts 1 and 2

Entries are the percentage who favor D over E, Observed and Expected

1. Rank of B↓; 2. AC vs. DE →	12	13	14	23	24	34
1	55,49	48,44	60,52	48,51	49,52	48,55
2	49,52	45,47	55,55	68,54	57,55	50,58
3	49,47	48,42	44,51	51,50	52,51	47,54
4	38,41	30,36	44,45	35,43	44,44	60,47
5	42,43	40,38	45,46	37,45	46,46	53,49

Table 9 contains the analogous results for Contrasts 1, 2 and 4, where the entries are the percentage who prefer D to E. The cell with the largest difference between observed and expected percentages, row 2 column 4, are those that rank B second and D and E first and fifth. They prefer D to E 15 points more than expected. In the row 4 column 6 cell, we have B ranked fourth and DE ranked first and second. Now 60% prefer D to E, 12 points over the expected. All the other differences are less than 10 points.

It is somewhat dangerous to read too much into these three-way interactions, but they do seem to yield some information. We already know that the AC people prefer C to A and E to D, while the DE people have reversed preferences. In Tables 8 and 9, we see that these differences are enhanced when Contrasts 1 and 2 suggest the feelings about AC or DE are strong, and are moderated when the feelings are less strong.

Finally, just to make sure nothing drastic has been missed, we found the $(Obs - Exp)/\sqrt{Exp}$ scores for the full table when fitting the model with the two 3-way interactions. Of the 120 values, the largest was 1.58, hence we need not look any further.

Complete rankings - Latent Class Analysis

The existence of two groups in this data set cries out for some sort of latent class analysis. With two latent classes, we fit the model with complete independence of the contrasts, and the model with all two-way interactions. The former model yielded a G of 343.42 on 96 d.f., the latter G=13.11 on 18 d.f., which fits quite well. We used the E-M algorithm, trying several different starting values and obtaining essentially the same results.

The two classes will be denoted AC and DE for obvious reasons. The estimated percentage in the AC group is 58%. Table 10 gives the estimated marginal distributions of the contrasts for the two classes using the two-way interaction model. The pairwise tables for the two groups could also be compared, but in the interest of space and sanity we will not do so.

Table 10: Latent marginal distributions of the contrasts

Entries are percentages of the latent class

	Contrast 1: Rank of B				
Latent Group↓	1	2	3	4	5
AC	14	17	30	23	17
DE	14	21	18	27	21

	Contrast 2: AC versus DE					
Latent Group↓	12	13	14	23	24	34
AC	43	18	12	16	4	8
DE	6	5	13	11	24	41

	Contrast 3		Contrast 4	
Latent Group↓	A	C	D	E
AC	37	63	39	61
DE	70	30	60	40

These distribution are consistent with the results of the previous subsection. The AC group is more likely to rank B third rather than first or second as compared to the DE group; each group overwhelmingly prefers its candidates to the other group's; the AC group much prefers C to A and D to E, while the DE group has the opposite preferences.

Partial Rankings

We will follow Diaconis and categorize the voters by the number q of candidates they ranked. Our main interest is to try to fit models to the partially ranked data, and thereby compare those data to the complete rankings.

We start by testing whether indeed the fully ranked data is different from the $q = 1$, 2 or 3 data by collapsing the former to be consistent with each of the latter. The resulting chi-squared statistics are, respectively, 52.88 on 4 d.f., 181.01 on 19 d.f., and 139.38 on 59 d.f. The contrasts in Model II of Table 4 will continue to be used to pinpoint the differences in the groups.

For the $q = 1$ data, we cannot estimate very much, and are unable to test any of the log-linear models as in Section 7.5 since there are not enough degrees of freedom. The percentages of voters who rank the candidates (A,B,C,D,E) first for $q = 1$ and $q = 5$ are, respectively, (17,17,23,22,20) and (18,14,28,20,20). Thus the $q = 1$ people like B and D slightly better, and C less well. In fact, the $q = 1$ voters are much closer to uniformity than the $q = 5$ voters, the G's being 79.65 and 310.26.

Although the above fairly well exhausts the comparison of the two groups, since our main purpose is to illustrate the use of orthogonal contrasts, we will show how to compare the marginal distributions of the contrasts for the groups. For Contrast 1, we only know how many ranked B first, and how many second through fifth. Categorizing the $q = 5$ data similarly, we obtain a 2×2 contingency table. Testing homogeneity gives $G = 27.66$, the percentages obtainable from above. For Contrast 2, we ignore the people who ranked B first since they give no information on the AC/DE comparison. The two categories remaining are those that rank either A or C first, and those that rank either D or E first. Testing homogeneity now yields $G = 18.65$: 53.6% of these $q = 5$ voters choose A or C, but only 49.1% of the $q = 1$ voters do. (Note that these percentages are conditional on B not being ranked first. Other comparisons made below are similarly conditional.) For Contrast 3, only those who rank either A or C first are used. Homogeneity is not far off here. The G=4.97 with 40% of the $q = 5$ voters and 43% of the $q = 1$ voters preferring A to C. Contrast 4 is even closer, with G=1.62, and 51% and 53%, respectively, preferring D to E. Basically, the voters who only list one candidate seem to be milder in their preferences, especially in choosing B first, and in the AC/DE controversy.

When $q = 2$, it is possible to estimate more of the marginal distributions of the contrasts, as well as to test whether the model that the contrasts are independent. The latter test yields G=194.19 on 12 d.f., which means independence in untenable. We still cannot test whether the model with all the two-way interactions fits.

The main difference between the $q = 5$ and $q = 2$ groups is that the latter are much more likely to rank A and C first and second, in either order: 21% for $q = 5$ to 34% for $q = 2$. We now look at the contrasts. Contrast 1 categorizes people into those who rank B first, second, and third through fifth. The G for homogeneity is 32.70, with the percentages of the three categories for $q = 5$ and $q = 2$ being, respectively, (14,19,68) and (12,14,74). The latter group seems to think less well of B. Contrast 2 is a little more complicated. People who rank A and C (D and E) first and second clearly have Contrast 2 being 12 (34). Those who rank one of A and C (D and E) first and one of D and E (A and C) second have Contrast 2 being 13 or 14 (23 or 24). Finally, if only one of A, C, D, E is ranked first or second, then one can only distinguish between Contrast 2 being 12, 13 or 14, or being 23, 24 or 34. Thus we have six categories, as in Table 11, where G=155.09.

Table 11: Comparison of $q = 5$ and $q = 2$ on Contrast 2

Entries are percentage of q-group

AC vs. DE→	12	13,14	12,13,14	34	23,24	23,24,34
$q = 5$	21	16	15	15	16	17
$q = 2$	34	12	12	14	14	14

The difference is in the "12" category, the other categories being fairly uniform.

There is very little difference in the groups for Contrast 3, with G=.69, and a slight difference for Contrast 4, with G=4.32: 48% of q=5 prefer D, while 52% of q=2 do.

Finally, turn to the $q = 3$ data. Each of the marginal distributions of the contrasts is estimable, except that the last two categories in Contrast

1 must be collapsed since there is no way to distinguish between those who rank B fourth and fifth. We fit both the full independence model and the two-way interaction model, obtaining G's of 272.55 on 74 d.f. and 30.94 on 17 d.f. The latter model fits reasonably well, so we will use the estimates from it to compare the q=5 and 3 groups. See Table 12. The q=3 people like DE better, and are more likely to prefer A to C, than the q=5 group. The two groups are fairly similar on their ranking of B, and preferences between D and E.

Table 12: Comparison of $q = 5$ and $q = 3$ groups

Entries are percentages of q-group

Contrast 1: Rank of B

	1	2	3	45
$q = 5$	14	19	25	43
$q = 3$	16	18	21	45

Contrast 2: AC versus DE

	12	13	14	23	24	34
$q = 5$	28	12	12	14	12	22
$q = 3$	24	12	11	11	16	26

	Contrast 3		Contrast 4	
	A	C	D	E
$q = 5$	50	50	48	52
$q = 3$	60	40	48	52

Recall that the "12" and "34" people in Contrast 2 are quite opinionated on Contrasts 3 and 4. Thus the marginal comparisons of Contrasts 3 and 4 in Table 12 may be a bit misleading since there are distinct differences between q=5 and 3 on Contrast 2, leading us to consider Table 13.

Table 13: Distributions of Contrasts 3 and 4 given Contrast 2

Entries are percentage of group and category who prefer A to C

C_2: AC versus DE→	12	13	14	23	24	34
$q = 5$	35	43	54	55	64	61
$q = 3$	43	49	68	54	75	71

Entries are percentage of group and category who prefer D to E

C_2: AC versus DE→	12	13	14	23	24	34
$q = 5$	46	41	50	49	49	52
$q = 3$	38	32	56	33	53	55

We see that the q=3 group does fairly uniformly (over the Contrast 2 categories) like A more than the q=5 group does. By contrast, the q=3 group is more favorable to E than the $q=5$ group when $C_2=12$, 13 or 23, but agrees with the $q=5$ group otherwise.

To summarize the comparisons, it appears that those who only give their first choice are less involved in the AC/DE controversy, being more favorable toward B, and more uniform overall. The people who rank their top two are strong AC supporters, while the people who rank three are somewhat more supportive towards DE.

7.6 Example: ANOVA

Chapman *et. al.* [4] carried out an experiment to study differences in the cognitive processes between novice and expert searchers using online computer catalogs. There were seven subjects in each group, with each subject performing the same twelve searches. A *search* is a sequence of online commands whose goal is to find a particular library holding. Each subject, for each search, was to try to memorize the search and then reconstruct it on the computer. If the reconstruction was not perfect, the subject had another chance to memorize the search. The process continued until the search was reproduced accurately. Among the data collected were the number of tries until successful. These are the data we will analyze here. See Table 19.

The twelve searches are categorized in three ways: $A \equiv$ Large (5-6 commands) versus Small (1-4 commands); $B \equiv$ LCS versus LCS & FBR (LCS means Library Circulation System and FBR means Full Bibliographic Record. These are two different databases, each with its own set of commands.); and $C \equiv$ Real versus Nonsense. The "Real" searches are logical sequences of commands, while the "Nonsense" searches use legitimate commands but the commands do not make a coherent search. The main hypothesis is that the experts should do much better on the Real than the Nonsense searches, while the Novices should do equally well (or poorly) on each. This hypothesis suggests that experts can encode information in chunks.

For each of the four $A \times B$ categories, there were two Real searches and one Nonsense search, as in Table 14. The data can be found in the Appendix. Note that all the numbers are between one and four. It happens that on the Nonsense searches, the subjects are limited to three attempts. Thus in the analysis that follows, we set all the fours to three on the Real searches.

Table 14: Arrangement of the searches

Search # ↓; Variable →	A	B	C
1	Large	LCS	Real
2	Large	LCS	Real
3	Large	LCS&FBR	Real
4	Large	LCS&FBR	Real
5	Large	LCS	Nonsense
6	Large	LCS&FBR	Nonsense
7	Small	LCS	Real
8	Small	LCS	Real
9	Small	LCS&FBR	Real
10	Small	LCS&FBR	Real
11	Small	LCS	Nonsense
12	Small	LCS&FBR	Nonsense

The experiment is taken to be a randomized block design with the subjects being the blocks. Within each block we have a $2 \times 2 \times 2$ fixed effect ANOVA. The subjects are grouped into Novices and Experts. A usual normal-theory or other location family model could be assumed, but one might feel uneasy assuming a continuous distribution when the data contain only three values. An alternative would be contingency table models, with the Number of Searches as fourth factor. However, one would then have two $3 \times 2 \times 2 \times 2$ tables, each with only seven observations. We will introduce a third approach that treats the discreteness of the data exactly, but with fewer parameters than the contingency table models. Specifically, we suppose the seven Novice 12×1 vectors are independent and identically distributed, as are the seven Expert, with distributions of the form (7).

It is necessary, of course, to choose the set of orthogonal contrasts. The object is to make comparisons on the variables, but in order to preserve orthogonality it must be done in a nested fashion. We start with A, and let Contrast 1 compare the Large and the Small Searches, (where s_i denotes the i^{th} search): $C_1 = (\{s_1, s_2, s_3, s_4, s_5, s_6\}, \{s_7, s_8, s_9, s_{10}, s_{11}, s_{12}\})$. Now B has two contrasts, one for each level of A: $C_2 = (\{s_1, s_2, s_5\}, \{s_3, s_4, s_6\})$ and $C_3 = (\{s_7, s_8, s_{11}\}, \{s_9, s_{10}, s_{12}\})$. Finally, variable C has one contrast for each of the four $A \times B$ categories: $C_4 = (\{s_1, s_2\}, \{s_5\})$, $C_5 = (\{s_3, s_4\}, \{s_6\})$, $C_6 = (\{s_7, s_8\}, \{s_{11}\})$ and $C_7 = (\{s_9, s_{10}\}, \{s_{12}\})$.

Table 15 contains the results of fitting the model. Since the sample size is fairly small and the data contains only three distinct values, we worried that the distributions of the Λ's under the null $\underline{\theta} = 0$ would not be well approximated by independent χ-squared variables. To check, we simulated the null distribution by randomly permuting the elements within each of the 14 vectors and refitting the model. It turned out that the estimated distributions of the Λ's were very close independent χ^2_1 variables, and in particular that the estimated p-values were very close the nominal values.

Table 15: Estimates from the model (7)
p-values based on 1000 simulations

		Novices	
Contrast↓	$\hat{\theta}_i(s.e.)$	$\hat{\tau}_i(s.e.)$	$\Lambda(p-value)$
1	0.25 (.08)	-0.49 (.12)	12.01 (.000)
2	-0.51 (.24)	0.52 (.19)	5.40 (.028)
3	-0.02 (.19)	0.02 (.22)	0.01 (.896)
4	$-\infty$ (—)	1.00 (—)	8.21 (.008)
5	0.91 (.91)	-0.53 (.41)	1.26 (.274)
6	1.01 (.68)	-0.58 (.29)	2.77 (.108)
7	-1.01 (.68)	0.58 (.29)	2.77 (.108)
Total			32.44 (.000)

		Experts	
Contrast↓	$\hat{\theta}_i(s.e.)$	$\hat{\tau}_i(s.e.)$	$\Lambda(p-value)$
1	0.42 (.11)	-0.70 (.10)	24.82 (.000)
2	0.02 (.19)	-0.02 (.22)	0.01 (.919)
3	-0.03 (.24)	0.03 (.28)	0.01 (.910)
4	$-\infty$ (—)	1.00 (—)	10.99 (.003)
5	-0.41 (.67)	0.27 (.41)	0.41 (.523)
6	0.48 (.72)	0.31 (.43)	0.47 (.493)
7	-1.13 (.93)	0.63 (.35)	2.03 (.178)
Total			38.73 (.000)

The largest effect for both the Novices and Experts is Contrast 1, which shows that the Large searches are more difficult to memorize than the Small ones. For the Novices, Contrast 2 shows a reasonably large effect, suggesting that the Large searches with both LCS and FBR are more of a challenge than those with just LCS. This difference is not evident in the Small searches, nor for the Experts. If we take the four variable C contrasts together, testing $\theta_4 = \theta_5 = \theta_6 = \theta_7 = 0$ yields Λ's of 15.02 and 13.98, which suggests there are Real/Nonsense differences. (In Contrast 4, which compares Real searches 1 and 2 to Nonsense search 5, for no one was the Nonsense search easier than either of the Real ones. This particular Nonsense search appears to be especially difficult. It had a mean of 2.64 while the others had means ranging from 1.14 to 2.29.) The interesting aspect of these four contrasts is that for the Novices, two show positive τ's and two negative, while for the Experts the τ's are all positive. Thus the main hypothesis above, that the Experts do better on Real than Nonsense while the Novices do equally well on both, is consistent with the data. The standard errors for the τ's are too large to make any definitive conclusions, but the results are certainly suggestive.

In the Introduction we implied that we could produce an ANOVA table that decomposes the overall Λ into main effects, two-way interactions and a three-way interaction. These effects can be defined easily using linear constraints on the θ_t's. Table 16 gives the null and alternative hypotheses for each effect. Testing each pair of hypotheses yields a one-degree-of-freedom test for the effect. The logic is as follows. For the main A effect, it is clear that θ_1 is the parameter to test. Before testing for the main B effect, we want to make sure that there is no A×B interaction. Since θ_2 measures the B effect for variable A being Large, and θ_3 for variable A being Small, the effect of B is independent of the level of A if $\theta_2 = \theta_3$. Now to find the B effect we test $\theta_2 = \theta_3 = 0$ assuming the parameters are equal. There are four parameters, θ_4 to θ_7, for the C effect, representing the C effect for different levels of A×B. The main C effect is then found as for the B effect by testing the parameters are zero versus their equality. The A×B, A×C and A×B×C effects are found using these same four θ_i's. No three-way interaction occurs if the difference in C effect between the two levels of B within the first level of A is the same as that within the second level of A. That is, $\theta_4 - \theta_5 = \theta_6 - \theta_7$. (Interchanging the roles of A and B yields the same result.) To isolate the A×C interaction, we need to assume that the C effect does not depend on the level of B, but may depend on the level of A. Thus the alternative hypothesis is $\theta_4 = \theta_5$ and $\theta_6 = \theta_7$. The null then equates the parameters. The B×C effect is defined similarly.

Table 16: ANOVA effects defined via the model parameters

Effect↓	Null hypothesis	Alternative hypothesis
A	$\theta_1 = 0$	$\theta_1 \neq 0$
B	$\theta_2 = \theta_3 = 0$	$\theta_2 = \theta_3$
C	$\theta_4 = \theta_5 = \theta_6 = \theta_7 = 0$	$\theta_4 = \theta_5 = \theta_6 = \theta_7$
A×B	$\theta_2 = \theta_3$	$\theta_2 \neq \theta_3$
A×C	$\theta_4 = \theta_5 = \theta_6 = \theta_7$	$\theta_4 = \theta_5$ and $\theta_6 = \theta_7$
B×C	$\theta_4 = \theta_5 = \theta_6 = \theta_7$	$\theta_4 = \theta_6$ and $\theta_5 = \theta_7$
A×B×C	$\theta_4 - \theta_5 - \theta_6 + \theta_7 = 0$	θ_i's arbitrary

Table 17, column "ABC," contains the results of the hypothesis tests from the model given in Table 15. As before, both Novices and Experts show strong A effect. It is interesting that for the Novices, the only other large effect is the three-way interaction. In particular, the main C effect is very small. For the Experts, the C effect is quite significant, confirming the original hypothesis of the experimenters. The Experts also show a possible three-way interaction, presumably due mainly to the extreme behavior of Contrast 4.

In Table 17, the "Sum" is the sum of the seven effect Λ's, while the "Overall" is the statistic for testing $\theta_1 = \cdots = \theta_7 = 0$. In a balanced ANOVA, the Sum and Overall values should be equal. The slight discrepancy here is due to the A×C and B×C effects. The linear constraints are orthogonal, but the lack of homoscedasticity of the $\hat{\theta}_i$'s causes some dependence in the linear combinations.

Finally, note that the nesting of the variables could have been done in any of the 6 orders of A, B and C. Table 17 contains the ANOVA decompositions corresponding to each of the orderings. Different orders do give different results, although the overall conclusions do not change much. The main effects remain fairly constant, and for the Novices the three-way effect stays large. The major differences occur for the A×B effect among the Novices, for which the Λ ranges from 2.66 to 6.32, and for the interaction among the Experts, in which the Λ- points wander about. However, none of the latter values are especially significant. We chose to focus on the ABC ordering since, given a significant three-way interaction, we would prefer to be able to investigate the Real versus Nonsense effects for the levels of the other variables, variable C being the main one of interest. Of course, the ordering BAC would have also worked, and in fact gives almost identical answers.

Table 17: ANOVA Decompositions
A=Large-Small, B=LCS-LCS & FBR, C=Real-Nonsense

			Novices			
Effect↓; Order→	ABC	BAC	ACB	CAB	BCA	CBA
A	12.01	11.47	12.01	9.41	7.66	7.66
B	2.75	2.48	2.88	2.88	2.25	2.16
C	0.53	0.53	1.09	0.93	0.52	0.93
A×B	2.66	2.98	4.96	4.96	6.32	6.32
A×C	0.82	0.82	1.05	2.09	1.07	1.07
B×C	0.04	0.04	0.00	0.00	0.00	0.00
A×B×C	13.57	13.57	9.27	9.27	9.63	9.63
Sum	32.39	31.90	31.26	29.55	27.46	27.80
Overall	32.44	31.72	31.27	29.57	27.30	27.64

			Experts			
Effect↓; Order→	ABC	BAC	ACB	CAB	BCA	CBA
A	24.82	26.61	24.82	22.06	22.98	22.98
B	0.00	0.14	0.17	0.17	0.14	0.46
C	8.68	8.68	8.44	6.45	6.33	6.45
A×B	0.02	0.09	0.10	0.10	0.10	0.10
A×C	0.50	0.50	0.48	3.29	4.16	4.16
B×C	0.80	0.80	3.02	3.02	0.70	1.80
A×B×C	4.03	4.03	4.06	4.06	0.00	0.00
Sum	38.86	40.85	41.09	39.15	34.41	35.96
Overall	38.73	40.73	41.01	39.07	34.34	35.88

7.7 Discussion of Contrasts

It is likely that the reader will ask, "How is the set of orthogonal contrasts chosen?" or "Must the contrasts be orthogonal?" The brief answers are

"We don't know" and "No." Longer answers follow.

In the Word Association and APA voting examples in Sections 7.4 and 7.5, as well as the Goldberg example ($m = 10$) as treated in Marden [24], the set of orthogonal contrasts was chosen by trying several sets and taking the one whose fit was most pleasing. For the Word example and for Goldberg data, we judged goodness by the strength of the likelihood ratio test statistic for testing uniformity versus the model. In other words, we treated the contrasts $\underline{C}$ in (4) as a parameter along with $\underline{\theta}$, and tried to maximize the likelihood with respect to both. For the APA voting data, we tried to minimize higher-order interactions among the contrasts.

The number of possible sets of contrasts rises rapidly with m, which makes an exhaustive search for the best $\underline{C}$ difficult. The number of sets of size $T = m - 1$ seems to grow a bit faster than $m!$, with 105 sets for $m = 5$ and 34.5 million sets for $m = 10$. Since we did not do an exhaustive search, we may not have hit the absolute maximum. An open problem is how to efficiently search through the sets of contrasts to find the best. This is a generalization of the problem of searching for the modal ranking in the Mallows' ϕ model, for which Critchlow [10] has an algorithm. In the examples we have attempted, this problem does not appear too serious. By using other analyses, such as the spectral decompositions in Diaconis [14, 15] or data analytic techniques as in Cohen [6], one can obtain reasonable guesses about the structure of the optimal $\underline{C}$. In fact, Section 7.4 shows that just calculating the average rank for each object can be quite informative.

Besides the computational concern, there is the question of statistical validity. The asymptotic results on $\underline{\theta}$ for these models assume $\underline{C}$ is known, but including $\underline{C}$ as a parameter changes the problem. Critchlow ([10], Chapter 6, Section 5) considers this problem for Mallows' ϕ model with partial rankings. He shows that since the maximum likelihood estimator of the modal ranking is consistent, the asymptotic chi-squared nature of the goodness-of-fit statistic holds as it would if the modal ranking were known. A similar result should hold for model (7) under appropriate conditions. The main requirement for consistent estimation of $\underline{C}$ would be that the model be identifiable, which will not hold unless there were as many distinct values among the true θ_i's as the assumed number T of contrasts.

In other examples, one may have *a priori* contrasts of interest. In the ANOVA example of Section 7.6, the computational problem is minor since there are only six relevant sets of contrasts to consider. We did not base our choice on the Λ, but rather on its interpretability given that factor C was the main one of interest. (It turned out that for the Novices, the model did maximize Λ.) In any case, it was easy enough to look at all six, and they all gave similar results. Croon [12, 13] presents results from an international survey in which people were asked to rank a set of political goals. Goals are categorized as "materialist" and "post-materialist," and within materialist there are goals concerned with social stability, and others concerned with economic stability. Thus at least three contrasts are suggested by context. Böckenholt [3] has an example in which people were asked to rank eight soft drinks as in our illustration in Sections 7.2 and 7.3. An interesting question is whether people's choices can be described by a set of orthogonal contrasts. For example, do people first divide the drinks into cola and noncola, or into diet and nondiet?

Our revised answer to the first question has to be conditioned on the situation. If there is a small number of sets of contrasts we know are of interest, all the models can be fit and inspected. If the number of sets is too large, one may have to engage in an extensive search, systematic or otherwise. Prior informal inspection of the data can provide good starting points.

The model (4) follows in a long tradition of exponential family models for ranked data. A number of interesting statistics are defined for a sample rank vector, and the model uses these as the natural sufficient statistics. Models include those based on paired comparisons, and on metrics on $\mathbf{P}_m$. See Critchlow [10], Critchlow, Fligner and Verducci [11], and Diaconis ([14], Chapter 9) for reviews of many such approaches. McCullagh [25] presents an additional model based on inversions. There is no reason why one cannot create an exponential family based on a set of contrasts that are not orthogonal. In fact, several models do consist of nonorthogonal contrasts. Babington Smith [1] takes the $m(m-1)/2$ pairwise contrasts of elements, $C_{ij} = (o_i, o_j)$. The first order inversions of McCullagh [25] are also these contrasts. A special case of the Babington Smith model given in Mallows [22] takes as sufficient statistic the rank of each object, so that in a sample the average ranks of the objects are sufficient. Here, the contrasts are $C_i = (0_i\{0_1, \ldots, 0_m\}, \ldots, \} - \{0_i\})$ for $i = 1, ..., m$ since $d(C_i(y)) = y_i - 1$. (In fact, Mallows *phi* model and the orthogonal contrast *phi* models are also special cases of Babington Smith.) The ANOVA situation of Section 7.6 can easily be formulated in terms of nonorthogonal contrasts. The main effect contrast for each variable would be defined as in Contrast 1 in Section 7.6, the two-way interaction contrasts would pit the ranks in the 0-0 and 1-1 cells versus those in the 0-1 and 1-0 cells, etc.

Why, then, restrict orthogonal contrasts? There are computational and conceptual reasons. For small values of m, say $m \leq 5$, general linear methodology can be effectively used fit any well-parametrized exponential family model. When m is even moderately large, say $m \geq 10$, the computations are both time-consuming as well as susceptible unacceptable accumulations of round-off error since the number of cells exceeds 3.6 million. At which point between five and ten the process breaks down is debatable, but it is not unusual to have $m \geq 10$, e.g., the Goldberg data has $m = 10$, the ANOVA data has $m = 12$, and the Draft Lottery example in Fienberg [17] has $m = 366$. Fortunately, when using orthogonal contrasts in (7) the likelihood computation breaks into T independent pieces, each using a sum of at most m fairly simple functions as in (5). Thus computations grow at the rate of m instead of $m!$. Further complications arise if ties are incorporated into the model with nonorthogonal contrasts, since the nice exponential family structure may be destroyed. (It may not, depending on how the ties are incorporated.) But with orthogonal contrasts, there is no essential increase in complexity. Additionally, although a few well-chosen nonorthogonal contrasts may indeed produce a better model than one with orthogonal contrasts, an unrestricted search through arbitrary sets of nonorthogonal contrasts quickly becomes a nightmare. The number of such sets is on the order of $2^{(3^m/2)}$, so that even $m = 6$ pushes us beyond 10^{100}. Restricting sets of k only reduces the rate $3^m/2$ choose k.

Orthogonal contrasts are easier handle conceptually here for the same reason that orthogonality is preferred in experimental designs and normal linear models: The interpretation of each individual contrast does not depend on the values of the other contrasts. It also follows that the individual parameters can be estimated more precisely. In our ranking case, when leaving the ϕ model, the data falls into a nice contingency table in which independence, conditional independence, multi-way interaction, etc., are of real interest. Thus standard methods yield a straightforward data-analytic framework. With nonorthogonal contrasts, independence is not a natural circumstance of interest.

To answer the second question, there are certainly many useful models that do not have orthogonality, but the purpose of this paper is convince the reader that modeling with orthogonal contrasts is an approach dealing with rank data that is flexible, easy interpret, and computationally accessible.

7.8 Appendix

Table 18 contains the fully ranked APA voting data arranged in the four-way table as in Section7.5. Contrast 1 gives the rank of candidate B, Contrast 2 gives the unordered ranks of A and C among (A,C,D,E), and Contrasts 3 and 4 compare A to C and D to E, respectively, the value "1" ("2") meaning the first (second) is preferred.

Table 18: APA Voting data, Full Rankings

Entries are numbers of voters

		Contrast 2					
$C_3 = 1, C_4 = 1$	$C_1 \downarrow$	12	13	14	23	24	34
	1	40	26	42	42	34	45
	2	30	27	35	75	64	66
	3	102	35	28	52	52	84
	4	45	24	48	28	53	133
	5	50	17	35	21	34	61
$C_3 = 2, C_4 = 1$							
	1	34	16	29	23	19	46
	2	74	40	50	51	24	44
	3	172	35	26	35	22	49
	4	96	35	28	28	24	67
	5	79	36	30	27	24	54
$C_3 = 1, C_4 = 2$							
	1	30	24	36	40	30	50
	2	28	29	34	34	41	58
	3	95	37	35	53	45	91
	4	70	51	52	44	63	107
	5	70	36	40	40	35	71
$C_3 = 2, C_4 = 2$							
	1	30	22	11	31	25	50
	2	82	52	35	24	26	54
	3	186	38	34	30	22	57
	4	162	87	43	62	37	29
	5	106	45	38	41	34	31

The raw data from the computer searches, Section 7.6, is presented in Table 19. See Table 14 for more detail on the searches.

Table 19: Search Data: Numbers of Attempts

Search	Novices							Experts						
1	2	2	3	2	3	1	1	3	2	1	1	1	3	1
2	2	2	2	1	1	1	1	3	1	1	1	1	3	2
3	3	4	3	3	2	2	2	2	2	2	1	1	2	3
4	2	2	4	3	3	2	2	4	2	1	2	2	3	2
5	2	2	3	3	3	2	3	3	3	2	3	2	3	3
6	3	2	3	3	2	1	2	2	2	2	2	2	2	3
7	3	3	2	3	3	2	1	3	1	1	1	1	1	2
8	1	1	2	2	1	1	1	1	1	1	1	1	1	1
9	1	1	2	3	1	2	1	1	1	1	1	1	1	1
10	1	2	1	1	1	1	2	1	1	2	2	1	1	1
11	2	2	1	1	1	1	1	1	3	1	1	1	2	1
12	1	3	3	2	2	2	1	2	2	2	1	1	1	1

Acknowledgements:

Partial support has been provided by grants AFOSR-87-0041 and NSF-DMS-88-02556.

7.9 References

[1] Babington Smith, B. Discussion of Professor Ross's paper. *J. Royal Statist. Soc.* B, **12**:53-56, 1950.

[2] Barton, D. E., David, F. N. and Mallows, C. L. "Non-randomness in a sequence of two alternatives: I. Wilcoxon and allied test statistics," *Biometrika* **45**:166-177, 1958.

[3] Böckenholt, U. "Thurstonian models for ranking data," *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[4] Chapman, L., Girard, A., Gravelle, M., Lewis, P. and Ricker, A. "Cognitive processes in expert and novice LCS searchers," Class project, University of Illinois at Urbana-Champaign, 1989.

[5] Chung, L. and Marden, J. I. "Use of nonnull models for rank statistics in bivariate, two-sample, and analysis-of-variance problems," To appear in *J. Amer. Statist. Assoc.* 1991.

[6] Cohen, A. "Analysis of large sets of ranking data," *Communications in Statistics - Theory and Methods* **11**:235-256, 1982.

[7] Cohen, A. "Data analysis of ranking data," *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[8] Cohen, A. and Mallows, C. L. "Analysis of ranking data," Bell Laboratories Memorandum, 1980.

[9] Cohen, A. and Mallows, C. L. "Assessing goodness of fit of ranking models to data," *The Statistician*, **32**:361-373 1983.

[10] Critchlow, D. *Metric Methods for Analyzing Partially Ranked Data*, Springer-Verlag, New York, 1985.

[11] Critchlow, D., Fligner, M. and Verducci, J. "Probability models on rankings," *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[12] Croon, M. A. "Latent class models for the analysis of rankings," in *New Developments in Psychological Choice Modeling.* Feger, Klauer and de Soete, editors, North-Holland, pp. 99-121, 1989.

[13] Croon, M. A. "Latent class models," *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[14] Diaconis, P. *Group Representations in Probability and Statistics* **11**, Institute of Mathematical Statistics Lecture Notes-Monograph, Series 1988.

[15] Diaconis, P. "A generalization of spectral analysis with application to ranked data," *Annals of Statistics*, **17**:949-979, 1989.

[16] Feigin, P. and Cohen, A. "On a model of concordance between judges," *Journal of the Royal Statistical Association* B, **40**:203-213, 1978.

[17] Fienberg, S. "Randomization and social affairs: The 1970 draft lottery," *Science* **171**:255-261, 1971.

[18] Fligner, M. A., and Verducci, J. S. "Distance based ranking models," *Journal of the Royal Statistical Association* B, **48**:359-369, 1986.

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

[20] Goldberg, A. I. The relevance of cosmopolitan/local orientations to professional values and behavior. *Sociology of Work and Occupation* **3**:331-356, 1976.

[21] Hare, T. *The Election of Representatives, Parliamentary and Municipal: A Treatise*, 3rd Edition, Logman, Roberts and Green, London, 1865.

[22] Mallows, C. L. "Non-null Ranking Models: I," *Biometrika*, **44**:114-130, 1957.

[23] Mann, H. B. "Non-parametric tests against trends," *Econometrica*, **13**:245-259, 1945.

[24] Marden, J. I. "Use of orthogonal contrasts in analyzing rank data," Technical Report, Department of Statistics, University of Illinois at Urbana-Champaign, 1990.

[25] McCullagh, P. "Permutations and regression models," *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[26] Schulman, R. R. "Ordinal Data: An alternative distribution," *Psychometrika*, **44**:3-20, 1979.

[27] Silverberg, A. R. "Statistical models for q-permutations," *Proceedings of the Biopharmaceutical Section, American Statistical Association*, 107-112, 1984.

[28] Smith, L. "A censored ranking problem," *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[29] Stern, H. "Probability models on rankings and the electoral process, *Proceedings of the Joint Summer Research Conference on Probability Models and Statistical Analysis for Ranking Data*, 1991.

[30] Thompson, G. L. "A note on the rank transform for interaction," *Biometrika*, in press, 1991.

[31] van Blokland-Vogelesang, R. A. W. "Unfolding and consensus ranking: A prestige ladder for technical occupations," in *New Developments in Psychological Choice Modeling.* Feger, Klauer and de Soete, editors, North-Holland, pp. 237-258, 1989.

8

Rank Correlations and the Analysis of Rank-Based Experimental Designs

M. Alvo[1]
P. Cabilio[2]

ABSTRACT The notion of distance between two permutations is used to provide a unified treatment for various problems involving ranking data. Using the distances defined by Spearman and Kendall, the approach is illustrated in terms of the problem of concordance as well as the problem of testing for agreement among two or more populations of rankers. An extension of the notion of distance for incomplete permutations is shown to lead to a generalization of the notion of rank correlation. Applications are given to the incomplete block design as well as to the class of cyclic designs.

8.1 Introduction

The subject of rank correlation has had a rich and extensive history. By viewing the rank correlation between two rankings in terms of distance functions, it is possible to define different measures of correlation, which include as special cases those defined by Spearman and Kendall [17]. The connection between average Spearman rank correlations and the Friedman test statistic for the problem of m rankings was noted by Kendall [17], while Ehrenberg [9] introduced average Kendall rank correlations for this problem. Durbin [8] extended the Friedman statistic to the balanced incomplete block design.

The distance based approach offered herein provides a unified treatment of such tests and allows for a general solution to the problem of m rankings, as well as that of testing for agreement between two or more populations of rankers. Many of these results, described in Sections 8.3 and 8.4, have been dealt with in greater detail in Alvo, Cabilio and Feigin [1], Alvo and Cabilio

[1] University of Ottawa, Ontario, Canada
[2] Acadia University, Nova Scotia, Canada

[2] and Feigin and Alvo [10]. In Section 8.5, an extension of the notion of distance applied to sets of permutations permits a generalization of the problem of m rankings to the case of balanced incomplete block designs, and allows for a new interpretation of the Durbin statistic. A detailed analysis of this problem appears in Alvo and Cabilio [4].

The definition of distance measures for incomplete rankings allows us to present some new results in the concluding sections. In Section 8.6 the Durbin test is extended to the more general situation of cyclic designs. In Section 8.7, measures of correlation between incomplete rankings are introduced and are related to the coefficient of concordance in a way that is parallel to the complete ranking situation.

8.2 Distance Based Measures of Correlation

Let $\mathcal{P} = \{\nu_j\}$ be the space of all possible permutations of the integers $1, 2, \ldots, t$ and let the column vectors

$$\nu_j = (\nu_j(1), \ldots, \nu_j(t))', \quad j = 1, 2, \ldots, t!$$

denote the $t!$ possible permutations. Thus $\mathcal{P}$ represents the collection of all possible rankings of t objects. For convenience these objects are ordered in some way and labelled $1, 2, \ldots, t$. A measure of the correlation between permutations μ and ν may be defined in terms of the distance $d(\mu, \nu)$ between them as:

$$\alpha(\mu, \nu) = 1 - \frac{2d(\mu, \nu)}{M} \tag{1}$$

where M is the maximum value of d taken over all possible pairs μ and ν in $\mathcal{P}$. Examples of distances over permutations may be found in Critchlow [6]. These include the distances associated with Spearman and Kendall (see Kendall [17]):

$$d_S(\mu, \nu) = \frac{1}{2}\sum_{i=1}^{t}[\mu(i) - \nu(i)]^2 \tag{2}$$

$$d_K(\mu, \nu) = \sum_{i<j}\{1 - sgn[\mu(i) - \mu(j)]sgn[\nu(i) - \nu(j)]\} \tag{3}$$

These distances have the property that the distance between two rankings remains unchanged under any permutation relabeling of the objects. That is, for any $\sigma, \mu, \nu \in \mathcal{P}$, $d(\mu, \nu) = d(\mu\sigma, \nu\sigma)$. This property is known as right invariance.

Let $\Delta = (d(\nu_i, \nu_j))$ denote the matrix of distances. If d is right invariant, then it follows that there exists a constant $c > 0$ for which

$$\Delta e = (ct!)e \tag{4}$$

where $e = (1, 1, \ldots, 1)'$ is of dimension $t!$ (see Feigin and Alvo [10]). Hence, c represents the average distance between pairs of permutations. For the Spearman and Kendall metrics, we have

$$c_S = \frac{t(t^2-1)}{12} \quad , \quad M_S = 2c_S \tag{5}$$

$$c_K = \frac{t(t-1)}{2} \quad , \quad M_K = 2c_K \tag{6}$$

(see Kendall [17]).
The correlation coefficients based on these distances are of the multiplicative type in the sense of Kendall [17]; that is, there exists a function g such that

$$\alpha(\mu, \nu) = k_\mu k_\nu \sum_{i=1}^{t} \sum_{j=1}^{t} g(\mu(i), \mu(j)) g(\nu(i), \nu(j)),$$

where k_μ, k_ν are normalizing constants. Kendall [17] distinguishes such coefficients into type a and type b depending on the choice of normalizing constants. For the distances in (2) and (3), the functions g are respectively

$$g_S(\mu(i), \mu(j)) = \mu(i) - \mu(j) \tag{7}$$

$$g_K(\mu(i), \mu(j)) = sgn[\mu(i) - \mu(j)]. \tag{8}$$

For a multiplicative index, the matrix $\Gamma = (\alpha(\nu_i, \nu_j))$ is necessarily positive semi-definite (see Quade [21]). Setting $Q \equiv (\frac{M}{2}J - \Delta)$, where $J = ee'$, this implies there exists a matrix T for which

$$Q = T'T. \tag{9}$$

In view of (4), it follows that for a right invariant metric

$$Qe = t!(\frac{M}{2} - c)e \tag{10}$$

Moreover, for the Spearman and Kendall distances, the relationships in (5) and (6) imply that $Q_S e = 0 = Q_K e$.

The matrices T corresponding to d_S and d_K are respectively:

$$T_S = (t_S(\nu_1), \ldots, t_S(\nu_{t!})) \tag{11}$$

of dimension $(t \times t!)$ where

$$t_S(\nu) = (\nu(1) - \frac{t+1}{2}, \ldots, \nu(t) - \frac{t+1}{2})' \tag{12}$$

the centered rank vector, and

$$T_K = (t_K(\nu_1), \ldots, t_K(\nu_{t!})) \tag{13}$$

of dimension $(\binom{t}{2} \times t!)$ where the qth element of $t_K(\nu)$ is

$$sgn[\nu(j) - \nu(i)] \tag{14}$$

$$q = (i-1)(t - \frac{i}{2}) + (j - i) \qquad 1 \leq i < j \leq t. \tag{15}$$

Here, $t_K(\nu)$ is the vector of pairwise concordances or discordances of the ranking ν with the identity rank permutation $(1, \ldots, t)'$ (see Feigin and Alvo [10]).

The notion of correlation between two permutations has previously been used in nonparametric tests of trend and of independence (see Randles and Wolfe [22]). In that context, it can be shown that the null distributions of α_S and α_K properly standardized are asymptotically normal as $t \longrightarrow \infty$. (see Kendall and Stuart [[19] p. 507] and Jirina [15]). Daniels [7] showed that the limiting joint distribution of α_S and α_K is bivariate normal whereas Hájek and Šidák [12] noted that, up to a factor, the Spearman correlation coefficient may be viewed as the projection of the Kendall coefficient into the family of linear rank statistics. As $t \longrightarrow \infty$, the two coefficients have a correlation which tends to 1 and hence are asymptotically equivalent.

8.3 The Problem of m Rankings

Suppose that m judges acting independently provide rankings $X_1, X_2, \ldots, X_m$ of t objects, each chosen according to a distribution $\pi = (\pi_1, \ldots \pi_{t!})'$ over $\mathcal{P}$; that is,

$$\pi_i = P(X = \nu_i) \qquad i = 1, \ldots, t!$$

The problem of m rankings consists of testing the null hypothesis H_0 : $\pi = \pi^0 \equiv (t!)^{-1}e$ against the alternative $H_1 : \pi \neq \pi^0$. This problem was first considered by Friedman [11] and later by Kendall and Babington Smith [18]. Friedman's result which may be presented in the context of the average pairwise correlation

$$\bar{\alpha} = \sum_{i<j} \alpha(X_i, X_j) / \binom{m}{2} \tag{16}$$

is that under H_0, as $m \longrightarrow \infty$.

$$(t-1)[(m-1)\bar{\alpha}_S + 1] \overset{\mathcal{L}}{\Longrightarrow} \chi^2_{t-1} \tag{17}$$

The test rejects H_0 for large values of $\bar{\alpha}_S$.

Noting that $\bar{\alpha}_S$, apart from some factors, can be expressed as a quadratic form in a multinomial $t!$ vector with parameters m and π, Alvo, Cabilio and Feigin [1] provide a proof of Friedman's result using the multivariate form of the central limit theorem (see also Quade [21]). Specifically, let

$f = (f_1, \ldots, f_{t!})'$ be the vector of frequencies of rankings and let $\hat{\pi} = f/m$. Defining $Z_m = m^{1/2}(f - m\pi^0)$, and using (10) and (5) one may write

$$\begin{aligned}(m-1)\bar{\alpha}_S + 1 &= \frac{1}{c_S}[Z_m' Q Z_m] \\ &= \frac{m}{c_S}[T_S(\hat{\pi} - \pi^0)]'[T_S(\hat{\pi} - \pi^0)]. \qquad (18)\end{aligned}$$

The form (18) may be simplified on noting that $T_S e = 0$, so that

$$(m-1)\bar{\alpha}_S + 1 = \frac{m}{c_S}(T_S\hat{\pi})'(T_S\hat{\pi}) \qquad (19)$$

From (11) and (12), it follows that the right side of (19) multiplied by $(t-1)$ is equal to the familiar form of the Friedman statistic,

$$\frac{12}{mt(t+1)} \sum_{q=1}^{t} [R_q - m(\frac{t+1}{2})]^2 \qquad (20)$$

where R_q is the sum of the ranks assigned to the qth object.

The representation in (19) provides a different interpretation for testing H_0 in terms of the characteristics defined by the matrix T_S and hence by the distance d_S. This opens the possibility of generalizing the approach. In particular, the test statistic associated with the Kendall metric would, in view of (9), have the form:

$$\begin{aligned}(m-1)\bar{\alpha}_K + 1 &= \frac{m}{c_K}[T_K(\hat{\pi} - \pi^0)]'[T_K(\hat{\pi} - \pi^0)]. \\ &= \frac{m}{c_K}(T_K\hat{\pi})'(T_K\hat{\pi}), \qquad (21)\end{aligned}$$

since $T_K e = 0$.

It has been shown by Alvo, Cabilio and Feigin [1] that under H_0, the asymptotic distribution ($m \to \infty$) of $m\bar{\alpha}_K + 1$ is given by

$$\frac{2}{3t(t-1)}\{(t+1)\chi^2_{t-1} + \chi^2_{\binom{t-1}{2}}\} \qquad (22)$$

where the two χ^2 variates are independent. The distribution in (22) is dominated by the first variate for all values of t. The statistics (19) and (21) may be viewed as sampling estimates of specific functions of the parameters $T_S\pi$, the vector of expected centered ranks, and $T_K\pi$, the vector of expected pairwise concordances. These parameters or characteristics thus replace the original parameter π, effecting a large decrease of dimensionality. Ehrenberg [9] was the first to suggest the use of average Kendall correlations as an alternative to $\bar{\alpha}_S$. He justified his preference for $\bar{\alpha}_K$ by arguing that pairwise concordances offer a deeper measure of agreement than that to be had from simply using the sum of the ranks. In this connection, it should be pointed out that if $t = m$, and the matrix of rankings forms a circulant

$$\begin{array}{cccccc} 1 & 2 & 3 & \cdots & m \\ m & 1 & 2 & \cdots & m-1 \\ m-1 & m & 1 & \cdots & m-2 \\ \vdots & \vdots & \vdots & & \vdots \\ 2 & 3 & 4 & \cdots & 1 \end{array}$$

then, with H_1 true,

$$\bar{\alpha}_S = -\frac{1}{m-1} \longrightarrow 0 \qquad \text{as} \qquad m \longrightarrow \infty$$

whereas

$$\bar{\alpha}_K = 1 - \frac{4}{m} \longrightarrow 1 \qquad \text{as} \qquad m \longrightarrow \infty$$

(see Ehrenberg [9]). Hence, under this alternative to H_0, the test based on $\bar{\alpha}_S$ would not reject. A similar difficulty arises when one half of the observed rankings are in natural order and the rest are in reverse natural order.

Tables for the exact null distribution of $\bar{\alpha}_K$ for small values of t and m can be found in Quade [21] and Alvo and Cabilio [2]. The latter provide comparisons of different approximations to the null distribution of $\bar{\alpha}_K$ given by Ehrenberg [9], Hays [13] and (22).

A different set of asymptotics arises if m is fixed and $t \longrightarrow \infty$ as the following theorem shows.

Theorem 8.3.1 Under H_0, as $t \longrightarrow \infty$,

$$a) \qquad \frac{\bar{\alpha}_K}{\sqrt{Var(\bar{\alpha}_K)}} \xrightarrow{\mathcal{L}} N(0,1)$$

and

$$b) \qquad \frac{\bar{\alpha}_S}{\sqrt{Var(\bar{\alpha}_S)}} \xrightarrow{\mathcal{L}} N(0,1)$$

where

$$Var(\bar{\alpha}_K) = [\binom{t}{2}\binom{m}{2}]^{-1}(\frac{2t+5}{9})$$

$$Var(\bar{\alpha}_S) = \binom{m}{2}^{-1}(\frac{1}{t-1})$$

Proof: First, it can be seen that if either result a) or b) holds, then the other must follow. Indeed, recall that

$$Var(\bar{\alpha}_K) = [\binom{m}{2}\binom{t}{2}]^{-1}(\frac{2t+5}{9}) \quad \text{and} \quad Var(\bar{\alpha}_S) = \binom{m}{2}^{-1}\frac{1}{(t-1)}.$$

Using a result from Hájek and Sǐdák ([12], p. 61), it can be shown that

$$E[\bar{\alpha}_K\bar{\alpha}_S] = \binom{m}{2}^{-1}\binom{t}{2}^{-1}\frac{(t+1)}{3};$$

consequently, as $t \longrightarrow \infty$,

$$E\left[\frac{\bar{\alpha}_K}{\sqrt{Var(\bar{\alpha}_K)}} - \frac{\bar{\alpha}_S}{\sqrt{Var(\bar{\alpha}_S)}}\right]^2 = 2 - 2E\frac{\bar{\alpha}_K}{\sqrt{Var(\bar{\alpha}_K)}}\frac{\bar{\alpha}_S}{\sqrt{Var(\bar{\alpha}_S)}} \longrightarrow 0$$

In the case where the rankings arise from the observation of t independent random vectors $X_i = [X_i^{(1)}, \ldots, X_i^{(m)}]'$, with independent components and continuous distributions, an application of the multivariate central limit Theorem 8.41 of Puri and Sen [20], leads to a demonstration of b). A more direct proof of a) for this case makes use of the central limit theorem for U-statistics.

In view of Theorem 8.3.1 as well as the comparisons described in Alvo and Cabilio [2], the asymptotic distributions provided in (17) and (22) are inappropriate when m, the number of rankers is not large compared to the number of objects t.

8.4 The Two Sample Problem

Consider now the problem of testing for agreement between two populations of rankers acting independently of one another. Letting $\pi^{(1)}, \pi^{(2)}$ be the two distributions over $\mathcal{P}$ associated with the populations, the two sample problem may be viewed as a test of the hypothesis $\pi^{(1)} = \pi^{(2)}$ against the alternative $\pi^{(1)} \neq \pi^{(2)}$. Schucany and Frawley [24] and subsequently Hollander and Sethuraman [14] suggested different interpretations of the notion of agreement. Feigin and Alvo [10], motivated by Rao's concept of diversity [23], proposed testing $T\pi^{(1)} = T\pi^{(2)}$ against the alternative $T\pi^{(1)} \neq T\pi^{(2)}$ where as before the matrix T serves to define certain population characteristics of interest and to reduce the dimensionality of π. Feigin and Alvo then proceeded by using standard multivariate methods. Their approach allows for a generalization to several populations and permits consideration of several problems in the analysis of variance when ranking data is involved.

8.5 The Problem of m Rankings for a Balanced Incomplete Block Design

A generalization of the problem of m rankings was first considered by Durbin [8]. A total of mb judges are presented $k < t$ objects to rank. The pattern of the objects presented follows m replications of a balanced incomplete block design of b blocks of k rankings of t objects. Within each basic design, every object is considered by r of b judges and each pair of objects is presented together to λ of these judges. For a balanced incomplete block design,

$$bk = tr \tag{23}$$

$$\lambda = r(k-1)/(t-1) \tag{24}$$

(see Cochran and Cox [5]).
Each block represents a different pattern of k objects to be ranked. Let

$$\nu_j^{(\ell)} = (\nu_j^{(\ell)}(1), \dots, \nu_j^{(\ell)}(k))', \quad j = 1, \dots, k!$$

represent the $k!$ possible permutations of the integers $(1, \dots, k)$ corresponding to all the possible k-partial rankings for each of the block patterns indexed by $\ell = 1, \dots, b$. A judge presented with k objects according to block pattern ℓ selects a ranking from $\{\nu_j^{(\ell)}\}$ according to the probability vector

$$\pi^{(\ell)} = (\pi_1^{(\ell)}, \dots, \pi_{k!}^{(\ell)})'.$$

The $(bk! \times 1)$ vector of probability vectors for the overall design is

$$\pi^* = (\pi_1^{(1)}, \dots, \pi_{k!}^{(1)}\Big|\dots\Big|\pi_1^{(b)}, \dots, \pi_{k!}^{(b)})'.$$

Setting

$$\pi_0^* = (k!)^{-1}(1, \dots, 1\Big|\dots\Big|1, \dots, 1)',$$

the null hypothesis to be tested is $H_0^* : \pi^* = \pi_0^*$ against the alternative $H_1^* : \pi^* \neq \pi_0^*$.

In order to clarify the notation, suppose $t = 3$ objects are presented $k = 2$ at a time. The complete rankings are labelled

$$\nu_1 = (123)', \nu_2 = (132)', \nu_3 = (213)', \nu_4 = (231)', \nu_5 = (312)', \nu_6 = (321)'.$$

The incomplete rankings are denoted by

$$\nu_1^{(1)} = (12\_)', \nu_2^{(1)} = (21\_)'; \quad \nu_1^{(2)} = (1\_2)', \nu_2^{(2)} = (2\_1)';$$

$$\nu_1^{(3)} = (\_12)', \quad \nu_2^{(3)} = (\_21)'$$

The following notion of compatibility between a complete and an incomplete ranking plays an important role in the sequel.

Definition 8.5.1 A complete ranking ν is said to be compatible with an incomplete ranking $\nu^{(*)}$ if the relative ranking of every pair of objects ranked in $\nu^{(*)}$ coincides with its relative ranking in ν.

Hence, the complete rankings ν_1, ν_2 and ν_4 above are compatible with the incomplete ranking $\nu_1^{(1)}$. In general, the number of complete rankings compatible with a specific incomplete ranking in a balanced design will be

$$a = t!/k!$$

The definition of compatibility is useful in extending the notion of distance to incomplete rankings.

Definition 8.5.2 The distance between the incomplete rankings $\nu_i^{(\ell)}$ and $\nu_j^{(p)}$, denoted by $d^*(\nu_i^{(\ell)}, \nu_j^{(p)})$ is defined to be the average of all values of $d(\mu, \eta)$ taken over all complete rankings μ, η compatible with $\nu_i^{(\ell)}$ and $\nu_j^{(p)}$ respectively.

Although the distance given in Definition 8.5.2 maintains the triangle inequality, it is not a metric since the distance between a ranking and its duplicate will not be zero. Another way of defining distances between incomplete rankings may be based on Hausdorff metrics as detailed in Critchlow [6] for the case of partial rankings. On the other hand, our approach could be extented to the situations studied by Critchlow. One justification for our approach in this case is that a gain in mathematical tractability offsets, in our opinion, the loss of the least important property of a metric. In any case, our approach leads to results which nicely parallel the complete case.

Ordering the complete rankings $\{\nu_j\}$ in some way, we may associate with every incomplete ranking $\nu^{(*)}$ a $(t! \times 1)$ compatibility vector, whose i^{th} component is 1 or 0 according to whether ν_i is compatible with $\nu^{(*)}$ or not. In this way, a matrix of compatibiity C_ℓ may be generated for block pattern ℓ whose j^{th} column is the compatibility vector associated with $\nu_j^{(\ell)}, j = 1, \ldots, k!$.

Finally, the overall compatibility matrix for the design is defined to be

$$C = (C_1 | C_2 | \ldots | C_b). \tag{25}$$

As an example for $t = 3, k = 2$, the compatibility matrix C is found to be

$$C = \begin{bmatrix} 10 & | & 10 & | & 10 \\ 10 & | & 10 & | & 01 \\ 01 & | & 10 & | & 10 \\ 10 & | & 01 & | & 01 \\ 01 & | & 01 & | & 10 \\ 01 & | & 01 & | & 01 \end{bmatrix}.$$

In matrix notation, if $\Delta^* = (d^*(\nu_i^{(\ell)}, \nu_j^{(p)}))$ represents the matrix of distances between pairs of incomplete rankings, it follows that

$$\Delta^* = a^{-2} C' \Delta C \tag{26}$$

Restricting attention to the Spearman and Kendall metrics, the representation in (9) leads to

$$\Delta^* = \frac{M}{2} a^{-2} C' J C - T^{*'} T^* \tag{27}$$

where $T^* = a^{-1} TC$. Consequently, in analogy with (19) and (21), the test statistic for testing H_0^* will be based on the sampling estimate of $(T^* \pi^*)'(T^* \pi^*)$.

The following theorem is proved in Alvo and Cabilio [4]. Let $f^{(\ell)} = (f_1^{(\ell)}, \ldots, f_{k!}^{(\ell)})'$ be the vector of frequencies of rankings for block pattern ℓ and set the $(bk! \times 1)$ vector

$$f = (f_1^{(1)}, \ldots, f_{k!}^{(1)} | \ldots | f_1^{(b)}, \ldots, f_{k!}^{(b)})',$$

Theorem 8.5.1 Under H_0^* as $m \longrightarrow \infty$,

$$(a) \qquad m^{-1} G_S \equiv (a\sqrt{m})^{-2} (T_S C f)'(T_S C f) \overset{\mathcal{L}}{\Longrightarrow} \alpha_1 \chi^2_{t-1} \tag{28}$$

$$(b) \qquad m^{-1} G_K \equiv (a\sqrt{m})^{-2} (T_K C f)'(T_K C f) \overset{\mathcal{L}}{\Longrightarrow} \alpha_2 \chi^2_{t-1} + \alpha_3 \chi^2_{\binom{t-1}{2}} \tag{29}$$

where

$$\alpha_1 = \frac{\lambda t (t+1)^2}{12(k+1)}, \quad \alpha_2 = \frac{\lambda (t+1)^2}{3(k+1)}, \quad \alpha_3 = \frac{\lambda [t(k-1)-2]}{3(t-2)(k+1)} \tag{30}$$

and the χ^2 variates in (29) are independent.

The result in (28) coincides with Durbin's [8]. In fact, it can be shown that the left side of (28) divided by α_1 is equal to

$$D = \frac{12}{m \lambda t (k+1)} \sum_{q=1}^{t} [R_q^* - mr(\frac{k+1}{2})]^2 \tag{31}$$

where R_q^* is the sum of the ranks assigned to the qth object. On the other hand, (29) provides a generalization to (22) in the case of the Kendall metric. Note that (31) becomes equal to (20) in the complete block situation when $k = t, b = r = 1$ and $\lambda = 1$.

A computational formula can be developed in order to facilitate the use of the Kendall metric (Alvo and Cabilio [4]). Specifically, for each pair of objects labelled $(q_1, q_2), q_1 < q_2$, the incomplete ranking μ_j^* of judge j is assigned a score $a_j(q_1, q_2)$ where

$$a_j(q_1, q_2) = \begin{cases} sgn(\mu_j^*(q_2) - \mu_j^*(q_1) & \text{if judge } j \text{ ranks both } q_1 \text{ and } q_2 \\ 1 - 2\frac{\mu_j^*(q_1)}{k+1} & \text{if judge } j \text{ ranks only } q_1 \\ \frac{2\mu_j^*(q_2)}{k+1} - 1 & \text{if judge } j \text{ ranks only } q_2 \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

The sum over all rankings of the scores for the pair (q_1, q_2) is

$$\frac{1}{a}(T_K Cf)_q = \sum_{j=1}^{mb} a_j(q_1, q_2) \tag{33}$$

the qth element of $\frac{1}{a}(T_K Cf)$, where q is defined by (15) with $i = q_1, j = q_2$.

Thus

$$G_K = \sum_{q_1 < q_2} [\sum_{j=1}^{mb} a_j(q_1, q_2)]^2. \tag{34}$$

This form is reminiscent of the one derived by Hays [13] for the complete ranking situation, and in fact reduces to it in that case.

8.6 The Problem of m Rankings for Cyclic Designs

In certain instances where the requirements of the balanced incomplete block design are too restrictive, it may be useful to consider cyclic designs instead. The properties of such designs are given in some detail in John [16]. As in the balanced incomplete block design the pattern is of m replications of a basic design of b blocks of k rankings of t objects, with each object presented to r of b judges. For such a design relation (23) continues to hold. A cyclic design is further characterized by a symmetric concurrence matrix (λ_{ij}) where λ_{ij} is the number of blocks in the design in which treatments i and j occur together. Unlike the balanced incomplete block design λ_{ij} is not a constant, but depends on $(j - i)$ and thus does not meet the requirements of (24). Specifically for a cyclic design

$$\lambda_{ij} = \begin{cases} \lambda_{j-i+1} & i \leq j - 1 \\ r & i = j \\ \lambda_{t+j-i+1} & i \geq j + 1 \end{cases}$$

for $1 \leq i \leq j \leq t$.

To illustrate, consider the following cyclic design with 4 treatments: $t = 4, k = 2, b = 4, r = 2$.

	Treatments			
	1	2	3	4
Blocks	*	*	—	—
	—	*	*	—
	—	—	*	*
	*	—	—	*

The concurrence matrix is given by

$$\begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}.$$

so that $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0, \lambda_4 = 1$.

In analogy to (28), it may be shown that for the C defined by this design

$$\begin{aligned}(a\sqrt{m})^{-2}(T_S Cf)'(T_S Cf) &= \frac{(t+1)^2}{m(k+1)^2}\sum_{q=1}^{t}[R_q^* - mr(\frac{k+1}{2})]^2 \\ &\xrightarrow{\mathcal{L}} \sum_{q=1}^{t-1} \alpha_i^* Z_i^2 \end{aligned} \tag{35}$$

where the $\{\alpha_i^*\}$ are the eigenvalues of $\frac{1}{a^2 k!}T_S CC'T_S'$ and $\{Z_i\}$ are independent standard normal variates. The principal change between this result and (28) is that the eigenvalues for the cyclic design are no longer all identical. This is a consequence of the fact that the cyclic design alters the balance in the compatibility matrix. What the actual eigenvalues are in general is unclear, however they may be calculated for specific situations.

In this example, $\alpha_1^* = \frac{25}{9}$ and $\alpha_2^* = \alpha_3^* = \frac{25}{18}$.

8.7 Measuring Correlation Between Incomplete Rankings

Motivated by the concepts of compatibility and average distance one may define the distance between μ_i^* and μ_j^*, the incomplete rankings of judges i and j respectively as

$$d^*(\mu_i^*, \mu_j^*) = c - \frac{1}{a^2}[C(\mu_i^*)]'T'T[C(\mu_j^*)] \tag{36}$$

where $C(\mu^*)$ is the compatibility vector of $\mu^*; c = c_S = \frac{t(t^2-1)}{12}$ in the Spearman case, and $c = c_K = \frac{t(t-1)}{2}$ in the Kendall case. Let ω_j be the vector formed by filling in the blanks of the incomplete ranking μ_j^* with the average rank $\frac{k+1}{2}$. That is

$$\omega_j(s) = \mu_j^*(s)\delta(s,j) + \left(\frac{k+1}{2}\right)(1 - \delta(s,j)) \tag{37}$$

where

$$\delta(s,j) = \begin{cases} 1 & \text{if judge } j \text{ ranks object } s \\ 0 & \text{otherwise} \end{cases}$$

Then a simplified form of the Spearman distance is

$$d_S^*(\mu_i^*, \mu_j^*) = \frac{t(t+1)(2t+1)}{6} - \left(\frac{t+1}{k+1}\right)^2 \sum_{s=1}^{t} \omega_i(s)\omega_j(s).$$

In the Kendall case the distance may be written as

$$d_K^*(\mu_i^*, \mu_j^*) = \frac{t(t-1)}{2} - \sum_{q_1<q_2} a_i(q_1,q_2)a_j(q_1,q_2),$$

where $a_j(q_1,q_2)$ is defined as in (32).

An application of the Cauchy-Schwarz inequality indicates that the upper bound of $[C(\mu_i^*)]'T'TC(\mu_j^*)$ is achieved when $C(\mu_i^*) = C(\mu_j^*)$, that is when $\mu_i^* = \mu_j^*$, and the lower bound is achieved when $TC(\mu_i^*) = -TC(\mu_j^*)$. If we let μ_j^* be the inverted ranking, that is $\mu_j^*(s) = k+1-\mu_i^*(s)$ when object s is ranked by i, then $\omega_j(s) = k+1-\omega_i(s)$ and $T_SC(\mu_j^*) = -T_SC(\mu_i^*)$. Further $a_j(q_1,q_2) = -a_i(q_1,q_2)$, and thus $T_KC(\mu_j^*) = -T_KC(\mu_i^*)$. A straightforward calculation of these distances using the incomplete ranking $(12\ldots k\_\ldots\_)'$ and its inversion shows the minimum (m^*) and the maximum (M^*) distances in the two cases to be

$$m_S^* = c_S - \frac{(t+1)^2}{12}\frac{k(k-1)}{(k+1)}, \quad M_S^* = c_S + \frac{(t+1)^2}{12}\frac{k(k-1)}{(k+1)}$$

$$m_K^* = c_K - \frac{(2t+k+3)}{6}\frac{k(k-1)}{(k+1)}, \quad M_K^* = c_K + \frac{(2t+k+3)}{6}\frac{k(k-1)}{(k+1)} \qquad (38)$$

Kendall [17] introduced the coefficient of concordance W, a measure of overall agreement amongst the judges in the complete ranking situation described in Section 8.3. This coefficient W may be written as

$$W = \frac{1}{m}(\bar{\alpha}_S(m-1)+1) \qquad (39)$$

and clearly achieves the maximum value of 1 when all judges agree. Durbin [8] extends this coefficient to the incomplete case. This measure, denoted by W_S, is related to G_S in (28) through

$$W_S = \frac{12(k+1)^2}{n^2\lambda^2 t(t-1)(t+1)^3} G_S. \qquad (40)$$

W_S is constructed so that it achieves the maximum value of 1 when the incomplete rankings share one compatible complete ranking. The property of W in (39), that it can be expresed as a linear function of the average of the Spearman rank correlations between all pairs of rankings, can be extended to the incomplete case by properly defining the correlation between two incomplete rankings. One possible approach is to define the correlation between the incomplete rankings μ_i^* and μ_j^* by

$$\alpha^*(\mu_i^*, \mu_j^*) = 1 - \frac{2(d^*(\mu_i^*, \mu_j^*) - m^*)}{M^* - m^*}. \tag{41}$$

Such a definition may be justified in various ways. For one, it is related to the average of the correlations between the corresponding complete compatible rankings in the following way. If this average is denoted by $\tilde{\alpha}(\mu_i^*, \mu_j^*)$, then

$$\tilde{\alpha}(\mu_i^*, \mu_j^*) = \frac{1}{a^2 c}[C(\mu_i^*)]'T'TC(\mu_j^*) = 1 - \frac{1}{c}d^*(\mu_i^*, \mu_j^*)$$

which on simplification becomes

$$\tilde{\alpha}(\mu_i^*, \mu_j^*) = \left(\frac{M^* - m^*}{M^* + m^*}\right)\alpha^*(\mu_i^*, \mu_j^*)$$

In the Spearman case a further justification for the definition (41) may be had by considering the vectors defined in (37). The Spearman distance (2) between ω_i and ω_j is found to be

$$d_S(\omega_i, \omega_j) = \left(\frac{k+1}{t+1}\right)^2 (d_S^*(\mu_i^*, \mu_j^*) - m_S^*) \tag{42}$$

so that the correlation between ω_i and ω_j defined by

$$corr(\omega_i, \omega_j) \equiv 1 - \frac{2d_S(w_i, w_j)}{Max(d_S)}$$

is exactly $\alpha_S^*(\mu_i^*, \mu_j^*)$. Further this correlation coincides with both type (a) and type (b) correlations as defined by Kendall [17]. Turning now to the average correlation between all pairs of incomplete rankings, we have

$$\bar{\alpha}^* = \frac{1}{mb(mb-1)}\sum_{i \neq j}^{mb}\left(1 - \frac{2(d^*(\mu_i^*, \mu_j^*) - m^*)}{M^* - m^*}\right). \tag{43}$$

With $G = a^{-2}(TCf)'(TCf)$, use of

$$\sum_{i \neq j}^{mb} d^*(\mu_i^*, \mu_j^*) = cm^2b^2 - m^*mb - G \tag{44}$$

gives

$$(mb-1)\bar{\alpha}^* = \frac{2}{mb(M^* - m^*)}G - 1. \tag{45}$$

In the Spearman case, use of (38) yields

$$(mb-1)\bar{\alpha}_S^* = \frac{m\lambda(t+1)}{(k+1)}W_S - 1, \tag{46}$$

so that

$$W_S = \frac{(k+1)}{m\lambda(t+1)}(\bar{\alpha}_S^*(mb-1)+1). \tag{47}$$

It may also be noted that use of (31) shows that the relationship between $\bar{\alpha}_S^*$ and the Durbin statistic D is

$$(mb-1)\bar{\alpha}_S^* = \frac{1}{t-1}D - 1$$

which is the same as the relatioship between $\bar{\alpha}_S$ and Friedman's statistic in the complete ranking case. For the Kendall case a measure of concordance may be defined to be

$$W_K = \frac{1}{\gamma}G_K$$

where γ is the value of G_K when all the incomplete rankings are compatible with one complete ranking which can be taken to be in natural order. However, the calculation of γ in this case is not as straightforward as is the case of G_S. (see Alvo and Cabilio [3]).

Acknowledgements:

These authors are indebted to Persi Diaconis for suggesting the consideration of cyclic designs and for several useful comments. This work was supported in part by the Natural Sciences and Engineering council of Canada, Grants Nos. A-9068 and A-9076 respectively.

8.8 References

[1] Alvo, M., Cabilio, P., and Feigin, P.D. Asymptotic theory for measures of concordance with special reference to average Kendall tau. *Ann. Statist.* **10**: 1269-1276, 1982.

[2] Alvo, M. and Cabilio, P. A comparison of approximations to the distribution of average Kendall tau, *Commun. Statist. Theor. Metho.* **13**: 3191-3213, 1984.

[3] Alvo, M., and Cabilio, P. On the balanced incomplete block design for rankings. Technical Report No. 130. Ottawa-Carleton Laboratory for Research in Statistics and Probability, 1989.

[4] Alvo, M., and Cabilio, P. On the balanced incomplete block design for rankings. *Ann. Statist.* **19**: 1597-1613, 1991.

[5] Cochran, W.G., and Cox, G.M. *Experimental Design.* Wiley, New York, 1957.

[6] Critchlow, Douglas E. *Metric Methods for Analyzing Partially Ranked Data.* Lecture Notes in Statistics. Springer-Verlag, Berlin, 1985.

[7] Daniels, H.E. The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33**: 129-135, 1944.

[8] Durbin, J. Incomplete blocks in ranking experiments. *Brit. J. of Psychology* IV, 85-90, 1951.

[9] Ehrenberg, A.S.C. On sampling from a population of rankers. *Biometrika* **39**:82-87, 1952.

[10] Feigin, P.D., and Alvo, M. Intergroup diversity and concordance for ranking data: An approach via metrics for permutations, *Ann. Statist.* **14**:691-707, 1986.

[11] Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32**:675-699, 1937.

[12] Hájek, J., and Šidák, Z. *Theory of Rank Tests.* Academic Press, New York, 1967.

[13] Hays, W.L. A note on average tau as a measure of concordance. *J. Amer. Statist. Assoc.* **55**: 331-341, 1960.

[14] Hollander, M., and Sethuraman, J. Testing for agreement between two groups of judges. *Biometrika* **65**:403-411, 1978.

[15] Jirina, M. On the asymptotic normality of Kendall's rank correlation statistic. *Ann. Statist.* **4**:214-215, 1976.

[16] John, J.A. *Cyclic Designs.* Chapman and Hall, London, 1987.

[17] Kendall, M.G. *Rank Correlation Methods.* Fourth Edition. Griffin, London,1975.

[18] Kendall, M.G., and Babington Smith, B. The problem of m rankings, *Ann. Math. Statist.* **10**: 275-287, 1939.

[19] Kendall, M.G., and Stuart, A. *The Advanced Theory of Statistics,* Vol. 2, Fourth Edition. Griffin, London, 1979.

[20] Puri, M.L. and Sen, P.K. *Nonparametric Methods in Multivariate Analysis.* John Wiley & Sons. New York, 1971.

[21] Quade, D. Average internal rank correlation. Technical Report, Mathematical Centre, Amsterdam, 1972.

[22] Randles, R.H., and Wolfe, D.A. *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York, 1979.

[23] Rao, C.R. Diversity and dissimilarity coefficients: A unified approach. *J. Theoret. Pop. Biol.* **21**:24-43, 1982.

[24] Schucany, W.R., and Frawley, W.H. A rank test for two group concordance. *Psychometrika* **38**: 249-258, 1973.

9

Applications of Thurstonian Models to Ranking Data

Ulf Böckenholt[1]

ABSTRACT Thurstonian models have proven useful in a wide range of applications because they can describe the multidimensional nature of choice objects and the effects of similarity and comparability in choice situations. Special cases of Thurstonian ranking models are formulated that impose different constraints on the covariance matrix of the objects' utilities. In addition, mixture models are developed to account for individual differences in rankings. Two estimation procedures, maximum likelihood and generalized least squares, are discussed. To illustrate the approach, data from three ranking experiments are analyzed.

Key words and Phrases: common factor model, normal orthant probabilities, mixture models, ranking data.

9.1 Introduction

This paper discusses a class of ranking models that is based on Thurstone's [33] random utility approach. Although the idea to analyze ranking data from a Thurstonian perspective is not a new one (Thurstone [34]; Daniels [13]), only recently have Thurstonian ranking models been developed that facilitate a parsimonious and flexible modeling of ranking data (Böckenholt [7]; Brady [8]). The main reason for this slow development is related to the computational complexity in estimating the model parameters. However, recent advances in evaluating the multivariate normal distribution make now the application of Thurstonian ranking models practical. This paper presents an overview of these models and introduces some new applications in the context of modeling individual differences.

As early as 1931, Thurstone suggested rankings as a substitute for paired comparison data. However, in his attempts to analyze these data he did not treat rankings as unit of analysis but converted them to paired comparison data to which he applied his influential 'law of comparative judgment' (Thurstone [33]). The crucial idea underlying Thurstone's approach is that

[1] Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois

a judgment about an object i can be represented by a random variable u_i which is composed of a fixed and a random part (Bock & Jones [5]),

$$u_i = \mu_i + d_i.$$

Thus, the location of object i is determined by its 'affective value', μ_i, and corresponds to the modal response of a homogeneous group of respondents to that object. Random components of the individual judgments are represented by d_i. When the joint distribution of d_i and d_j associated with objects i and j is bivariate normal with a zero mean vector and covariance matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ij} & \sigma_j^2 \end{bmatrix},$$

the probability of preferring object i over object j is given by the difference process

$$P(i,j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\,\sigma_{ij}}}\right),$$

where Φ denotes the standard normal distribution function. In the next section, Thurstone's law of comparative judgment is extended to ranking judgments.

9.2 The Ranking Model

In a ranking task, a homogeneous group of respondents is asked to rank $t > 2$ distinct objects with respect to some criterion. The respondents' rankings are assumed to be independent of each other. According to the random utility approach, an object i is ranked first if its value, u_i, is largest; it is ranked second if its value is second largest, etc. Note that although a judge may compare the objects in any order, the affective values remain fixed during the ranking trial. Thus, it is assumed that a judge does not 'resample' affective values when s/he rank orders the objects.

Ranking outcomes are represented by the ordering vector $\boldsymbol{s} = (i, j, \ldots, k)$ which denotes that object i is judged superior to object j which in turn is judged superior to the remaining objects in the choice set, with the last preferred object being k. The probability of observing $\boldsymbol{s} = (i, j, k, \ldots, s, t)$ is

$$P\{\boldsymbol{s} = (i,j,k,\ldots,s,t)\} = P\{(u_i - u_j > 0) \cap (u_j - u_k > 0) \cap \cdots \cap (u_s - u_t > 0)\}.$$

To compute this probability, it is assumed that the joint distribution of $\boldsymbol{d} = (d_i, d_j, \ldots, d_t)'$ is multivariate normal with covariance matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{d} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

and

$$\boldsymbol{u} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

As a result, the probability of a particular rank order can be determined by evaluating a $(t-1)$ variate normal distribution function. For example, when $t = 3$,

$$P(s = i, j, k) = \Phi_2 \left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}}}, \frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2 - 2\sigma_{jk}}}, \rho_{ij,jk} \right),$$

where Φ_2 is the standard bivariate normal distribution function with integration limits, $\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}}}$ and $\frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2 - 2\sigma_{jk}}}$, and correlation coefficient

$$\rho_{ij,jk} = \frac{\sigma_{ij} - \sigma_{ik} - \sigma_j^2 + \sigma_{jk}}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}}\sqrt{\sigma_j^2 + \sigma_k^2 - 2\sigma_{jk}}}.$$

The covariance terms describe the perceived relationship between two objects, and the variance components describe the rankers' heterogeneity in their judgments. For instance, assume that three objects to be ranked have the same average popularity ($\mu_i = \mu_j = \mu_k$) but that the discriminal dispersion of object i is much larger than of the two other objects. In the limit,

$$P(s = i, j, k) = P(s = i, k, j) = P(s = j, k, i) = P(s = k, j, i) = 0.25,$$

and

$$P(s = j, i, k) = P(s = k, i, j) = 0.$$

Thus, in this extreme case of heterogeneity, object i is either ranked first or last although there are no differences in the mean rank values of the objects. To illustrate the effect of non zero covariances, assume that object i and j are very similar but have the same average popularity as object k. In the limit,

$$P(s = i, j, k) = P(s = j, i, k) = P(s = k, i, j) = P(s = k, j, i) = 0.25,$$

and

$$P(s = i, k, j) = P(s = j, k, i) = 0.$$

Note that the rank positions of object i and object j are always adjacent although there are no differences in the average rank values of each object. In general, by relaxing the assumption that the mean utilities are independently distributed, Thurstonian models allow for a straightforward representation of effects of heterogeneity and perceived similarity between choice objects.

Thurstone [33] suggested several simple constraints to be imposed on the covariance matrix $\boldsymbol{\Sigma}$ that have been applied successfully in the modeling of paired comparison data. For example, in his Case III formulation, Thurstone allowed the variances to vary, while in his Case V formulation, Thurstone posited that the variances are equal. Both constraints reduce considerably the number of parameters to be estimated. In Case V, $(t-1)$ mean values have to determined; in Case III $(t-1)$ mean values and $(t-1)$ variance components have to be estimated. Daniels [13] suggested the Case V constraint for the analysis of ranking data. When the number of objects to be ranked is large, modeling the covariance matrix $\boldsymbol{\Sigma}$ may also prove useful. This is the topic of the next section.

9.3 Modeling Σ

Under the assumption that respondents use for their rankings a set of r dimensions that characterize the choice alternatives, the covariance matrix may be modeled by a common factor structure,

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},$$

where the loading matrix $\boldsymbol{\Lambda}(t \times r)$ and the diagonal matrix $\boldsymbol{\Psi}(t \times t)$ are all parameters. Thus, $\boldsymbol{u}$ can be written in the form

$$\boldsymbol{u} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{d}.$$

Under the assumptions that the common factors $\boldsymbol{f}(r \times 1)$, and $\boldsymbol{d}(t \times 1)$ are multinormally distributed,

$$\boldsymbol{f} \sim N(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{d} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}),$$

and $\boldsymbol{f}$ and $\boldsymbol{d}$ are independent,

$$\boldsymbol{u} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}).$$

Indeterminancies in the estimation of the common factor model can be handled by imposing constraints on $\boldsymbol{\Lambda}$ and $\boldsymbol{\mu}$ (Bartholomew [2]). The ranking probabilities are invariant under a transformation of the loading matrix by any orthogonal matrix of order r and addition of a constant r-dimensional vector. Similarly, the ranking probabilities are not affected by any linear transformation of the mean utilities. The factor analytic approach has proven useful in the analyses of a wide variety of paired and multiple comparison data (Arbuckle & Nugent [1]; Brady [8]; Manski & McFadden [23]).

An alternative to the linear factor model is provided by the ideal point factor model (Brady [8]; Takane [31]). This model assumes that respondents can be represented by their ideal points, $\boldsymbol{f}$, on a latent continuum. In this case the 'affective value' of a choice object is determined as,

$$u_i = \mu_i - (\boldsymbol{\lambda}_i - \boldsymbol{f})'(\boldsymbol{\lambda}_i - \boldsymbol{f}) + d_i,$$

and by making the same distributional assumption as for the linear factor model, we obtain

$$\boldsymbol{u} \sim N(\boldsymbol{\mu} - \ \mathrm{diag}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')\mathbf{1}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}),$$

where $\mathbf{1}$ is a t-dimensional unit vector. Thus, although the mean vectors of the linear and the ideal point factor model differ, their covariance structures are identical.

In their analyses of paired comparison data, Takane [30] and Heiser and DeLeeuw [18] assumed that the unique variances can be ignored and introduced a principal component type of constraint with $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. This model requires the estimation of $(t - 1)$ mean utilities and $(t\ r - r(r + 1)/2 - 1)$ effective parameters to determine $\boldsymbol{\Sigma}$. Other approaches that constrain the structure of the covariance matrix and the mean utilities include the *Wandering Vector* model (DeSoete & Carroll [14]) and the *Wandering Ideal Point* model (I. Böckenholt & Gaul [6]; DeSoete, Carroll, DeSarbo [15]). It seems likely that these approaches developed for modeling paired comparison data will also prove useful in the analysis of ranking data.

9.4 Subpopulations

The assumption of a preference structure that is common to all rankers may not always be adequate and can be rather restrictive. For example, it is easy to imagine that in an election there are several groups of voters with very different opinions about the political candidates. In this case it may be more appropriate to analyze the data of each group separately. This approach is straightforward if group membership is known for every ranker. Group differences can be investigated systematically by comparing the groups' scale values and covariance structures. An application of this approach is presented in Example II.

However, if group membership is not known, a mixture approach may be utilized for modeling population differences in ranking data. Recently, Croon [12] demonstrated that a special case of mixture models, latent class analysis, in combination with Luce's [21] or the Pendergrass-Bradley [26] ranking model, can prove useful in the exploration of individual differences in ranking data. Consistent with the Thurstonian framework, a multivariate normal mixture approach is adopted here that assumes that the heterogeneous population of rankers can be regarded as a finite mixture of more homogeneous subpopulations. Each subpopulation is characterized by its distinct mean ranking of the choice alternatives. Consequently, the probability of observing a ranking pattern is a weighted sum of each subgroup's ranking probability, $P_g(s)$

$$P(s) = \sum_{g=1}^{w} \pi_g P_g(s).$$

The relative size of each group is denoted by π_g and because the subpopulations are assumed to be exhaustive and disjunctive,

$$\sum_{g=1}^{w} \pi_g = 1 \text{ with } \pi_g > 0.$$

For example, the probability of observing the ordering vector $s = (1,2,3)$ in the g-th subpopulation is given by

$$P_g(s = 1,2,3) = \Phi_2 \left(\frac{\mu_{1g} - \mu_{2g}}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}}, \frac{\mu_{2g} - \mu_{3g}}{\sqrt{\sigma_2^2 + \sigma_3^2 - 2\sigma_{23}}}, \rho_{12,23} \right).$$

Although (theoretically) the covariance matrix of each subgroup may be of arbitrary form, it is restricted to be equal for all subpopulations. This constraint is introduced because the log-likelihood of a mixture of normal distributions is unbounded in the case of unequal covariance matrices (McLachlan & Basford [24]). In general, parameters of finite mixtures of multivariate normal distributions are identified (see Teicher [32]). However, in the analysis of ranking data, it is necessary to impose parameter constraints to solve for the location, scale, and rotational indeterminancies of each subgroup's estimated parameter values as discussed in the previous section. A further essential condition for identifiability is that the total number of free parameters does not exceed the number of independent rankings.

In the application of mixture models reported in Example III, use is made of the Expectation-Maximization (E-M) algorithm. The implementation of this algorithm is straightforward and not further discussed here because it is well documented in the literature. For example, Hathaway [17] and Lwin and Martin [22] provide a detailed discussion of this algorithm for estimating normal mixture models. However, it should be noted that the E-M algorithm has slow convergence properties. As a result, the E-M algorithm becomes impractical for the estimation of mixture models for rankings with more than four choice alternatives because of the repeated evaluation of the ranking model at each Maximization-step. More detail on the estimation of the ranking models is presented in the next section.

9.5 Model Estimation and Tests

Two general approaches, maximum likelihood and general least squares, are presented for estimating the parameters of a Thurstonian ranking model. The maximum likelihood approach models the probability of observing a ranking pattern. In contrast, the generalized least squares (GLS) approach takes into account only binary choice probabilities or trinary rankings, and, as a result, is more attractive from a computational viewpoint.

Maximum likelihood estimation: Parameter estimates of the Thurstonian ranking models can be obtained by using maximum likelihood methods. Assuming random sampling of N subjects for the ranking task, the log-likelihood function is specified as

$$ln\ L = c + \sum_{l=1}^{t!} r(\boldsymbol{s}_l) ln\ P(\boldsymbol{s}_l),$$

where c is a constant and $r(\boldsymbol{s}_l)$ is the observed frequency of the ordering vector $\boldsymbol{s}_l$. The estimation of $P(\boldsymbol{s}_l)$ requires the computation of normal orthant probabilities. In the case of a covariance matrix with a common factor structure, numerical evaluation of the multivariate normal integral can be performed by Gauss-Hermite quadrature to any practical degree of accuracy (Stroud & Sechrest [29]). Because of the assumption of a common factor structure, this approach is computationally unproblematic for a small number of factors. Gauss-Hermite quadrature has proven practical for a large number of items in the evaluation of the multivariate normal distribution of person parameters in full information factor analysis (Bock & Aitkin [3]; Bock, Gibbons, & Muraki [4]). If the covariance structure is general and not limited to a common factor model, Schervish's [28] error bounded algorithm can be employed for evaluating a normal distribution. To ensure the positive definiteness of $\boldsymbol{\Sigma}$ in this case, it is useful to estimate the Cholesky factor $\boldsymbol{V}$ with $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{V}'$. Unfortunately, the Schervish algorithm is rather slow for practical applications in a 6- or higher normal variate case.

When the number of objects to be ranked is small, large sample tests of fit are available based on the likelihood-ratio (LR) chi-square statistic:

$$G^2 = 2\sum_{l=1}^{t!} r(\boldsymbol{s}_l) ln \frac{r(\boldsymbol{s}_l)}{r^e(\boldsymbol{s}_l)},$$

where $r^e(\boldsymbol{s}_l)$ refers to the expected ordering frequencies under the model to be tested. Asymptotically, for small t, if the ranking model provides an adequate description of the data, then G^2 will be distributed with $(t!-h-1)$ degrees of freedom, where h refers to the number of parameters to be determined. Moreover, nested ranking models can be compared by computing the difference between their LR statistics. This difference is asymptotically distributed as a chi-square statistic with the number of degrees of freedom equal to the difference between the number of parameters in the unrestricted and restricted model. For large t only a small subset of all possible ranking patterns may be observed and, as a result, there is little justification that the LR statistic will follow a chi-square distribution. Instead, the model fit may be assessed by comparing standardized differences between the observed and fitted ranking probabilities in different partitions of the ranking data (Cohen & Mallows, [9]).

Generalized Least Squares Estimation: A computationally attractive procedure for estimating the parameters of the Thurstonian ranking models is provided by the generalized least squares (GLS) principle. In the analysis of binary data, Christofferson [10] and Muthen [25] demonstrated the usefulness of the GLS estimators for determining the common factor model and structural equation models. Brady [8] suggested GLS estimation for the analysis of ranked data and provided a detailed discussion of GLS estimation for binary choices and trinary rankings. GLS estimators are defined by minimizing

$$F = \boldsymbol{e}'\boldsymbol{S}^{-1}\boldsymbol{e},$$

where $\boldsymbol{e} = \boldsymbol{P} - \boldsymbol{P}^*$and $\boldsymbol{S}$ is a sample estimate of the covariance matrix of $\boldsymbol{e}$. If only binary choice probabilities are extracted from the ranking data, $\boldsymbol{P}$ is a $\binom{t}{2} \times 1$ vector of the observed binary choice probabilities and $\boldsymbol{P}^*$ contains the binary choice probabilities expected under the ranking model.

The information contained in the binary choice probabilities may be rather limited. Brady showed that more efficient estimators are obtained if trinary rankings are used. In this case, $\boldsymbol{P}$ is a vector containing nonredundant trinary ranking probabilities and $\boldsymbol{P}^*$ contains the corresponding expected trinary probabilities. Further improvements in efficiency may be obtained by considering higher-order rankings. However, in this case, the GLS approach loses its computational attractiveness. The GLS estimators for binary choices and trinary rankings are consistent. Model tests may be based on (min F) which follows asymptotically a chi-square distribution with degrees of freedom equal to the difference between the number of independent probabilities and the number of free parameters.

9.6 Applications

Example 1: Analysis of Incomplete Rankings

In the first example a data set with incomplete rankings presented by Pendergrass and Bradley [26] is reanalyzed. These authors suggested an extension of the Bradley-Terry-Luce model (Luce [21]) for modeling trinary ranking data. As a numerical illustration of their model, they collected ranking data according to a balanced incomplete block design (BIBD) with four objects presented three at a time so that each object is considered three times by four judges and each pair of objects is considered by two judges. This whole pattern is replicated forty times. The data are displayed in Table 1.

Table 1
Ranking Frequencies (Pendergrass and Bradley [26])

Triples ijk	ijk	ikj	jik	jki	kij	kji
123	10	8	8	6	4	4
124	12	8	8	6	4	2
134	10	8	8	8	4	2
234	8	6	6	8	6	6

The initial model fitted to these data is the Thurstonian model with identity covariance matrix. This model requires the estimation of three parameters and yields a $G^2 = 6.57$ on 17 degrees of freedom. The parameter estimates are $\hat{\mu}_1 = .317$, $\hat{\mu}_2 = .066$, $\hat{\mu}_3 = -.046$, and $\hat{\mu}_4 = -.336$. Compared to the likelihood ratio statistic, $G^2 = 23.02$ (df $= 20$), under the hypothesis of no differences between the four objects ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$), a significant reduction in the fit statistic is obtained ($\Delta G^2 = 23.02 - 6.57 = 16.45, \Delta$df$= 20 - 17 = 3$) which clearly indicates that the judges have an order preference.

The blocking effect may be investigated by first estimating a model with mean utilities that do not differ among the four blocks of triple comparisons and then comparing the fit of this model against the fit of a ranking model without this constraint. The difference in the fit statistics follows asymptotically a χ^2-distribution with 5 degrees of freedom ($\Delta G^2 = 6.57 - 4.22 = 2.35, \Delta$df$= 17 - 12 = 5$ and is not substantial which suggests that the four blocks do not need to be modeled separately. Finally, note that the differences between the estimated means for the second and the third object are quite small and probably negligible. To test this 'post-hoc' hypothesis, the two means were constrained to be equal 0, and a $G^2 = 7.08$ on 19 degrees of freedom was obtained. Thus, this restriction seems consistent with the data ($\Delta G^2 = 7.08 - 6.57 = 0.51, \Delta df = 2$) and the estimation of only one parameter ($\hat{\mu}_1 = .327$) is required.

In summary, subjects seem to distinguish only between the 'first' and the 'last' of the four objects. It is interesting to note that the most parsimonious Thurstonian model (with one parameter) still provides a somewhat better fit than the Pendergrass-Bradley extension of the Bradley-Terry-Luce models to triple rankings. The likelihood ratio statistics of this model is 7.94 on 17 degrees of freedom.

Example II: Testing for Agreement Among Groups of Judges

One procedure for testing for agreement among groups of judges is to determine the likelihood ratio statistic for a ranking model with equal scale values for all groups and compare this likelihood ratio statistic to the fit of a model without this constraint. This difference in the test statistics is compared to a χ^2-distribution with $[(w-1)(t-1)]$ degrees of freedom where w refers to the number of groups. This procedure is illustrated with a data set in Hollander and Sethuraman [19] who obtained the data from an unpublished dissertation by G. Sutton. Sutton asked 14 white females (WF) and 13 black females (BF) to rank order the gender with which they preferred to spend their leisure time. The females were between 70 and 79 years old and their three choice alternatives were (1) 'male', (2) 'female', and (3) 'both sexes'. The data are displayed in Table 2 with the first position denoting the most preferred alternative and the last position denoting the least preferred alternative. For example, five black females specified the preference rank order 'both sexes', 'male', and 'female'. Note that the following χ^2- statistics should be interpreted with care because the ranking frequencies are quite small.

Table 2
Ranking Frequencies (Hollander and Sethuraman [19])

Triples	123	132	213	231	312	321
W. Females	0	0	1	7	0	6
B. Females	1	1	0	0	5	6

The likelihood ratio test $G^2 = 21.12$ of a model with equal means for both groups and an identity covariance matrix has 8 degrees of freedom for this data set. The χ^2-statistic indicates that the means of the two groups may be different. A ranking model that allows for different mean utilities for each group yields a G^2 =4.91 on 6 degrees of freedom. This statistic may be decomposed into a $G^2 = 1.22$ for the WF group and a $G^2 = 3.69$ for the BF group and both LR statistics have three degrees of freedom. The difference between the likelihood ratio statistics for the constrained and the unconstrained model $\Delta G^2 = 21.12$ - $4.91 = 16.21$ on 2 degrees of freedom supports the hypothesis that there is little agreement between both groups. The estimated mean utilities for the WF group are $\hat{\mu}_{WF} =$ (-1.695, 1.015, .680) and for the BF group $\hat{\mu}_{BF} =$ (-.502, -.665, 1.167). Clearly, the rank orders of the three choice alternatives are different. The WF group does not like to socialize with only 'males' and seems to be indifferent between a 'female' companion and 'both sexes'. In contrast, the BF group prefers 'both sexes' and seems to be indifferent between a 'female' and a 'male' companion. A test of this 'post-hoc' hypothesis yields a $\Delta G^2 = 5.54$ - $4.91 = .63$ on 2 degrees of freedom with the scale values $\hat{\mu}_{WF} = (-1.677, .839, .839)$ and $\hat{\mu}_{BF} = (-.583, -.583, 1.167)$.

Cohen and Mallows [9] applied the ranking model suggested by Luce [21] for separate analyses of the two groups. The resulting LR-tests are compared to a χ^2 distribution with 3 degrees of freedom. Luce's model provided a somewhat less satisfactory fit of the data than the Thurstonian model with $G^2 = 3.19$ for the WF group and $G^2 = 6.39$ for the BF group. However, the conclusions reached by Luce's ranking model are similar to the ones presented above.

The assumption that the three choice alternatives are judged independently seems doubtful when considering that the 'both sexes' category includes 'female' and 'male' companions. However, the goodness-of-fit statistic of a Thurstonian ranking model with an identity covariance matrix indicates that the impact of the correlations between the responses is quite small. Of course, the Thurstonian approach is not limited to this independence assumption and can deal naturally with situations where the independence assumption is violated and ranking data from several groups (of unequal sample sizes) are collected (Fligner & Verducci [16]; Pettitt [27]).

Example III: A comparison of GLS and ML

The third example is taken from Croon [12] who reported the rankings of 2262 German respondents in a survey about the desirability of the four political goals: 1. Maintain order in the nation; 2. Give people more say in the decisions of the government; 3. Fight rising prices; 4. Protect freedom of speech. Table 3 contains the ranking data. In his detailed analysis of the data set, Croon made use of Ingelhart's [20] distinction between a materialistic and a post-materialistic value orientation. According to this theory, respondents with a materialistic value orientation prefer the first and the third item, while respondents with a post-materialistic orientation prefer the second and the fourth item. Croon found support for Ingelhart's distinction when modeling the ranking data with Pendergrass-Bradley's and Luce's ranking model in a latent class framework.

In a first analysis of the data set, a Thurstonian ranking model was fitted with an identity covariance matrix. Both ML and GLS estimation methods were used to determine the mean utilities of the ranking model. No support was found for this simple Thurstonian ranking model. Maximum likelihood estimation led to the LR-statistic $G^2 = 256.28$ with df = 20. GLS estimation was performed based on trinary and quaternary rankings. The corresponding chi-square statistics were 179.08 (df = 11) and 195.24 (df = 20), respectively. The degrees of freedom for the trinary rankings are determined as the difference between the number of independent probabilities (14) and the number of free parameters (3). To test the hypothesis that the independence structure of the covariance matrix caused the poor fit of the ranking model, a common one-factor model, $\boldsymbol{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}$, was estimated. This modification led to a substantial reduction in the LR-statistic, $G^2 = 60.38$ (df = 18). Similarly, the GLS chi-squared statistics were 41.59 (df =

9) and 60.15 (df = 18) for trinary and quaternary rankings, respectively. Although the fit of the one-factor ranking model is still far from being satisfactory, the decrease in the goodness-of-fit statistics indicates that the political goals are not perceived independently of each other.

Substantial agreement was found between the mean values and the factor loadings estimated by ML and GLS methods. Mean utilities estimated by ML methods are $\hat{\boldsymbol{\mu}}_{ML} = \{.514, -.449, .578, -.643\}$. Mean utilities estimated by GLS methods are for trinary rankings $\hat{\boldsymbol{\mu}}_{GLS3} = \{.533, -.441\ .558, -.650\}$ and for quaternary rankings $\hat{\boldsymbol{\mu}}_{GLS4} = \{.530, -.443, .561, -.649\}$. Clearly, these results point to the materialistic value orientation of the respondents. Moreover, the loadings of the one-factor ranking model are quite similar to the mean scale values. For example, $\hat{\boldsymbol{\lambda}}_{ML} = \{.62, -.59, .33, -.44\}$ which indicates that the materialistic item pair 1 and 3 and the post-materialistic item pair 2 and 4 are positively correlated, but negatively correlated with each other.

Table 3
Ranking order of Political Goals (Croon [12])

1234	137	2134	48	3124	330	4123	21
1243	29	2143	23	3142	294	4132	30
1314	309	2314	61	3214	117	4213	29
1341	255	2341	55	3241	69	4231	52
1423	52	2413	33	3412	70	4312	35
1432	93	2431	59	3421	34	4321	27

Mixture analysis: According to Ingelhart's [20] theory, there should be two distinct subpopulations of respondents, one with a materialistic and one with a post-materialistic value orientation. To test this hypothesis, mixture models with two and three groups were fitted to the ranking data by ML methods. For both analyses, the covariance matrix was set equal an identity matrix. The results indicated that a two group solution ($G^2 =$ 32.37, df = 16) may be preferable to a three-group solution ($G^2 = 28.18$, df = 12). Note, however, that the overall fit of the two-group model is poor. A residual analysis of this model showed that the assumption of equal variances may be too strong. This result was supported by the fit improvement obtained by a two-group model with a variance component estimated for the first item ($G^2 = 27.51$, df = 15; $\hat{\sigma}_{11} = .81$). Although this finding of more homogeneous responses to item 1 may not be a stable one, it is interesting to observe that item 1 occupies an extreme position in both groups. The estimated mean values for the first group are $\hat{\boldsymbol{\mu}}_1 = (.75, -.62, .69, -.82)$ and for the second group, $\hat{\boldsymbol{\mu}}_2 = (-.88, .62, .04, .22)$. While the difference between the third and the fourth item in the second group is small, the overall pattern of means corresponds roughly to Inglehart's expectation regarding a materialistic and post-materialistic value structure. However, the relative group sizes differ significantly ($\hat{\pi}_1 = .83$, $\hat{\pi}_2 = .17$), supporting the initial finding of a dominant materialistic value orientation.

9.7 Discussion

The approach presented here does not require a simultaneous ranking of all t objects. In particular, when the number of choice alternatives is large, incomplete rankings of $m(2 \leq m \leq t)$ out of t alternatives provide an attractive alternative to a complete ranking of all t choice alternatives (Critchlow [11]). For instance, in Example I subjects were asked to rank three (out of four) alternatives at a time. Other data collection methods such as paired comparisons and the method of first choices are special cases of an incomplete ranking task and can be treated in the same framework. Given this wide range of data collection devices, the choice of a particular ranking method can depend solely on the needs and requirements of an application. For example, the method of first choices seems more natural in the context of consumers' purchase decisions than other (incomplete) ranking methods. The paired comparison technique may be the method of choice if the experimenter wants to impose minimal constraints on a response behavior of the subject.

Imposing restrictions on the parameters of a ranking model facilitate straightforward testing of a variety of hypotheses, for instance, regarding differences between subpopulations and treatment effects. Moreover, by relaxing the assumption that the random variables representing the choice objects are independently and identically distributed, apart from location shifts, one can describe the multidimensional nature of choice objects and effects of similarity and comparability. Recent advances in evaluating the multivariate normal distribution facilitate the straightforward analysis of choice sets with at least eight alternatives by maximum likelihood methods. In addition, the application of the GLS method in Example III demonstrated that this estimation technique compares favorably with a ML approach. If this result can be supported in further research, GLS may be the method of choice in estimating the parameters of ranking models because of its computational attractiveness. Thus, the main reason, computational complexity, that hampered the application of Thurstonian ranking models may have lost its force, rendering a parsimonious and flexible modeling of ranking data.

Acknowledgements:

The author is grateful to Douglas Critchlow, Mike Fligner, Joe Verducci and two reviewers for helpful comments on this research.

9.8 References

[1] Arbuckle, J., & Nugent, J. H. A general procedure for parameter estimation for the law of comparative judgment. *British Journal of*

Mathematical and Statistical Psychology, **26**:240-260, 1973.

[2] Bartholomew, D. J. *Latent variable models and factor analysis.* London: Oxford University Press, 1987.

[3] Bock, R. D., & Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, **46**:443-459, 1981.

[4] Bock, R. D., Gibbons, R., & Muraki, E. Full-information item factor analysis. *Applied Psychological Measurement*, **12**:261-280, 1988.

[5] Bock, R. D., & Jones, L. V. *The measurement and prediction of judgment and choice.* San Francisco: Holden-Day, 1968.

[6] Böckenholt, I., & Gaul, W. Analysis of choice behavior via probabilistic ideal point and vector models. *Applied Stochastic Models and Data Analysis*, **2**:209-226, 1986.

[7] Böckenholt, U. Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, in press, 1992.

[8] Brady, H. Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika*, **54**:181-202, 1989.

[9] Cohen, A., & Mallows, C. L. Assessing goodness of fit of ranking models to data. *The Statistician*, **32**:361-373, 1983.

[10] Christoffersson, A. Factor analysis of dichotomized variables. *Psychometrika*, **40**:5-32, 1975.

[11] Critchlow, D. E. *Metric methods for analyzing partially ranked data.* New York: Springer, 1985.

[12] Croon, M. Latent class models for the analysis of rankings . In G. DeSoete, H. Feger, & K. C. Klauer (Eds.) *New developments in psychological choice modeling* (pp. 99-121). Elsevier: Holland, 1989.

[13] Daniels, H. E. Rank correlation and population models. *Journal of the Royal Statistical Society* B, **12**:171-181 1950.

[14] DeSoete, G., & Carroll, J. D. A maximum likelihood method for fitting the wandering vector model. *Psychometrika*, **48**:553-566 1983.

[15] DeSoete, G., Carroll, J. D., & DeSarbo, W. S. The wandering ideal point model: A probabilistic multidimensional unfolding model for paired comparison data. *Journal of Mathematical Psychology*, **30**:28-41, 1986.

[16] Fligner, M. A., & Verducci, J. S. Aspects of two group concordance. *Communications in Statistics, Theory and Methods*, 1479-1503 1987.

[17] Hathaway, R. J. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, **13**: 795-800, 1985.

[18] Heiser, W., & De Leeuw, J. Multidimensional mapping of preference data. *Mathematiques et Sciences humaines*, **19**:39-96, 1981.

[19] Hollander, M., & Sethuraman, J. Testing for agreement between groups of judges. *Biometrika*, **65**:403-411 1978.

[20] Inglehart, R. *The silent revolution.* Princeton: Princeton University Press, 1977.

[21] Luce, R. D. *Individual choice behavior.* New York: Wiley. 1989.

[22] Lwin, T., & Martin, P. J. Probits of mixtures. *Biometrics*, **45**:721-732, 1959.

[23] Manski, C., & McFadden, D. Alternative estimators and sample designs for discrete choice analysis. In C. Manski, & D. McFadden (Ed.) *Structural analysis of discrete data with econometric applications.* MIT Press: Cambridge, 1981.

[24] McLachlan, G. J., & Basford, K. E. *Mixture models.* New York: Marcel Dekker, 1989.

[25] Muthén, B. Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, **22**: 43-65, 1983.

[26] Pendergrass, P. N., & Bradley, R. A. Ranking in triple comparisons. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp.331-351). Palo Alto, CA: Stanford University Press, 1960.

[27] Pettitt, A. N. Parametric tests for agreement amongst groups of judges. *Biometrika*, **69**:365-375, 1982.

[28] Schervish, M. Algorithm AS 195. Multivariate normal probabilities with error bound. *Applied Statistics*, **33**:81-94, 1984.

[29] Stroud, A. H., & Sechrest, D. *Gaussian quadrature formulas.* New York: Prentice Hall, 1966.

[30] Takane, Y. Maximum likelihood estimation in the generalized cases of Thurstone's law of comparative judgment. *Japanese Psychological Research*, **22**:188-196, 1980.

[31] Takane, Y. Analysis of covariance structures and probabilistic binary choice data. *Cognition and Communication*, **20**:45-62, 1987.

[32] Teicher, H. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, **34**:1265-1269, 1963.

[33] Thurstone, L. L. A law of comparative judgment. *Psychological Review*, **15**:284-297, 1927.

[34] Thurstone, L. L. Rank order as a psychophysical method. *Journal of Experimental Psychology*, **14**:187-201, 1931.

10

Probability Models on Rankings and the Electoral Process

Hal Stern [1]

ABSTRACT Multicandidate elections with a single winner suggest several questions about the manner in which the preferences of a group of individual voters are aggregated into a single social choice. Obvious examples are the national presidential primaries in the major political parties. However, nonpolitical exercises such as the ranking of job applicants or college football teams provide other examples. If an individual's preference is viewed as a ranking of the available choices then the literature on probability models for rankings (see the survey by Critchlow, Fligner and Verducci [11]) may be used to analyze methods for combining preferences. Several probability models are used to analyze the results of a five candidate presidential election of the American Psychological Association. In addition, simulated data generated by parametric probability models is used to consider the merits of a variety of voting systems.

10.1 Introduction

Political scientists, sociologists and philosophers are among the many to consider appropriate mechanisms for social choice. Consider an election with more than two candidates but only a single winner. The nomination of a political party's single presidential candidate in the United States is an example of such an election. The population of voters includes a variety of different constituencies. Each voter's preference is assumed to be recorded as a ranking of the set of candidates. An electoral system is defined as a mechanism for aggregating the preferences of a population of voters. It is well known that different electoral systems may lead to different outcomes in multicandidate elections (Rae [31], for example). Arrow's [1] result that it is impossible to construct a multicandidate voting system that satisfies four seemingly straightforward and desirable criteria has motivated many comparisons of electoral systems. Arrow's criteria are: monotonicity - a

[1] Department of Statistics, Harvard University, Cambridge, MA 02138

candidate should not be hurt by a gain of support or aided by a loss of support, nondictatorship - no one voter determines the winner, nonimposition - the outcome cannot be imposed independently of the voters' preferences, and independence from irrelevant alternatives - the relative standing of two candidates in an election consisting of k candidates should not change if one of the other $k-2$ candidates is removed from consideration. Research has focused on two questions: 1) for a given distribution of preferences what electoral system leads to the best social choice and 2) how can the preferences of a population be modeled.

Recent comparisons of electoral systems include Bordley [4], Merrill [28] [29], Chamberlin and Featherston [8] and Hill [24]. A variety of election paradoxes are exhibited in these studies. One example is the failure of electoral systems to elect a candidate that is preferred to each of the others in pairwise comparisons; such a candidate is called the Condorcet candidate (Condorcet [9]). Voting systems in which a decision is reached by successively eliminating candidates may fail to achieve Arrow's monotonicity condition. It is possible that voters can help their favorite candidate by placing him/her lower in their ranking (Fishburn and Brams [20]). A system for which such paradoxes occur frequently would seem to be undesirable.

The statistical and psychometric literature describe a variety of techniques for the analysis of data that is in the form of rankings. Critchlow, Fligner and Verducci [11] review four categories of parametric ranking models and their properties. Diaconis [15] describes the use of spectral analysis for ranked data. Probability models for rankings are applied to the election context in the remainder of the paper. In the next two sections, seven electoral systems and three parametric ranking models are described in more detail. Following these descriptions, data from the 1980 presidential election of the American Psychological Association is used to illustrate the voting systems and the probability models. Finally, the electoral systems and ranking models are considered under a variety of conditions using simulated election data.

10.2 Electoral Systems

To illustrate the seven voting systems, consider an election among 3 candidates (A,B,C) in which 200 voters have the following preferences:

$$38:\mathrm{ABC} \quad 33:\mathrm{ACB} \quad 30:\mathrm{BAC} \quad 33:\mathrm{BCA} \quad 28:\mathrm{CAB} \quad 38:\mathrm{CBA}.$$

This data set is a random sample of 200 voters from a population of voters with preferences described by the uniform distribution on the permutations. Two common summaries of a set of candidate rankings are given below. The table on the left gives $n(i,j)$, the number of voters ranking candidate i in position j, for each candidate i and position j. The table on the right gives the results of pairwise elections in which candidates other than i, j are ignored.

	Ranking		
Candidate	1	2	3
A	71	58	71
B	63	76	61
C	66	66	68

	Candidate		
Candidate	A	B	C
A	-	99	101
B	101	-	101
C	99	99	-

Thus 76 voters have candidate B ranked second. If candidate A is ignored then 101 voters prefer candidate B to candidate C.

The most commonly used electoral system is the plurality voting system. Each voter is allowed to cast one vote for exactly one candidate. Candidate A would be elected with 71 votes to 66 for C and 63 for B. Notice that those voters with preference BCA might choose to abandon candidate B in favor of candidate C in order to defeat their least favorite candidate A. This type of strategic voting is sometimes referred to as insincere voting because those voters that abandon B do not rank the candidates in the same order that their personal preference would suggest. Of course such voters would need to have detailed information about the preferences of the rest of the population in order to determine the best voting strategy. Such strategic voting issues are ignored here; each voter is assumed to cast one vote for his/her most preferred candidate. In multicandidate elections, if no candidate has a majority of the votes, then the plurality system is often supplemented with a runoff election between the two top candidates. In this case A would defeat C in the runoff election 101-99. We refer to this electoral system as the plurality with runoff voting system.

The Borda [3] rank system requires each voter to rank all of the candidates. If there are k candidates then the voter casts $k-1$ votes for his/her first choice, $k-2$ votes for his/her second choice, down to 0 votes for his/her least favorite candidate. For the above voter preferences, candidate B is the winner with 202 votes to 200 for A and 198 for C. One interpretation of this system is that each candidate gets one vote for every opponent to whom they are preferred.

Approval voting (Brams and Fishburn [6]) allows voters to cast one vote for as many candidates as the voter desires. The candidate with the most votes is elected. The system seems to have favorable properties for getting people to vote sincerely according to their beliefs. There is still little information about the voting strategies of individuals under approval voting. In the absence of such information, we assume that in a k- candidate election, voters are equally likely to choose $1, \ldots, k/2$ candidates. This corresponds roughly to current political theory (Brams and Nagel [7]) which suggests that the average number of approval votes cast will be less than half the number of candidates. In the simulations of 3 and 4 candidate elections described in Section 10.5, each voter is given a 50% chance of voting for one candidate and a 50% chance of voting for two candidates; in 5 candidate simulations the voter is equally likely to cast one, two or three votes. Given the number of votes to be cast, the voter is assumed to vote for the highest ranked candidates according to the voter's preference. The winner of the election according to this randomized implementation of approval voting depends on the randomly generated voting patterns. Either A, B, or C could be elected given the distribution of preferences in this example.

The Hare [23] system of preferential voting is used in elections for the Australian House of Representatives. Each voter ranks all of the candidates. If no candidate has a majority of the first place votes, then the candidate with the fewest first place votes is eliminated and votes for that candidate are distributed to their second favorite candidate. Candidate B would be eliminated and then candidate A would be elected in the second round with 101 votes. In three candidate elections, the Hare voting system will always elect the same candidate as a plurality election followed by a runoff. An interesting account of some of the things that can go wrong under this electoral system is given by Fishburn and Brams [20]. The Hare system has been quite successful in elections with many candidates and more than one winner. It is typically called the single-transferable vote in that context and is quite effective at achieving proportional representation. The Coombs [10] elimination system is a modification of the Hare system in which the candidate with the most last place votes is eliminated at each stage. Here, candidate A is eliminated and B defeats C with 101 votes. In practice, many voters fail to rank all of the candidates. For the American Psychological Association presidential election described in Section 10.4, 65% of the ballots were incomplete. The Hare system is still feasible in such a case. If all of a voter's ranked candidates are eliminated, then the voter's ballot is removed. The Coombs system is not feasible because it requires some assumption about the least favorite candidate of the incomplete ballots.

A family of techniques have been proposed that choose a candidate that defeats each of the other candidates in pairwise elections if such a candidate exists. This candidate is called the Condorcet candidate (Condorcet [9]) and the voting systems are called Condorcet completion methods. One such system, derived from the writings of Dodgson [17] (also known as Lewis Carroll), elects the candidate that is closest to being a Condorcet candidate. In this paper, the Dodgson system refers to a minimax election criterion. The results of two candidate elections between each pair of candidates are determined by ignoring the other candidates in every voter's ranking. The candidate whose worst defeat in any two candidate election is the smallest is elected. A Condorcet candidate, who wins each two candidate election by definition, is always elected by this system. Candidate B is elected by the Dodgson system in the sample election. An alternative interpretation of the same Dodgson proposal is described by Bartholdi, Tovey and Trick [2]. They determine the candidate that requires the minimum number of pairwise adjacent transpositions to be a Condorcet candidate. The distinction between the interpretation used in this paper and the Bartholdi interpretation is similar to the distinction between Cayley's distance and Kendall's distance (Diaconis [14]). Note however that the interpretation of the Dodgson voting system used here is based only on the results of pairwise elections and does not use the complete rankings explicitly.

For a given set of permutations, each election system can be used to determine a single winner and a consensus ranking of the candidates. Voting systems are compared in subsequent sections with regard to the frequency with which they elect the same candidate and with regard to the similarity of the consensus rankings. The consensus ranking is obtained by ranking candidates according to the number of votes received in the approval voting, plurality and Borda systems. When two or more candidates have the same number of votes, the tie is broken by a random selection of one of the tied candidates. For elimination systems, like the Hare and Coombs systems, candidates eliminated early are ranked last.

10.3 Models for Permutations

Probability models on rankings can be used to analyze election data and to generate simulated election data for studying the effectiveness of voting systems under various assumptions about the population. A permutation of k objects is represented as either a ranking π or an inverse ranking π^{-1} in the remainder of the paper. Let $\pi = (\pi(1) \cdots \pi(k))$ represent the ranking of k objects or candidates where $\pi(i)$ is the rank of candidate i. The inverse ranking or ordering is $\pi^{-1} =< i_1 \cdots i_k >$ where candidate i_j has rank j. As an example, $\pi = (3421)$ and $\pi^{-1} =< 4312 >$ are alternative representations of the preference of a voter whose first choice is candidate 4, followed by candidates 3, 1 and 2.

Three probability models are used throughout this article. The Bradley-Terry-Luce model (Luce [26], Bradley and Terry [5]) is an example of a Thurstone [33] order statistics model. For each voter, a random variable is associated with each candidate. The random variable X_i can be interpreted as the voter's reaction to candidate i on a linear scale. The random variables X_i are assumed to be independent of each other with common distribution F but different location parameters μ_i, $X_i \sim F(x - \mu_i)$. The probability of a ranking is the probability that the associated random variables are ordered according to that ranking. Thurstone developed this model for studies in which subjects ranked pairs of psychophysical stimuli; he considered the case in which F is taken to be the normal distribution. The Bradley-Terry-Luce (BTL) model is the result of using the Gumbel or extreme value distribution for F. The BTL is the only Thurstone model used in the simulations. If the candidates in an election are represented by the integers from 1 through k, then the probability of the ranking π with inverse ranking $< i_1 i_2 \cdots i_k >$ under the Bradley-Terry-Luce model is

$$P_{BTL}(\pi) = \prod_{r=1}^{k-1} \frac{p_{i_r}}{\sum_{j \in B_r} p_j}$$

where B_r is the set of candidates remaining after the first $r-1$ are chosen. The p_i are just a transformation of the location parameters of the random variables in the Thurstone model, $\mu_i = -\ln p_i$. The probability of a permutation is unchanged if all of the p_i are multiplied by a positive constant, or equivalently, if each of the μ_i is shifted by the same amount. By convention, the p_i are taken to have sum one, thus the BTL model contains $k-1$ free parameters. If each $p_i = 1/k$, then the BTL model is equivalent to the uniform distribution on all $k!$ rankings. The assumption of independent univariate $X's$ is quite restrictive. Mixture models, described later, are one mechanism for circumventing this restriction. Estimates of the parameters p_i of the BTL model are obtained by using Newton's method to maximize the multinomial likelihood. Steepest descent steps are also used in the algorithm to ensure that the estimates remain feasible (each $p_i \geq 0$) at each step. Other order statistics models used in the analysis of ranking data include the Thurstone [33] -Mosteller [30] -Daniels [12] model, based on a location family of normal random variables, and gamma models (Stern [32]), based on a scale family of gamma random variables.

The other models used in the simulations are special cases of the multistage ranking models of Fligner and Verducci [21]. These models are derived by assuming that each voter ranks the candidates sequentially, with the choice at each stage independent of the other stages. The probability of a ranking π is determined by how closely the ranking matches a central ranking π_0 at each stage of the process. Specifically, let $V_i = m$ if at the ith stage the $(m+1)$st best of the remaining candidates is selected. The variables V_i indicate the degree of discordance of the ranking π with respect to the central ranking π_0 at the ith stage. If $\pi_0 = (2314)$ and $\pi = (4312)$ then $V_1 = 0$ (since π has candidate 3 ranked first in agreement with π_0), $V_2 = 2$ (since candidate 4 is ranked second according to π ahead of candidates 1 and 2 that are preferred by the central ranking) and $V_3 = 1$ (candidate 2 is ranked ahead of candidate 1 in disagreement with π_0). In the general stagewise model,

$$P(\pi) = \prod_{\beta=1}^{k-1} p(V_\beta, \beta)$$

where $p(V_\beta, \beta)$ is the probability that the $V_\beta + 1$ st best candidate of the $k+1-\beta$ candidates remaining at stage β is chosen. We restrict attention to a subset of the stagewise models for which

$$p(\alpha,\beta) \propto \exp(-\alpha\theta_\beta)$$

where $\theta_\beta \geq 0$. The parameters θ_β measure the sensitivity of the probability model at each stage of the ranking. A large value of θ_β indicates that there is little disagreement at stage β, rankings with incorrect choices at this stage are unlikely. This model is called the $\phi-$component model by Fligner and Verducci, however the authors' initials FV are used to refer to the model here. The uniform model results if all of the θ_β are set to zero. The FV model is a generalization of Mallows' [27] ϕ model. Mallows' model results when $\theta_\beta = \theta$, $\beta = 1,\ldots,k-1$ and therefore $P(\pi) = \exp(-\theta K(\pi,\pi_0))$ where K is Kendall's distance (Kendall [25]). The use of the letter ϕ refers to an alternative parameterization of the model; the abbreviation PHI is used to refer to this model. Models which are more flexible than the FV models are obtained by replacing α with $f(\alpha)$ in the expression for $p(\alpha,\beta)$, where $f(\cdot)$ is a nonnegative and strictly increasing function with $f(0) = 0$ and $f(1) = 1$. Mallows' original description suggests that π_0 is known and $P(\pi)$ describes variability around the known central ranking. In the application of the FV and PHI models to election data, the central ranking is usually estimated from the data. The approach of Fligner and Verducci [21] is used to obtain maximum likelihood estimates of the continuous parameters θ_β and the central ranking π_0 in the FV and PHI models. The impact of the extra parameter π_0 is considered as part of the simulation study. This parameter is not a continuous parameter and the usual results concerning degrees of freedom do not apply. Many examples and simulations (see Critchlow, Fligner and Verducci [11] for one example) indicate that a wide variety of models, including the three considered here, provide a similar fit to ranking data.

The models described thus far use relatively few parameters to describe a distribution on $k!$ permuations. In addition, ranking data from elections would be expected to be multimodal, as a result of the different constituencies in the population. Mixture models can be used to describe such heterogeneous populations. A mixture distribution with M components has the form

$$P_{MIX}(\pi) = \sum_{i=1}^{M} \lambda_i P_i(\pi)$$

where λ_i is a parameter indicating the proportion of the population with preferences described by the distribution $P_i(\pi)$ and $\sum \lambda_i = 1$. For the mixture distributions considered here, the P_i are assumed to be distributions from a single family (BTL, FV, or PHI) but with different parameters. The mixture models have some of the properties of the spatial models used by political scientists. Each component of the mixture can be viewed as a single dimension or issue; the parameters of that component indicate the locations of the candidates on that dimension or issue.

Maximum likelihood estimates of the mixing parameters λ_i and the parameters of the component distributions $P_i(\pi)$ are estimated using the EM algorithm (Dempster, Laird and Rubin [16]). The EM algorithm is an iterative procedure for obtaining maximum likelihood estimates in situations with incomplete data. In this case, the data is incomplete because the appropriate mixture component for each individual voter is unknown. If the correct component for each voter was known, then maximum likelihood estimation would be straightforward using the procedures described above for each component. During each iteration of the EM algorithm, the E-step updates estimates of the probabilities that a voter with a particular preference ranking is from each component of the mixture given the current estimates of the component parameters. The E-step is followed by the M-step that computes new maximum likelihood estimates of the parameters of each component assuming that the E-step results are accurate. The starting values for the mixture proportions in the EM algorithm are chosen such that all of the λ_i are equal. Starting values for the component parameters are determined from the sample summaries $n(i, j)$, the number of voters ranking candidate i in position j. For the BTL model, the starting parameter values for the jth mixture component are taken to be a renormalized version of the vector $n(\cdot, j)$. The starting values are bounded away from zero. An initial estimate for the central ranking of the jth mixture component for the FV and PHI models is obtained from the ranking of the elements of the vector $n(\cdot, j)$. Starting values for the θ parameters in the FV and PHI models are chosen on an ad hoc basis, the first mixture component is assumed to be most highly peaked (large θ), and subsequent mixture components are less highly peaked (smaller θ). The maximum likelihood calculation converged from these starting values in all examples, however no formal investigation of the starting values has been carried out.

10.4 The American Psychological Association Election

Diaconis [15] provides the results of the 1980 American Psychological Association (APA) presidential election. There were 5 candidates, and voters were asked to rank order all of the candidates. Of the roughly 15,000 voters, only 5738 cast complete ballots. These complete ballots are considered here. Table 1 provides two summaries of the election data, the number of voters ranking each candidate in each position, and the results of pairwise elections between each pair of candidates. The complete data is provided by Diaconis. The Hare system was used to decide the winner of the election, candidate C. Plurality voting or plurality voting with a runoff election would also elect C. However, the Borda sum of ranks system and the

Table 1. APA Presidential Election 1980

			Ranking		
Candidate	1	2	3	4	5
A	1053	1519	1313	1002	851
B	775	1077	1415	1416	1055
C	1609	960	793	1050	1326
D	1172	972	1089	1164	1341
E	1129	1210	1128	1106	1165

			Candidate		
Candidate	A	B	C	D	E
A	—	3318	2897	3129	3053
B	2420	—	2593	2853	2711
C	2841	3145	—	3031	2935
D	2609	2885	2707	—	2745
E	2685	3027	2803	2993	—

Coombs elimination system would elect candidate A. In fact, despite being ranked first by fewer than 20% of the voters, candidate A is the Condorcet candidate.

Table 2 indicates the fit obtained by the models described in the previous section. The last column of the table is twice the difference between the log likelihood of the specified model and the uniform model (a model with no parameters). The saturated model refers to a multinomial model that incorporates one parameter for each of the $k!$ permutations subject to the usual multinomial restriction. The number of continuous parameters ignores the central ranking in the PHI and FV models. Results obtained by fitting 2 and 3 component mixture models are included in Table 2. Four and five component mixtures of BTL models are also included. The usual likelihood ratio tests for comparing nested models do not apply to mixture models because of difficulties with the asymptotic normality of the log likelihood under the null hypothesis (see Everitt and Hand [19] for more details). It is also difficult to reach conclusions about appropriate models because the large sample size produces large changes in the likelihood for even minor improvements in the model.

Mixtures of BTL models seem to provide a better fit to the data than mixtures of FV models. Table 3 gives the parameter estimates for the BTL mixture models with up to five components. In the two component mixture, the large group (roughly 72% of the voters) is described by a roughly uniform distribution with a slight preference for the order *DEBAC* . The

Table 2. Log-Likelihood of Ranking Models for APA Data

Model	Number of Mixture Components	Log Likelihood	Number of Continuous Parameters	Twice the LogLikelihood vs Uniform
Uniform	-	-27470.63	0	—
Saturated	-	-26611.87	119	1717.5
BTL	1	-27402.29	4	136.7
BTL	2	-26838.72	9	1263.8
BTL	3	-26774.68	14	1391.9
BTL	4	-26720.85	19	1499.6
BTL	5	-26707.38	24	1526.5
FV	1	-27338.76	4	263.7
FV	2	-26999.28	9	942.7
FV	3	-26978.13	14	985.0
PHI	1	-27408.49	1	124.1
PHI	2	-27347.88	3	245.5
PHI	3	-26997.86	5	945.5

smaller group has a highly peaked distribution about the preference order $CABED$ (the reverse order of the first group). The three component mixture provides an even more interpretable picture. A partisan minority that prefers candidates D, E by a wide margin over the others is removed from the large, approximately uniform distribution. Additional components find smaller groups with slightly different preferences. Note that Table 3 does not suggest the winner of the election, only the makeup of the voting population. The partisan group supporting the candidates A, C is much larger than the partisan group supporting the candidates D, E, suggesting that either candidate A or C be chosen. The smaller two mixture components in the three group model apparently correspond to the partisan voters from the two main groups in the APA at the time, clinical psychologists and academic psychologists. The large component, approximately uniformly distributed, includes the less partisan voters and voters from other constituencies.

The structure obtained here corresponds closely to the picture obtained by Diaconis' spectral analytic approach. Diaconis finds considerable structure in the data unaccounted for by the first order results of Table 1. Candidates A and C appear in positions 1 and 2, in either order, more often than would be expected by looking at the number of first and second place votes of each candidate. The same result holds for candidates D and E. Diaconis considers a spectral decomposition of the vector of 120 counts representing the number of voters who rank the candidates according to each permutation. Diaconis calculates the sum of squared residuals for the vector after accounting for the mean (the mean is 5738/120), 104384, the

Table 3. B-T-L Parameter Estimates for APA Data

No. of Mixture Components	Population Proportion	Candidate Parameters A	B	C	D	E
1	1.00	0.232	0.188	0.200	0.182	0.198
2	0.72	0.185	0.190	0.134	0.248	0.243
	0.28	0.234	0.059	0.640	0.028	0.039
3	0.67	0.196	0.212	0.162	0.215	0.215
	0.24	0.222	0.044	0.682	0.021	0.031
	0.09	0.052	0.032	0.013	0.481	0.422
4	0.59	0.185	0.199	0.159	0.197	0.260
	0.27	0.234	0.053	0.652	0.027	0.034
	0.09	0.060	0.035	0.014	0.492	0.399
	0.05	0.081	0.185	0.028	0.693	0.013
5	0.49	0.164	0.223	0.153	0.226	0.234
	0.28	0.229	0.054	0.652	0.029	0.036
	0.11	0.072	0.039	0.018	0.470	0.400
	0.08	0.308	0.088	0.121	0.059	0.424
	0.04	0.083	0.168	0.022	0.718	0.009

sum of squared residuals after accounting for the first order (Table 1) structure, 68593, and the sum of squared residuals after accounting for the structure of unordered pairs, 13495. This is quite similar to the sum of squared residuals obtained from the three component BTL mixture, 14300.

10.5 Simulation Results

Election data for which the preferences of the population are known provide a means for comparing voting systems. Simulations for 3, 4, and 5 candidates, with 25, 201 and 1001 voters (odd numbers are desirable for avoiding ties in pairwise elections) are carried out. The simulations with 25 voters are intended to represent elections or rankings by small committees. The larger populations are intended to illustrate the effects of samples that approach town and city size. For elections with three and four candidates, 25 voter elections are replicated 10000 times, 201 voter elections are replicated 2500 times and 1001 voter elections are replicated 1000 times. With five candidates the number of replications is 1000, 500 and 200. A sample of voter preferences, in the form of counts of voters selecting each of the $k!$

possible rankings, are generated from either the BTL, PHI or FV model, or a mixture model. The behavior of voting systems is investigated using the simulated data. In addition to the seven voting systems of Section 10.2, the modal rankings obtained from the maximum likelihood estimates of the three parametric ranking models are also considered as voting systems.

There are two interpretations of the simulations. The interpretation here is that each simulated set of voters is a population for which a social choice is required. Thus the election is viewed as a problem in aggregation of individual preferences rather than an estimation problem. An alternative interpretation is to view the probability model as describing the population. Each simulated data set is a sample from which an estimate of the appropriate population social choice is to be estimated. We return to this distinction in discussing simulations based on the APA data of the previous section.

Voting systems are evaluated in several ways in the remainder of this section. It is frequently held that the Condorcet candidate, the candidate who could defeat all others in pairwise elections, should be chosen, if such a candidate exists. Thus, the proportion of elections for which a system elects the Condorcet candidate, when one exists, provides one means of evaluation. The similarity of two electoral systems can be judged by the proportion of elections for which the two systems choose the same winner. A more comprehensive measure of similarity is the average value of Kendall's distance between the final consensus rankings of the two systems. The tendency of systems to elect a particular candidate, for example, a partisan minority candidate, is examined by considering the frequency with which each candidate is ranked first, second, etc. Political scientists also consider the ease with which a system can be manipulated by insincere voting. This measure is not reproduced here; Chamberlin and Featherston [8] find that the most difficult systems to manipulate are the Hare and Coombs electoral systems by virtue of the complicated sequence of calculations involved.

Simulations of electoral systems have also been carried out by Merrill [28] and Bordley [4]. Bordley generates a vector of candidate utilities for each voter from either the uniform distribution or a multivariate normal distribution. Voters rank the candidates in order of decreasing utility. Merrill generates voter preferences according to the uniform model or a spatial model (Downs [18]). In the spatial model, voters rank candidates according to the distance from each candidate to the voter's position in some issue space. Candidate and voter positions in the issue space are observations from multivariate normal distributions. Each author uses many of the voting systems described in Section 10.2. The results presented here for data from the uniform model match the results of Merrill and Bordley. The mixture models and Merrill's spatial models seem to produce similar effects.

Table 4 compares the voting systems with regard to the frequency with which Condorcet candidates are elected when voter preferences are from the uniform model. The table also includes the proportion of elections for which a Condorcet candidate exists. The results are similar to those in Table 2.1 of Merrill [29]. For example, in 3 candidate elections with 201 voters from the uniform model, 90.5% of the elections had a Condorcet winner. The proportion of elections with a Condorcet candidate can be computed directly when voters preferences follow the uniform distribution, assuming the number of voters is large. Let n_{ij} represent the number of voters preferring candidate i to candidate j in a pairwise election. Under the uniform model with independent voters, n_{ij} is a binomial random variable with n voters and probability of success $1/2$. The probability that candidate 1 is a Condorcet candidate is the probability that n_{12} and n_{13} are both greater than $n/2$. For a large number of voters, the two binomial variables are well approximated by two correlated normal random variables (correlation $= 1/3$). The probability that n_{12} and n_{13} are both greater than $n/2$ can be computed using a result of David [13] (see also Gibbons [22]) as 0.304. By symmetry, the probability of a Condorcet winner for a three candidate election is three times this quantity, 0.912. David also gives a formula for trivariate normal probabilities that can be used to determine the probability of a Condorcet candidate among four candidates in a large population with uniform preferences, 0.825. Table 4 seems to indicate that the normal approximation is adequate for even 25 voters. Exact calculations for five or more candidates require high dimension numerical integration.

With 201 voters and four candidates, the Hare system elected the Condorcet candidate 93.6% of the time while approval voting chose the Condorcet candidate 71.5% of the time. As mentioned earlier, the results for four candidate elections with 201 voters are based on 2500 replications. The Dodgson election system and Mallow's ϕ model always find a Condorcet candidate if one exists. The elimination systems, Hare and Coombs, are next best. The simpler vote accumulation systems, approval voting and plurality are least effective. They elect the Condorcet candidate less than 70% of the time with five candidates. The electoral systems are more likely to elect Condorcet candidates in the presence of a large voting population. The plurality with runoff system does better than approval voting but this advantage disappears as the number of candidates increases (Merrill [29]).

Table 5 illustrates simulated results for permutations generated from the BTL order statistics model. These simulations use 201 voters from the BTL model with the given parameters. The parameters of the BTL model correspond to the proportion of voters ranking each candidate first. The behavior of the voting systems is quite similar to the behavior under the uniform model. The results in this table, which we expect to be more representative of real elections than the results from the uniform model, indicate much better performance by all voting systems.

Table 4. Proportion of Condorcet Winners Elected for a Uniform Model

	25 voters			201 voters		
	Number of Candidates					
Voting System	3	4	5	3	4	5
Approval Voting	0.753	0.707	0.681	0.764	0.715	0.649
Borda	0.903	0.876	0.852	0.908	0.882	0.847
Plurality	0.790	0.694	0.610	0.765	0.676	0.569
Plurality with Runoff	0.959	0.901	0.807	0.960	0.899	0.782
Hare	0.959	0.925	0.884	0.960	0.936	0.880
Coombs	0.965	0.936	0.905	0.968	0.932	0.905
Dodgson	1.000	1.000	1.000	1.000	1.000	1.000
PHI max likelihood	1.000	1.000	1.000	1.000	1.000	1.000
FV max likelihood	0.868	0.814	0.796	0.869	0.798	0.747
BTL max likelihood	0.873	0.826	0.792	0.888	0.841	0.792
% Condorcet winners	0.915	0.836	0.758	0.905	0.824	0.798

Table 5. Proportion of Condorcet Winners Elected
Preferences from Bradley-Terry-Luce Model with 201 voters

	Number of Candidates		
	3	4	5
	$p =$	$p =$	$p =$
Voting System	$(.38, .32, .30)$	$(.28, .26, .24, .22)$	$(.24, .22, .20, .18, .16)$
Approval	0.881	0.773	0.768
Borda	0.949	0.883	0.895
Plurality	0.866	0.721	0.681
Plur w Runoff	0.986	0.924	0.916
Hare	0.986	0.957	0.966
Coombs	0.994	0.972	0.971
Dodgson	1.000	1.000	1.000
PHI max like.	1.000	1.000	1.000
FV max like.	0.935	0.850	0.871
BTL max like.	0.941	0.859	0.848
% Condorcet	0.979	0.936	0.948

The similarity of election systems can be judged by comparing the average distance between the consensus rankings of the different systems, or by comparing the frequency with which systems elect the same candidate. Table 6 shows the proportion of elections in which systems elect the same candidate below the diagonal and the average distance between the consensus rankings of electoral systems above the diagonal for the five candidate elections used to generate the last column of Table 5. As a first means of comparison of the voting systems, consider the frequency with which systems elect the same candidate. The Dodgson and Coombs votes agree quite often, choosing the same winner in 94.2% of the elections. The Mallows PHI maximum likelihood ranking produces the same winner as the Dodgson system in 98.6% of the elections. Approval voting and plurality voting, the systems that do not require complete rankings from the voters, are different from the other systems and from each other. The winner changes between approval voting and plurality voting in approximately one-third of the elections. Approval voting is more likely to agree with the ranking methods.

Table 7 lists the number of times that each candidate is elected in the 500 simulated elections. Plurality voting fails to elect the most popular candidate, candidate A by design here, 44% of the time. Plurality is also the only system for which candidate E, the least popular candidate according to the probability model, is ever elected. Approval voting fails to elect the most popular candidate in 34.6% of the elections. Both systems are extremely easy to implement and Brams and Fishburn [6] show that approval voting for three or four candidates leads to sincere voting by the electorate. The systems that require complete ranking of the candidates tend to elect the best candidate approximately 70% of the time. As expected, choosing the candidate ranked first according to the maximum likelihood estimates of the parameters of the Bradley-Terry-Luce model, the model used to generate the data, agrees with the designated winner most often.

The similarity of voting systems can also be judged by comparing the average Kendall's distance between the consensus rankings obtained by the voting systems, given above the diagonal in Table 6. The standard error of the averages is approximately 0.04. Considering average distances may be misleading since voting systems are typically designed to choose a winner and no claim is made about the ordering of the other candidates. Naturally the plurality system is extremely similar to the plurality system with runoff as they necessarily agree on the ranking of the last $k - 2$ candidates. The plurality system with a runoff election between the top two candidates is extremely similar to the Hare system. The two systems are identical for three candidate elections. Even with five candidates the difference is an average of only 0.65 pairwise adjacent transpositions with a standard error of 0.04 transpositions. Both systems eliminate candidates with few first place votes. Also striking is the similarity of the ranking based on Borda vote totals (sum of the ranks) and the ranking based on the maximum likelihood estimates of the BTL parameters. The average difference is 0.31 pairwise adjacent transpositions with a standard error of 0.025 transpositions. The largest differences in Table 5 are for systems that are qualitatively different.

Table 6. Similarity of Electoral Systems
Proportion of Agreement (below diagonal)
Average Distance Between Rankings (above diagonal)
5 candidates, 201 voters, BTL model p = (0.24,0.22,0.20, 0.18,0.16)

Voting								PHI	FV	BTL
System	Appr	Bord	Plur	Runoff	Hare	Coomb	Dodgs	mles	mles	mles
Approv	—	0.91	1.65	1.55	1.39	1.34	1.16	1.07	1.31	1.08
Borda	.796	—	1.51	1.34	1.11	0.74	0.73	0.46	0.75	0.31
Plural	.668	.688	—	0.27	0.86	1.93	1.58	1.61	1.83	1.73
Plur Run	.752	.836	.732	—	0.65	1.68	1.36	1.35	1.63	1.56
Hare	.748	.850	.686	.906	—	1.43	1.09	1.07	1.33	1.31
Coombs	.728	.850	.642	.860	.906	—	0.87	0.67	0.83	0.58
Dodgson	.758	.884	.668	.884	.934	.942	—	0.83	1.09	0.83
PHI	.756	.882	.660	.892	.938	.946	.986	—	0.57	0.60
FV	.706	.814	.622	.782	.832	.840	.860	.862	- -	0.80
BTL	.760	.920	.632	.784	.804	.818	.834	.836	.778	—

Table 7. Number of Times Each Candidate Elected in 500 Trials
5 candidates, 201 voters, BTL model p = (0.24,0.22,0.20, 0.18,0.16)

	Number of Times Elected				
Voting System	A	B	C	D	E
Approval Voting	327	131	37	5	0
Borda	367	113	19	1	0
Plurality	280	148	53	18	1
Plurality with Runoff	336	128	31	5	0
Hare	346	123	29	2	0
Coombs	353	114	29	4	0
Dodgson	358	114	27	1	0
PHI max likelihood	362	113	24	1	0
FV max likelihood	349	126	23	2	0
BTL max likelihood	380	104	15	1	0

Thus, the plurality voting system which uses only the voters' first choices is markedly different than the Coombs and Dodgson methods which use more information.

Tables 4-7 above are selected to illustrate the main results from a large number of simulations. The relative order of the voting systems, in terms of their ability to elect Condorcet candidates, remains the same when the number of candidates and the number of voters are varied. The observed similarities among systems are also consistent. The results do not seem to change when the BTL model is replaced by the FV model (Fligner and Verducci's ϕ-component model) or the PHI model (Mallow's ϕ model). In fact if appropriate parameter values are selected, the probability models produce extremely similar simulation results. Larger differences between the probability models might be expected when mixtures are considered in place of the one-dimensional voting populations.

Table 8 illustrates the results from a two-component mixture, with each portion of the population having preferences consistent with a BTL model. The majority of the population has preferences from a distribution that favors A and B, followed by C and D, with candidate E least preferred. There is a minority (30% of the population) that tend to favor C over the other candidates. The first column in the table gives the proportion of elections in which each voting system elected the Condorcet candidate (there was a Condorcet candidate in 489 of the 500 trials). As before, plurality and approval voting are least effective at electing Condorcet candidates. The Borda system is a slight improvement over these systems, however the elimination systems seem to perform best. The remaining columns indicate the number of times each candidate was elected. As with the simple one class models, plurality and approval voting are most likely to elect a minority candidate.

Table 8. Simulation of 500 Elections with 201 Voters - BTL Mixture
70% of the population $p = (0.36, 0.34, 0.15, 0.12, 0.03)$
30% of the population $p = (0.14, 0.13, 0.50, 0.12, 0.11)$

		Number of Times Elected				
Voting System	% Condorcet Winners Elected	A	B	C	D	E
Approval Voting	0.767	313	167	20	0	0
Borda	0.883	336	150	14	0	0
Plurality	0.753	284	153	63	0	0
Plurality with Runoff	0.947	333	156	11	0	0
Hare	0.949	334	153	13	0	0
Coombs	0.990	345	153	2	0	0
Dodgson	1.000	343	150	7	0	0
PHI max likelihood	1.000	342	149	9	0	0
FV max likelihood	0.890	333	158	9	0	0
BTL max likelihood	0.857	342	151	7	0	0

To consider a more realistic example, the three component BTL mixture that was fit to the APA presidential election data (parameter values in Table 3) is used to generate 1000 simulated elections with 5738 voters. The simulation results are summarized in Table 9. There was a Condorcet winner in 987 of the 1000 simulated elections. The Condorcet winner in the data set, candidate A, was the Condorcet winner only 22% of the time. Candidate C was the Condorcet winner 78% of the time. The first column of Table 9 is the first indication that the relative ranking of the voting systems is population dependent. The Borda system, which usually outperforms plurality and approval voting, is inferior in these simulations. The last five columns show that the Borda system elects candidate A in 997 of the 1000 simulated elections. The plurality and Hare systems elect candidate C over 99% of the time. The Dodgson and Coombs system are most effective at electing Condorcet candidates. Note that the BTL parameter estimates are ineffective in identifying Condorcet candidates, despite the fact the data is from a mixture of BTL models. The BTL model uses information about the complete ranking and always elects candidate A. The FV model appears to place more emphasis on the first stage and therefore favors candidate C.

Table 9 examines each simulated data set of 5738 voters as if it were the complete voting population. As described earlier, an alternative interpretation of the simulated data views the mixture distribution as describing the preferences of the APA population and each simulated data set as a sample from that population. Candidate C is the Condorcet candidate of the population described by the mixture distribution (C is preferred to A by 50.3% of the population). The Borda and Coombs systems applied to the population distribution elect candidate A. The plurality and Hare voting

Table 9. Simulation of 1000 APA-like Elections
(data generated by 3 component BTL mixture of Table 3)

		Number of Times Elected				
Voting System	% Condorcet Winners Elected	A	B	C	D	E
Approval Voting	0.766	320	0	680	0	0
Borda	0.223	997	0	3	0	0
Plurality	0.780	0	0	1000	0	0
Plurality with Runoff	0.877	97	0	899	0	4
Hare	0.780	0	0	996	0	4
Coombs	0.969	254	0	746	0	0
Dodgson	1.000	222	0	778	0	0
PHI max likelihood	1.000	228	0	772	0	0
FV max likelihood	0.759	89	0	911	0	0
BTL max likelihood	0.220	1000	0	0	0	0

systems applied to the population distribution elect candidate C. Under this alternative interpretation, the results of Table 9 illustrate the degree to which electoral system results vary under sampling. The sample Condorcet candidate does not match the "population" Condorcet candidate (candidate C) in 22% of the elections. The plurality and Hare voting systems elect the population Condorcet candidate C even in samples for which A is the Condorcet candidate. With larger sample sizes the Condorcet candidate in each sample would be expected to agree with the population Condorcet candidate.

This analysis suggests a plan of attack for organizations attempting to choose a voting system. Probability models may be used to model the population and generate simulated election data. Election systems can then be compared on the simulated data sets. The evidence from this APA election suggests that while the Hare system is competitive with plurality voting and approval voting, improvement would be obtained by using the Coombs system or a technique that checks for a Condorcet candidate first. Of course there are other considerations, as the Coombs system would invalidate the ballots of those voters that did not rank all candidates (2/3 of the voters in 1980).

Comparison of the probability models with each other is also possible based on the simulations. For each simulated data set, the log likelihood is computed for the BTL, FV and PHI models. Naturally, each probability model fits best on data generated by that model. As the PHI model is a subset of the FV model, the latter always provide a better fit as measured by the likelihood. However, the extra parameters provide no significant improvement when the data is actually generated by a Mallows model. The effect of estimating the central ranking in the FV and PHI models can be determined by considering simulations of uniform populations. For 1000 simulated elections with 3 candidates and 1001 voters with uniform preferences, the average difference between the saturated multinomial log likelihood and the BTL log likelihood is 2.96 with a standard error of 0.08. This mean difference matches quite closely the difference in degrees of freedom between the multinomial model (5 parameters) and the BTL model (2 free parameters). The average for the PHI model is 3.06 and the average for the FV model is 1.90 (approximately the same standard errors as above), both suggesting that the central ranking is the equivalent of one parameter in this case. However, for data generated using a highly peaked distribution (Mallows model with $\theta = 0.25$), the average difference between the saturated multinomial log likelihood and the FV log likelihood is 2.98 and the average for the PHI model is 3.91, as would be expected if the central ranking is not counted as a parameter. Evidently, the central ranking π_0 is obvious from the data in this case and the central ranking does not appear to be the equivalent of a parameter. Evidence from elections with more than 3 candidates leads to similar conclusions. The empirical evidence suggests that when the central ranking must be estimated, typically in populations with roughly uniform preferences, it is roughly equivalent to $k-2$ parameters in a k candidate election.

10.6 Conclusions and Summary

The data analysis and simulations in this study reinforce some obvious notions. Systems that require voters to completely rank all of the candidates in an election perform extremely well at electing Condorcet candidates. Within this group, it is interesting to note that the Coombs system outperforms the more popular Hare system. The Hare system is however more forgiving of incomplete ballots (a certainty in any election). If an organization agrees that electing Condorcet candidates is a reasonable goal, then a Condorcet completion method like the Dodgson election system is recommended.

The simpler balloting systems, approval voting and plurality, are qualitatively different than the ranking systems and each other. Due to the limited information obtained from the voter, neither system elects Condorcet candidates as often as the ranking systems. Results regarding approval voting should be viewed with skepticism due to the ad hoc rule used in the simulations to decide how many votes are cast by approval voters. Plurality elections tend to elect inferior candidates (as specified by the probability model) more often than any other system. Adding a runoff election when no majority candidate exists, a procedure currently used in many state and local elections in the United States, tends to alleviate this problem. Plurality elections followed by a runoff are more effective than approval voting at electing Condorcet candidates (although this advantage disappears when the number of candidates is large). The runoff election is however an added cost to the organization and these results suggest that approval voting may deserve more consideration.

The final decision of an electoral system will frequently depend on properties of the system that are not considered here: cost, ease of implementation, legal restrictions, and ease of manipulation. Simulations of APA-like data indicate that the structure of the voting population should be investigated to determine the consequences of the various election systems for a particular organization.

10.7 Acknowledgements

The author thanks Laurel Smith for the argument concerning the asymptotic probability of a Condorcet winner in the uniform model. The editors and two referees provided several thought provoking comments and helpful references. This work was partially supported by National Science Foundation grant SES-8805433 and Office for Naval Research Grant N00014-86K-2046. The author thanks Donald Rubin and Herman Chernoff for this support.

10.8 References

[1] K. J. Arrow. *Social choice and individual values*, 2nd ed., 1963. New York: John Wiley.

[2] J. J. Bartholdi III, C. A. Tovey, and M. A. Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare* **6**, 157-165, 1989..

[3] J. C. deBorda. Memoire sur les elections au scrutin. *Histoire de l'Academie Royale des Sciences,* 1781. Paris.

[4] R. F. Bordley. A pragmatic method for evaluating election schemes through simulation. *American Political Science Review* **77**, 123-141, 1983.

[5] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39**, 324-345, 1952.

[6] S. J. Brams and P. C. Fishburn *Approval Voting*, 1983. Boston:Birkhauser.

[7] S. J. Brams and J. H. Nagel. Approval voting in practice. to appear in *Public Choice*, 1990.

[8] J. R. Chamberlin and F. Featherston. Selecting a voting system. *The Journal of Politics* **48**, 347-369, 1986

[9] Marquis de Condorcet (also known as J.A.N. de Caritat). *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la plurailite des voix*, 1785. Paris.

[10] C. Coombs. *A theory of data*, 1964. New York: John Wiley.

[11] D. E. Critchlow, M. A. Fligner, and J. S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, **35**:294-318, 1991.

[12] H. E. Daniels. Rank correlation and population models. *J. Roy. Statist. Soc. Ser. B* **12**, 171-181, 1950.

[13] F. N. David. A note on the evaluation of the multivariate normal integral.
Biometrika **40**, 458-459, 1953.

[14] P. Diaconis. *Group representations in probability and statistics*, IMS Lecture Notes, Volume 11, 1988.

[15] P. Diaconis. A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17**, 949-979, 1989.

[16] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1-38 (with discussion), 1977.

[17] C. L. Dodgson. A method of taking votes on more than two issues, 1876. presented in the appendix of *The theory of committees and elections*, D. Black (1958), University Press, Cambridge.

[18] A. Downs. *An economic theory of democracy*, 1957. New York:Harper and Row.

[19] B. S. Everitt and D. J. Hand. *Finite mixture distributions*, 1981. London:Chapman and Hall.

[20] P. C. Fishburn and S. J. Brams. Paradoxes of preferential voting. *Mathematics Magazine* **56**, 207-214, 1983.

[21] M. A. Fligner and J. S. Verducci. Multistage ranking models. *Jour. Amer. Statist. Assoc.* **83**, 892- 901, 1988.

[22] J. D. Gibbons. *Nonparametric statistical inference*, 2nd ed., pg 209, 1985. New York:Marcel Dekker.

[23] T. Hare *The election of representatives, parliamentary and municipal: a treatise*, 3rd ed., 1865. London:Longman, Roberts and Green.

[24] I. D. Hill. Some aspects of elections – to fill one seat or many. *J. Roy. Statist. Soc. Ser. A* **151**, 243-275 (with discussion), 1988.

[25] M. G. Kendall. *Rank correlation methods*, 4th ed., 1970. New York:Hafner.

[26] R. D. Luce. *Individual choice behavior*, 1959. New York:John Wiley.

[27] C. L. Mallows. Non-null ranking models. I. *Biometrika* **44**, 114-130, 1957.

[28] S. Merrill III. A comparison of efficiency of multicandidate electoral systems. *American Journal of Political Science* **28**, 23-48, 1984.

[29] S. Merrill III. *Making multicandidate elections more democratic*, 1988. Princeton:Princeton University Press.

[30] F. Mosteller. Remarks on the methods of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika* **16**, 3-9, 203-206, 207-218, 1951.

[31] D. W. Rae. *The political consequences of election laws*, 1971. New Haven:Yale University Press.

[32] H. Stern. Models for distributions on permutations. *Jour. Amer. Statist. Assoc.* **85**, 558-564, 1990.

[33] L. L. Thurstone. A law of comparative judgment. *Psychol. Rev.* **34**, 273-286, 1927.

11

Permutations and Regression Models

Peter McCullagh [1]

ABSTRACT A class of exponential-family models on the set of permutations of k objects or items is described. The null or uniform model gives probability $1/k!$ to each of the $k!$ possible permutations. The first-order inversion model has as sufficient statistic the $k \times k$ matrix listing the number of times that each pair of candidates was ranked in that order, i.e. the number of times that candidate a was preferred over candidate b for all ordered pairs a and b. In the second-order inversion model the sufficient statistic is a similar listing for each ordered triplet of three candidates. Interesting sub-models are identified and used to help in the analysis of the APA election data.

11.1 Introduction

A class of factorial models for permutations is constructed as follows. For any permutation y we construct a potential function $\lambda(y)$ such that the probability of observing the permutation y is proportional to $\exp(\lambda(y))$. Thus $\lambda(y) = 1$ gives rise to the null or uniform model on the set of permutations. In the class of models considered here, the simplest non-null potential function is a sum of pairwise effects or first-order inversions. More complicated potential functions are sums over the items taken three at a time. Thus the higher-order models contain the lower-order models as sub-models. In this way, the total sum of squares or total deviance can be decomposed into first-order effects, second-order effects and so on. This decomposition is very similar to the decomposition of the total sum of squares for factorial models into main effects, first-order interactions, and so on. Unlike Diaconis's [5, 6] decomposition, our decomposition is not in any sense complete or maximal.

The first-order model is sometimes called the Babington Smith model: see Babington Smith [1] or Mallows [10]. The higher-order models and some of their sub-models seem to be new. A quite different class of models based on rankings is described by Critchlow [3]. Further probability models,

[1] Department of Statistics, University of Chicago

including the Babington Smith model, are discussed in Critchlow, Fligner and Verducci [4].

In the analysis of the APA election data, the first-order model is found to be unsatisfactory. A more satisfactory fit is found using a sub-model of the second-order inversion model. This model has a potential function that is a sum of pairwise potentials proportional to the unsigned rank difference of each pair of candidates. High positive coefficients identify pairs that are usually ranked far apart: high negative coefficients indicate pairs that tend to be ranked adjacent. In the context of the APA election data, where candidates and voters tend to be either academicians or clinicians, this model makes sense. Most voters tend to rank the academician candidates together, whether high or low: similarly for the clinical candidates.

11.2 Models for Random Permutations

Notation

Suppose for the moment that four contestants, here denoted by the letters a, b, c, d are ranked independently by each of n judges. For clarity of exposition in this section, it is helpful to assume that the n judges form a homogeneous set, so that the probabilities for the 4! rankings are the same for each judge. In subsequent sections, we deal with the case where it is required to compare two or more groups of judges or, more generally, to take account of covariates measured on each of the judges.

In the case of a homogeneous group of n judges, the data comprise n independent and identically distributed observations, here denoted by $y_1, \ldots, y_n$, where each y_i is one of the 4! permutations of *abcd*. Thus, the observations are non-numerical. If n is large, it is often more convenient to exhibit the data in a condensed form as 24 counts, one for each of the 4! permutations of *abcd*. In other words, we write $(y_{(1)}, w_1), \ldots (y_{(4!)}, w_{4!})$, where $y_{(1)}, \ldots, y_{(4!)}$ are the distinct permutations in alphabetical order and $w_1, \ldots, w_{4!}$ are the associated counts. In either case, the ys are non-numerical. In what follows, the data can be taken in either form.

For convenience of notation, we denote by 1, the permutation $abc\ldots$ in standard order. Probabilities are denoted by $\pi(y)$ or by $\pi(abdc)$ when we wish to refer to a particular permutation. As usual, when dealing with probabilities, it is more convenient and natural to work on the logistic scale, and so we write

$$\lambda(y) = \log\{\pi(y)/\pi(1)\}$$

for the relative log odds in favour of y over 1. Thus, $\lambda(1) = 0$ and the remaining $4!-1$ logits are unconstrained. The probabilities may be recovered from the logits via the expression

$$\pi(y) = \frac{\exp\{\lambda(y)\}}{\sum_j \exp\{\lambda(j)\}},$$

where the sum in the denominator runs over all 4! permutations.

Our aim then is to construct a useful class of non-null probability distributions on the set of permutations. To do so, we express $\lambda(y)$ linearly in terms of 'first-order inversions', 'second-order inversions', and so on, in much the same way that a class of multivariate discrete distributions can be expressed in terms of independence, conditional independence and so on. In other words, the absence of second- and higher-order inversions has much the same meaning as the absence of two-factor interaction in a log-linear model. In a similar manner, the absence of third- and higher-order inversions has much the same meaning as the absence of three-factor interaction in a log-linear model for the joint distribution of three or more discrete response variables.

The first step in this program is to describe what we mean by inversions of higher order.

INVERSIONS

An inversion in a permutation is measured relative to an agreed standard order, here taken to be alphabetical. Two letters, not necessarily adjacent in the sequence, are either in standard order, for example *bc*, or else inverted, *cb*. Three letters may occur in any of six orders, each order associated with one or more of the three inversions, *ba*, *cb*, *ca*, as shown in Table 1. Such inversions, involving only two letters at a time are called 'first-order inversions' in the remainder of this paper. Every permutation is uniquely identified by its list of first-order inversions, but there are lists of first-order inversions for which no permutation is possible, for example *cb*, *ba* omitting *ca*, in a list of three.

In addition to the three first-order inversions of three letters, we now introduce the notion of a second-order inversion. A second-order inversion is an ordered triple of letters, not necessarily adjacent in the sequence, none of which occurs in its natural position relative to the other two. Thus, there are exactly two second-order inversions of *abc*, namely *bca* and *cab* as shown in Table 1. In the case of four letters, there are two second-order inversions for each subset of three letters taken from the four available. The list of 8 second order inversions so constructed is given in Table 2.

Table 1: Incidence matrix of inversions of *abc*.

Permutation	*First-order inversions*			*Second-order inversions*	
	cb	*ba*	*ca*	*bca*	*cab*
abc					
acb	+				
bac		+			
bca		+	+	+	
cab	+		+		+
cba	+	+	+		

By extension, a third-order inversion is an ordered sequence of four letters, none of which occurs in its natural relative position in the sequence. Thus, *badc* and *bcda* are two third-order inversions. Oddly, *badc* contains no three-letter sequences that are second-order inversions, whereas *bcda* contains *bca*, *bda*, and *cda*. The third-order inversion *cdab* contains four second-order inversions. Table 2 gives a complete list of all inversions of *abcd*.

Table 2: Inversions of *abcd*

Order	*No.*	*Inversions*
1	6	*dc*, *cb*, *db*, *ba*, *ca*, *da*
2	8	*cdb*, *dbc*, *bca*, *bda*, *cda*, *dac*, *cab*, *dab*
3	9	*badc*, *bcda*, *bdac*, *cadb*, *cdab*, *cdba*, *dabc*, *dcab*, *dcba*

According to this decomposition of effects, for k letters there are $\binom{k}{2}$ first-order inversions, one for each distinct pair of letters. For each subset of three letters chosen from the k available, there are two second-order inversions as shown in Tables 1 and 2. This gives a total of $2\binom{k}{3}$ second-order inversions. By extension of this argument, it is easily seen that there are $9\binom{k}{4}$ third-order inversions.

The general expression for the number of inversions of order $j - 1$, involving exactly $j \geq 2$ of the k letters, is

$$I_{kj} = \binom{k}{j} j! \left(\frac{1}{2!} - \frac{1}{3!} + \cdots \pm \frac{1}{j!} \right)$$

which may be derived by extension of the argument in Feller (1968, p.101). The same formula gives the number of permutations in which exactly $k - j$ of the items occur in their 'home' position. Summation from $j = 2$ to $j = k$ gives

$$\sum_{j=2}^{k} I_{kj} = \sum_{j=2}^{k} \sum_{i=2}^{j} \binom{k}{j} j! \, (-1)^i / i!.$$

By transformation of variables, we find

$$\sum_{j=2}^{k} I_{kj} = k! - 1.$$

In other words, the total number of inversions of all orders is exactly $k! - 1$, the same as the number of independent probabilities required to specify an arbitrary probability distribution over the set of permutations.

For large k, the great preponderance of inversions are of near maximal order since

$$I_{kk} \simeq I_{k,k-1} \simeq k!\, e^{-1}$$
$$I_{k,k-j} \simeq k!\, e^{-1}/j! \quad for \quad j \geq 1.$$

In other words, roughly 74% of the inversions are of orders $k - 1$ and $k - 2$, while 98% are of order $k - 4$ or more. These approximations are quite accurate even for $k = 4$ as can be seen from Table 2.

Incidence Matrices

By an incidence matrix is meant a matrix $\mathbf{R}$ of zeros and ones, whose rows are indexed by observations or permutations and whose columns are indexed by effects or inversions. Such a matrix is given in Table 1 for $k = 3$. This incidence matrix differs from the incidence matrix for a factorial design in three important respects. First, the constant vector or column of ones is missing, though if we had chosen to work with log probabilities rather than logits, it would have been essential to include the constant vector in $\mathbf{R}$. This augmented matrix is henceforth denoted by $\mathbf{R}^*$. Second, the 'main effects' or first-order inversions are not orthogonal as they are in the case of balanced factorial designs. Third, the 'interactions' or second-order inversions cannot be obtained by elementwise multiplication of pairs of main-effect vectors. Despite these obvious differences, the analogy with main effects and interactions in factorial models is useful and helpful for understanding and interpreting effects.

We use the terms *first-order incidence matrix, second-order incidence matrix* and so on, meaning that only selected inversions up to the stated order are included in $\mathbf{R}$. Thus, a complete first-order incidence matrix is of order $k! \times \binom{k}{2}$; a complete second-order incidence matrix of order $k! \times \{\binom{k}{2} + 2\binom{k}{3}\}$, and so on.

It is by no means obvious that the incidence matrices so constructed have full rank equal to the number of effects included. I have checked the rank condition for $k \leq 4$ and in all cases I find that the augmented $k! \times k!$ incidence matrices have unit determinant. In fact, by a suitable re-arrangement of rows and columns, $\mathbf{R}^*$ may be reduced to lower triangular form with unit values along the diagonal. I believe that these properties must extend to complete incidence matrices of all orders, though I have been unable to prove this conjecture.

Factorial Models

Linear models for the logits of the permutation probabilities may be constructed by writing

$$\boldsymbol{\lambda} = \mathbf{R}\boldsymbol{\rho} \tag{1}$$

for a suitably chosen incidence matrix $\mathbf{R}$ and coefficients $\boldsymbol{\rho}$. We use $\mathbf{R}$ and $\boldsymbol{\rho}$ here rather than the more familiar $\mathbf{X}$ and $\boldsymbol{\beta}$ because the effects involved are *internal* to a single multinomial response vector. In keeping with standard statistical usage, $\mathbf{X}$ and $\boldsymbol{\beta}$ are reserved for describing differences between groups or populations. The latter contrasts are sometimes said to be *external.* Stated in an equivalent way, (1) implies that the vector $\boldsymbol{\lambda}$ is required to lie in the column space of the chosen incidence matrix. The trivial null distribution is obtained by taking $\mathbf{R}$ to be degenerate, or equivalently, $\boldsymbol{\rho} = \mathbf{0}$. This gives $\boldsymbol{\lambda} = \mathbf{0}$ and $\pi(y) = 1/k!$ for each permutation y.

More interestingly, if $\mathbf{R}$ is any first-order incidence matrix that includes the inversion ba, then (1) implies that

$$\log\left(\frac{\pi(*_1\, ba\, *_2)}{\pi(*_1\, ab\, *_2)}\right) = \lambda(*_1\, ba\, *_2) - \lambda(*_1\, ab\, *_2) = \rho_{ba}. \tag{2}$$

In the above, and in subsequent expressions, $*_1$ and $*_2$ are so-called 'wildcard' characters that match any string, possibly degenerate. All occurrences of $*_1$ and $*_2$ in one equation refer to the same sets of strings. In particular, for $k = 3$, any first-order model including the inversions ba and ca implies that

$$\begin{aligned}\log\left(\tfrac{\pi(cba)}{\pi(cab)}\right) &= \log\left(\tfrac{\pi(bac)}{\pi(abc)}\right) = \rho_{ba}\\ \log\left(\tfrac{\pi(bca)}{\pi(bac)}\right) &= \log\left(\tfrac{\pi(cab)}{\pi(acb)}\right) = \rho_{ca}\end{aligned} \tag{3}$$

and so on. These conclusions are independent of whether the inversion cb is included in $\mathbf{R}$ or not.

If a and b are not adjacent, switching the two letters triggers inversions involving the intervening letters. For example, if c intervenes between a and b we have

$$\log\left(\frac{\pi(*_1\, bca\, *_2)}{\pi(*_1\, acb\, *_2)}\right) = \lambda(*_1\, bca\, *_2) - \lambda(*_1\, acb\, *_2) = \rho_{ba} + \rho_{ca} - \rho_{cb},$$

as can be seen from the first-order incidence matrix in Table 1.

The justification for drawing a strong analogy between the first-order inversion model and a model of 'no interaction' is the following. Given that a and b are adjacent, the odds that b occurs first is $\exp(\rho_{ba})$, independently of the position of c. Similarly, the odds that c occurs before a is $\exp(\rho_{ca})$ independently of whether b occurs first or last position. Thus, the first-order inversion model does indeed imply the absence of interaction of an easily understood type.

By extension, in the case of a second-order inversion model, we have

$$\log\left(\frac{\pi(*_1\, bac *_2)}{\pi(*_1\, abc *_2)}\right) = \text{const}(*) + \rho_{ba},$$

$$\log\left(\frac{\pi(*_1\, cba *_2)}{\pi(*_1\, cab *_2)}\right) = \text{const}(*) + \rho_{ba} - \rho_{cab},$$

again independently of the two sequences $*_1$ and $*_2$. The constant term may depend on $*_1$ and $*_2$, but the same constant occurs in both cases. Stated in another way, the prior occurrence of c reduces the odds of b preceding a by the factor $\exp(\rho_{cab})$.

The occurrence of second-order inversions may be interpreted as heterogeneity among judges of the type that we might expect among voters in elections. To take a drastically over-simplified example, suppose that in a constituency for which there is a number of seats available, voters are required to rank the three candidates, Messrs Left, Centre and Right, in decreasing order of preference. One would expect that voters who rank Mr Left in first place would tend to prefer Mr Centre to Mr Right and that those who rank Mr Left in third place would prefer the reverse order. This is a simple instance of a second-order inversion explained by heterogeneity among the voters.

Marginality

The log odds ratio

$$\log\left(\frac{\pi(cba)\,\pi(abc)}{\pi(cab)\,\pi(bac)}\right) = \log\left(\frac{\pi(cba)}{\pi(cab)}\right) - \log\left(\frac{\pi(bac)}{\pi(abc)}\right) = -\rho_{cab} \quad (ba \text{ vs. } c)$$

is a measure of the change in the ba effect explained by the position of c. It would not normally make sense, therefore, to include in $\mathbf{R}$, the second-order inversion, cab, without the ba inversion. The interpretation given to a non-zero value of ρ_{cab} implies the existence of a ba effect. We say that ba is marginal to cab in the sense of Nelder [11]. See also McCullagh and Nelder [9], (section 3.5). The usual marginality constraint on linear models is that every term included in a model must be accompanied by all marginal terms. Wilkinson and Rogers's [12] model-formula notation for the specification of linear models automatically enforces the marginality conditions, at least for factorial models. In this section we describe the marginality conditions appropriate for models based on inversions. A model formula notation is developed to enforce these conditions in an automatic way.

The remaining second-order contrasts,

$$\log\left(\frac{\pi(acb)\,\pi(bca)}{\pi(abc)\,\pi(cba)}\right) = \rho_{bca} \qquad (cb \text{ vs. } a)$$

$$\log\left(\frac{\pi(cab)\,\pi(bac)}{\pi(acb)\,\pi(bca)}\right) = \rho_{cab} - \rho_{bca}, \qquad (ca \text{ vs. } b)$$

are interpreted as the change in the *cb* effect explained by the position of *a*, and the change in the *ca* effect explained by the position of *b* respectively. If the *bca* inversion were excluded from the model then the interpretation of ρ_{cab} would be ambiguous: it could be taken either as the (*ba* vs. *c*) effect or as the (*ca* vs. *b*) effect. If the above interpretations as interactions are to be maintained, it follows that the second-order inversions *bca* and *cab* must be included as a pair. Furthermore, if this pair of second-order inversions is included in **R** all three marginal second-order inversions must also be included.

For $k \geq 4$ the above argument can be applied directly to each subset of three items or letters. By, extension all nine third order inversions involving a given set of four letters must be handled en bloc and not teased apart except, perhaps, for interpretation.

To summarize, therefore, for $k = 3$, there are nine models that satisfy the marginality conditions (Fig. 1). In the remainder of this paper, such models are called factorial although, in the literature on discrete data, hierarchical is unfortunately sometimes used for the same purpose. The notation adopted in Fig. 1 is such that AB denotes the space spanned by the constant and the *ba* inversion. By the marginality convention, ABC includes the inversions *bca* and *cab* as well as the three marginal first-order inversions and the constant vector. The order of appearance of the letters is therefore immaterial, so no distinction is made between AB and BA, nor between ABC, BAC, CAB,... . These factors refer to the same spaces, though they might use different basis vectors. With this notation a factor with r upper-case letters has $r!$ levels, one for each permutation of the letters.

The number of factorial models is considerably less than the $2^5 = 32$ possibilities available if the marginality conditions are ignored. Among the factorial models, there is a partial ordering based on the relationship of nesting or sub-model. For instance, the null model (0), is a sub-model of *ca*, which, in turn, is a sub-model of both $cb + ca$ and $ba + ca$. The lattice diagram in Figure 1 depicts all the order relationships among the various factorial models for $k = 3$.

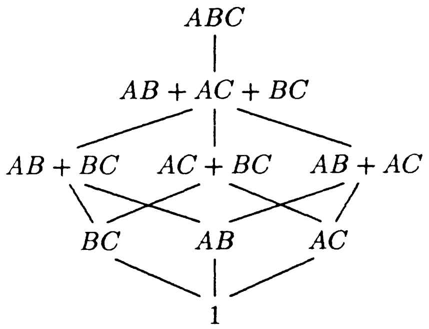


Fig. 1: Lattice diagram of factorial models for $k = 3$.

In the case of four letters, the number of factorial models is considerably larger. So far as I can determine, there are 114 models of 16 different types (20 including sub-types). A part of the lattice, which has 12 rows in all, is shown in Fig. 2.

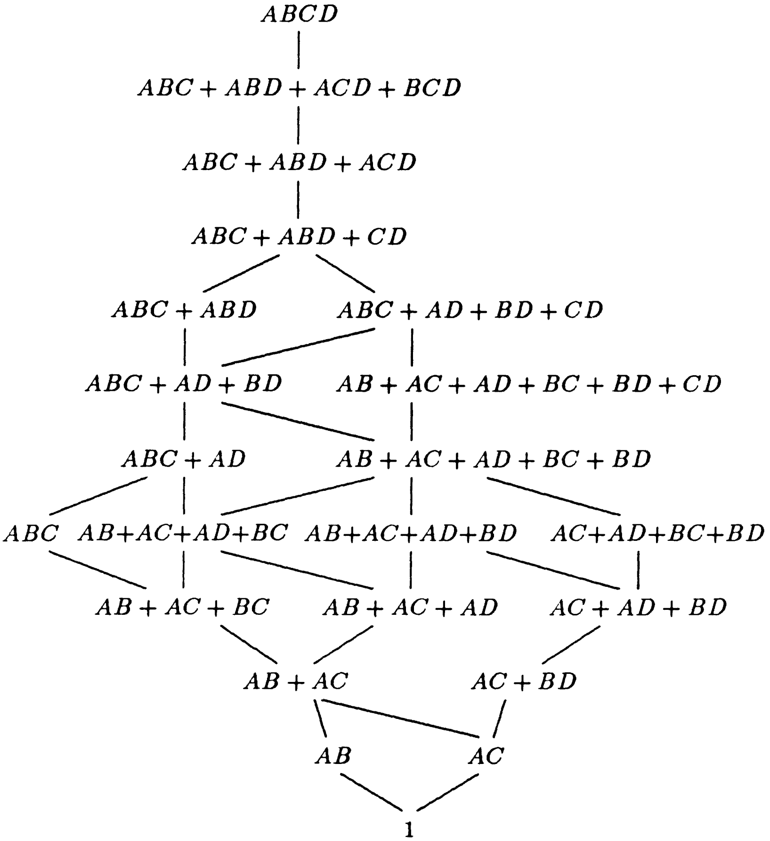


Fig. 2: Part of lattice diagram of factorial models for $k = 4$.

For instance, in addition to the equi-probable null model, there are $2^6 - 1 = 63$ first-order inversion models obtained by selecting up to 6 of the first-order inversions. In addition to these, there are models such as $ABC + BCD + AD$, meaning that all first- and second-order inversions of a, b, c, all first- and second-order inversions of b, c, d as well as the first-order inversion da are included. Of these 114 factorial models, only 17 include all six first-order inversions.

I have been unable to determine, for an arbitrary number of letters, the number of models that satisfy the marginality constraints. Evidently, the number increases rapidly with k. The counting problem is essentially the same as determining the number of factorial models that, in the notation of Wilkinson and Rogers [12], contain no singleton letters, i.e. the number of elements in the free distributive lattice on k generators that contain no singleton letters.

11.3 Sufficient Statistics and Log-linear Models

In this section we apply some of the ideas discussed in the previous section to help analyse the APA election data (Diaconis [6]) from a new perspective. We begin with an examination of the sufficient statistic for the complete first- and second-order inversion models. This process helps to pinpoint the major sources of variation in these data, and serves as a guide in formulating a suitable model.

Sufficient Statistics

In this section we consider initially log-linear models in which the incidence matrix $\mathbf{R}$ contains all inversions up to a given order, say $d \leq k$, and no higher-order inversions. Such models are invariant under item re-labelling, but they are not the only useful models with that property. If $w_1, \ldots, w_{k!}$ are the observed counts for the various permutations, then $\mathbf{R}^T\mathbf{W}$ is the sufficient statistic for the model under consideration. For example, if $k = 5$ and $\mathbf{R}$ is the first-order incidence matrix, $\mathbf{R}^T\mathbf{W}$ may be presented as a $k \times k$ matrix in which the (i, j) entry is the number of times that item i received a lower rank than item j. In the case of the APA election data

(Diaconis [6]), we have

$$\mathrm{S} = \begin{array}{ccccccc} & a & b & c & d & e & \text{Total} \\ a & -- & 3318 & 2897 & 3129 & 3053 & 12397 \\ b & 2420 & -- & 2593 & 2853 & 2711 & 10577 \\ c & 2841 & 3145 & -- & 3031 & 2935 & 11952 \\ d & 2609 & 2885 & 2707 & -- & 2745 & 10946 \\ e & 2685 & 3027 & 2803 & 2993 & -- & 11508 \\ & & & & & & \\ \text{Tot} & 10555 & 12375 & 11000 & 12006 & 11444 & \end{array} \tag{4}$$

Note that in all cases $S_{ij} + S_{ji} = 5738$, the total number of votes cast. Furthermore, $S_{i\cdot}$ is the linear score for candidate i, with a score of 4 for first place, 3 for second place and so on, the score being equal to the number of candidates beaten. Thus popular candidates tend to have large values of $S_{i\cdot}$: unpopular candidates have low values. What is remarkable about the APA election is that the entries in the above table are so nearly equal for all candidates. Candidate a has the highest score, but only by a very small margin. This first-order analysis is more or less consistent with the hypothesis that all votes were cast at random and that the voters have no strong preferences among the candidates. As we shall see, however, this simple hypothesis is not supported by a second-order analysis, which examines the structure of the data in finer detail.

For the second-order model, the sufficient statistic may be presented as a $k \times k \times k$ table, each margin indexed by the candidates in the same order. Thus S_{ijl} gives the total number of votes in which candidates i, j and l were ranked in the specified order. Evidently, cell (i, j, l) is empty if any pair of indices is equal, leaving 60 non-empty entries among the 125 cells for $k = 5$. It is difficult to find a satisfactory way to present a 3-way symmetrically indexed array, particularly when, as here, more than half the cells are empty. Ideally we would like a format in which all two-way and one-way marginal tables are equally apparent, but I have been unable to devise anything better than Tables 3 and 4 below.

Table 3: *Three-way table of counts for APA data*

Candidate	A	B	C	D	E	Total
* * A						* + A
A	—	—	—	—	—	—
B	—	—	665	709	717	2091
C	—	803	—	630	681	2114
D	—	951	758	—	925	2634
E	—	870	813	903	—	2586
+ * A	—	2624	2236	2242	2323	9425
* * B						* + B
A	—	—	969	985	1059	3013
B	—	—	—	—	—	—
C	1373	—	—	950	1029	3352
D	949	—	707	—	954	2610
E	1098	—	854	1066	—	3018
+ * B	3420	—	2530	3001	3042	11993
* * C						* + C
A	—	976	—	728	799	2503
B	952	—	—	772	805	2529
C	—	—	—	—	—	—
D	1221	1228	—	—	1074	3523
E	1191	1144	—	1009	—	3344
+ * C	3364	3348	—	2509	2678	11899
* * D						* + D
A	—	1384	948	—	1233	3565
B	760	—	593	—	920	2273
C	1453	1488	—	—	1264	4205
D	—	—	—	—	—	—
E	857	1007	720	—	—	2584
+ * D	3070	3879	2261	—	3417	12627
* * E						* + E
A	—	1161	907	1039	—	3107
B	833	—	644	926	—	2403
C	1347	1262	—	1047	—	3656
D	781	865	624	—	—	2270
E	—	—	—	—	—	—
+ * E	2961	3288	2175	3012	—	11436
First marginal table: **+						* + +
A	—	3521	2824	2752	3091	12188
B	2545	—	1902	2407	2442	9296
C	4173	3553	—	2627	2974	13327
D	2951	3044	2089	—	2953	11037
E	3146	3021	2387	2978	—	11532
+ * +	12815	13139	9202	10764	11460	57380
Second marginal table: *+*						* + +
A	—	3013	2503	3565	3107	12188
B	2091	—	2529	2273	2403	9296
C	2114	3352	—	4205	3656	13327
D	2634	2610	3523	—	2270	11037
E	2586	3018	3344	2584	—	11532
+ + *	9425	11993	11899	12627	11436	57380
Third marginal table: +**						+ * +
A	—	3420	3364	3070	2961	12815
B	2624	—	3348	3879	3288	13139
C	2236	2530	—	2261	2175	9202
D	2242	3001	2509	—	3012	10764
E	2323	3042	2678	3417	—	11460
+ + *	9425	11993	11899	12627	11436	57380

Table 4: *Three-way table of partial residuals for APA data*

Candidate	A	B	C	D	E	Total
**A						*+A
A	—	—	—	—	—	—
B	—	—	−138.0	−105.8	−71.3	−315.1
C	—	−58.0	—	−271.3	−214.5	−543.8
D	—	199.4	−104.6	—	115.6	210.3
E	—	59.5	−94.9	19.2	—	−16.1
+*A	—	200.9	−337.5	−357.9	−170.2	−664.7
**B						*+B
A	—	—	−138.0	−105.8	−71.3	−315.1
B	—	—	—	—	—	—
C	196.0	—	—	−94.9	−41.8	59.3
D	−93.6	—	−300.8	—	−16.8	−411.1
E	11.8	—	−205.8	40.5	—	−153.5
+*B	114.2	—	−644.5	−160.3	−129.9	−820.4
**C						*+C
A	—	−58.0	—	−271.3	−214.5	−543.8
B	196.0	—	—	−94.9	−41.8	59.3
C	—	—	—	—	—	—
D	375.9	395.7	—	—	209.2	980.8
E	309.4	247.5	—	73.9	—	630.8
+*C	881.3	585.2	—	−292.3	−47.1	1127.0
**D						*+D
A	—	199.4	−104.6	—	115.6	210.3
B	−93.6	—	−300.8	—	−16.8	−411.1
C	375.9	395.7	—	—	209.2	980.8
D	—	—	—	—	—	—
E	−134.8	−23.7	−283.1	—	—	−441.6
+*D	147.5	571.4	−688.5	—	308.0	338.4
**E						*+E
A	—	59.5	−94.9	19.2	—	−16.1
B	11.8	—	−205.8	40.5	—	−153.5
C	309.4	247.5	—	73.9	—	630.8
D	−134.8	−23.7	−283.1	—	—	−441.6
E	—	—	—	—	—	—
+*E	186.3	283.4	−583.8	133.7	—	19.6
First marginal table: **+						*++
A	—	200.9	−337.5	−357.9	−170.2	−664.7
B	114.2	—	−644.5	−160.3	−129.9	−820.4
C	881.3	585.2	—	−292.3	−47.1	1127.0
D	147.5	571.4	−688.5	—	308.0	338.4
E	186.3	283.4	−583.8	133.7	—	19.6
+*+	1329.0	1641.0	−2254.0	−676.7	−39.2	0.0
Second marginal table: *+*						*++
A	—	−315.1	−543.8	210.3	−16.1	−664.7
B	−315.1	—	59.2	−411.1	−153.5	−820.4
C	−543.8	59.3	—	980.8	630.8	1127.0
D	210.3	−411.1	980.8	—	−441.6	338.4
E	−16.1	−153.5	630.8	−441.6	—	19.6
++*	−664.7	−820.4	1127.0	338.4	19.6	0.0
Third marginal table: +**						+*+
A	—	114.2	881.3	147.5	186.3	1329.0
B	200.9	—	585.2	571.4	283.4	1641.0
C	−337.5	−644.5	—	−688.5	−583.8	−2254.0
D	−357.9	−160.3	−292.3	—	133.7	−676.7
E	−170.2	−129.9	−47.1	308.0	—	−39.2
++*	−664.7	−820.4	1127.0	338.4	19.6	0.0

Table 3 shows that of the $N = 5738$ votes cast, 665 had candidates a, b and c in the order bca, whereas 803 had them in the order cba. With some effort on the part of the reader one can begin to unravel some of the strong effects that are present in this three-way table. For a start, the counts in the table range from 593 to 1488, a factor of 2.5. Second, all of the counts in the column labelled 'C' are less than the average of 956. On the other hand, the counts in the rows labelled 'C' are mostly larger than average, though this effect is not so easy to detect because of the way that the table has been laid out.

The three two-dimensional marginal tables of counts are also shown. These are labelled $**+$, $*+*$ and $+**$, with $+$ indicating the index that has been summed out. Each marginal table is a weighted version of the first-order sufficient statistic, with weights depending on the ranks of the two candidates. For example, in the permutation $edcba$, the inversion ec receives weights of 2, 1, and 0 in the three marginal tables. These weights are equal to the number of items that can be substituted for the '+'. The one-dimensional $*++$-table gives a weighted ranking of the candidates, with weights $(12, 6, 2, 0, 0)$ for the various ranks. Similarly, the $*++$-table gives a weighted ranking of the candidates, with weights $(0, 6, 8, 6, 0)$. The weights are the number of ordered pairs that can be substituted for the two '+'s: in both cases the weights are quadratic functions of the ranks.

Table 4 is the result of an attempt to remove some of the clutter from Table 3. In addition to removing the overall mean, all first-order effects have been removed by regression. In other words, Table 4 has been constructed in the same way as Table 3, but instead of starting with the raw counts I have used the residual counts from the first-order inversion model. The residuals have not been standardized in any way. As a result, the table now contains only 30 distinct numbers, reverse permutations being equal. In addition, cyclic sums of three are zero, so there are effectively only 20 linearly independent combinations in the Table. We now see that all of the entries in the column labelled 'C', and most of the large values in column 'D' are negative. Conversely, most of the large positive values occur in columns 'A' and 'B'.

In the marginal $*++$ and $+*+$-tables, candidate 'C' has the largest totals, in the first case positive and in the second negative. In other word candidate 'C' improves his relative position when greater weight is given to higher positions in the ranking. The negative value in the $+*+$-table shows that few voters ranked 'C' in the middle. A positive value for the (i, j)-entry in the marginal $**+$-table indicates that candidates i and j tend to be ranked as ij when both are ranked high, and as ji when both are ranked low.

If we denote by M_1, M_2, M_3, the three two-dimensional marginal tables it can be seen that $M_3 = M_1^T$, $M_2 = M_2^T$, and $M_2 + M_3 = -M_1$. Thus all the information in the three two-dimensional marginal tables resides in the symmetric and asymmetric parts of M_1, which are the upper and lower triangles of the matrix below.

$$\begin{array}{cccccc} & A & B & C & D & E \\ A & - & 315.1 & 543.8 & -210.3 & 16.1 \\ B & -86.7 & - & -59.2 & 411.1 & 153.5 \\ C & 1218.8 & 1229.7 & - & -980.8 & -630.8 \\ D & 505.4 & 731.7 & -396.2 & - & -441.6 \\ E & 356.5 & 413.3 & -536.7 & -174.3 & - \end{array} \tag{5}$$

Each entry in the above table is associated with a pair of candidates. The value in cell (i, j) is a weighted sum of residuals, with weights depending on the ranks of i and j, here denoted by $r(i)$ and $r(j)$. In the symmetric part (upper triangle) of the table, the weight is $k - \max(r(i), r(j))$. Since the residuals are orthogonal to $r(i)$, this weight is equivalent to $-\frac{1}{2}|r(i) - r(j)|$, i.e. minus one half the rank difference between the two candidates. In the asymmetric part, the weight is $\{k - \max(r(i), r(j))\}\text{sign}(r(j) - r(i))$, i.e. the interaction of the symmetric part with the (i, j) inversion. Since the residuals are orthogonal to $r(i)$, this weight is equivalent to $-\frac{1}{2}(r(i) + r(j))\text{sign}(r(i) - r(j))$.

It follows then that large positive values in the upper triangle indicate pairs who tend to have similar, or adjacent, ranks. Conversely, large negative values indicate pairs that tend not to be ranked together. Thus there is a tendency for c and d, and also c and e not to be ranked together. Conversely, ab and ac tend to be given similar rankings. These second-order unordered effects are much larger than any first-order effects, and are reminiscent of the 'unordered pairs subspace' in Diaconis's [5, 6] analysis. Large positive values in the asymmetric lower triangle of the table indicate pairs whose order tends to be reversed when their average rank is increased. Thus, the orders ca, cb, ce and db are favoured when the average rank is low, but the opposite orders are favoured when the average rank is high. This effect is again reminiscent of Diaconis's ordered pairs subspace.

Models and their Interpretation

Armed with the information we have obtained from the tables in the previous section, it is helpful to clarify aspects of our interpretation by fitting a sequence of well-chosen models. It should be abundantly clear that the first-order inversion model does not fit the APA data. This model gives a deviance of 1527.9 on 109 degrees of freedom, which is only a slight reduction over the uniform null model. The second-order model does considerably better, giving a residual deviance of 246.5 on 89 degrees of freedom. In a context such as this, where voters almost certainly do not vote independently, we should expect over-dispersion relative to the Poisson or multinomial distributions. If we accept, at least tentatively, the second-order model, our dispersion factor is estimated as 2.77, not an unrealistic figure for data of this sort.

The majority of the second-order inversions are quite large in absolute terms. For example the estimated inversion coefficients for b, c, d in the second-order model are

Parameter	*Estimate*	*Nominal s.e.*
cb	0.7768	0.080
db	−0.9210	0.104
dc	−0.2438	0.092
cdb	0.3580	0.090
dbc	0.9224	0.085

Thus, by way of example, the fitted probabilities satisfy

$$\log\left(\frac{\hat{\pi}(*_1\, cdb\, *_2)}{\hat{\pi}(*_1\, dcb\, *_2)}\right) = \log\left(\frac{\hat{\pi}(*_1\, bcd\, *_2)}{\hat{\pi}(*_1\, bdc\, *_2)}\right) + 0.3580,$$

corresponding to an odds ratio of 1.430.

It should be clear from our examination of the sufficient statistics in the previous section that both the first-order and the second-order inversion models have interesting sub-models whose sufficient statistics are the marginal tables in (4) and (5). For example, the Bradley-Terry model has as its sufficient statistic the vector of rank sums, which form the row totals of **S** in (3.1). By extension, the second-order inversion model has a sub-model whose sufficient statistic is (4) together with the symmetric half of (5).

In order to simplify the discussion of the various sub-models, it is helpful to introduce suitable notation. First, R_a denotes the rank of a; R_{ab} denotes the average rank of a and b. These are both $k! \times 1$ vectors. Similarly $D_{ab} = |R_a - R_b|$, the unsigned rank difference. These quantities occur as weights in the formation of sufficient statistics in (4) and (5). In what follows, we use standard model-formula notation augmented by the bracket summation convention. Thus,

$$R_a[5] = R_a + R_b + R_c + R_d + R_e$$

is a model formula in which the model matrix of order $5! \times 5$ specifies the rank vectors for the five candidates. Since the sufficient statistic for this model is the vector of rank sums, it follows that $R_a[5]$ is equivalent to (a reparameterization of) the Bradley-Terry model. Likewise $AB[10]$ denotes the complete first-order inversion model, also called the Babington Smith model. Table 6 below shows the residual deviances for selected sub-models of the second-order inversion model. The fit of the third-order inversion model is also given for comparative purposes.

Table 6: *Analysis of deviance for a sequence of models fitted to the APA election data*

	Residual		*Reduction*		*Mean*
Model	*deviance*	*d.f.*	*deviance*	*d.f.*	*Deviance*
1	1717.5	119			
$R_a[5]$	1566.3	115	151.2	4	37.8
$AB[10]$	1527.9	109	38.4	6	6.4
$+ D_{ab}[10]$	269.4	100	1258.5	9	139.8
$+ AB.R_{ab}[10]$	249.2	94	20.2	6	3.4
$ABC[10]$	246.5	89	2.7	5	0.6
$ABCD[5]$	58.8	44	187.7	45	4.2
$D_{ab}[10]$	431.6	110			
$AB + D_{ab}[10]$	290.5	109	141.1	1	141.1

The model $D_{ab}[10]$ accounts for a very large fraction of the total deviance, almost as much as the complete second-order inversion model, but using only 10 parameters rather than 30. This model asserts that the log probability of any permutation is a sum of pairwise potentials, one potential for each pair of candidates regardless of their order in the permutation. The potentials are proportional to the unsigned rank difference, with a coefficient that depends on the pair of candidates. It follows then that reversed permutations have equal probability according to this model. The estimated coefficients in the model $AB + D_{ab}[10]$, with the single inversion AB added, are shown in Table 7. By way of illustration, the fitted log probability of the permutation *abcde* is

$$\begin{aligned}\log \pi(abcde) &= 1.242 + 0.316 + 0.036 \times 1 - 0.039 \times 2 + 0.180 \times 3 \\ &\quad +0.106 \times 4 + 0.137 \times 1 - 0.018 \times 2 + 0.101 \times 3 \\ &\quad +0.388 \times 1 + 0.286 \times 2 + 0.000 \times 1 \\ &= 3.844.\end{aligned}$$

The fitted probability for the reverse permutation is computed in the same way except that the *ab* inversion is omitted. Apart from the effect of the *ab* inversion, pairs for which the potential coefficient is high tend to be well separated. Conversely, pairs having low or negative potential coefficients tend to be adjacent in the rankings. Note that any constant could be added to each of the potential coefficients without altering the fitted model. The convention employed here is to set the coefficient of D_{de} to zero, although a zero sum constraint would make it slightly easier to distinguish the attractive potentials from the inhibitory ones.

Table 7: *Parameter estimates in the model* $AB + D_{ab}[10]$

Parameter	*Estimate*	*Parameter*	*Estimate*
1	1.242	D_{bc}	0.137
AB	0.316	D_{bd}	−0.018
D_{ab}	0.036	D_{be}	0.101
D_{ac}	−0.039	D_{cd}	0.388
D_{ad}	0.180	D_{ce}	0.286
D_{ae}	0.106	D_{de}	0.000

These conclusions are similar to, but are not entirely in agreement with the informal analysis used in the previous section. In particular, the asymmetric part of the matrix (5) strongly suggests that the interaction term $AB.R_{ab}[10]$ should be rather substantial, but the model-fitting analysis shows that the term has little effect. Thus, it appears that the large values in the lower triangle of (5) must be caused by the confounding effect of the symmetric part of the table.

11.4 Conclusions

This paper began by considering models based solely on inversions of various orders. By examining the sufficient statistics produced by these inversion models, we are led to consider a variety of sub-models that turn out in some cases to be more interesting and more readily interpretable than the inversion models themselves. The Bradley-Terry model, which has a simple interpretation in terms of consensus and linear ordering, is the most obvious example of this phenomenon. It is a sub-model of the first-order inversion model whose sufficient statistic is the column of marginal totals of $\mathbf{S}$ in (4).

Unfortunately, in the case of the APA election data, none of these first-order models fit the data, even approximately. The second-order inversion model is more successful in the sense that it fits the data reasonably well, but the parameters are not especially easy to interpret. By examining the marginal tables of the sufficient statistic, we were led to consider a sub-model based on undirected distance-related pairwise potentials, namely the model $D_{ab}[10]$. This 10-parameter model, which is invariant under relabelling, accounts for 75% of the total deviance. Of the first-order inversions, only AB makes a noticeable improvement to the fit. The parameter estimates in Table 7 show that there is a tendency for a and b to occur together, and a strong tendency for a to precede b. It is as if b were the junior partner or running mate of a. The largest inhibitory effects are c

versus *d* and *e*. It is as if *c* were at the opposite end of the psychological/political spectrum from *d* and *e*.

It is undoubtedly true that our final model leaves some small higher-order systematic effects unaccounted for. Indeed, it can be sen from Table 6 that the third-order inversion model is a statistically significant improvement over the second-order model. Nevertheless our final model is strikingly simple in its interpretation, and seems to fit the data reasonably well. Unless we have very specific questions in mind concerning the third-order effects, the simpler model based on undirected pairwise potential seems adequate for most purposes.

11.5 References

[1] Babington Smith, B. Discussion of Professor Ross's paper. *J. Roy. Statist. Soc.* B **12**:53-56, 1950.

[2] Bradley, R.A. and Terry, M.A. Rank analysis of incomplete block designs I. *Biometrika* **39**:324-45, 1952.

[3] Critchlow, D.E. *Metric Methods for Analyzing Partially Ranked Data* Lecture Notes in Statistics, **34**, Springer-Verlag, N.Y. 1985.

[4] Critchlow, D.E., Fligner, M.A. and Verducci, J.S. Probability models on rankings. Technical Report No. 406, Dept of Statistics, Ohio State University. 1989.

[5] Diaconis, P. *Group Representations in Probability and Statistics.* I.M.S. Monograph Series **11**: Hayward, CA. 1988.

[6] Diaconis, P. A generalization of spectral analysis with application to ranked data. *Annals of Statistics* **17**, 949-79, 1989.

[7] Feller, W. *An Introduction to Probability Theory and its Applications.* New York: Wiley. 1968.

[8] Fienberg, S.E. *The Analysis of Cross-classified Categorical Data.* Cambridge, MIT Press. 1977.

[9] McCullagh, P. and Nelder, J.A. *Generalized Linear Models.* London: Chapman and Hall. 1989.

[10] Mallows, C.L. Non-null ranking models. I. *Biometrika* **44**:114-130, 1957.

[11] Nelder, J.A. A reformulation of linear models (with Discussion). *J. Roy. Statist. Soc.* A, **140**:48-77, 1977.

[12] Wilkinson and Rogers, C.E. Symbolic description of factorial models for analysis of variance. *Appl. Statist.* **22**: 392-399, 1973.

12

Aggregation Theorems and the Combination of Probabilistic Rank Orders

In D.E. Critchlow, M.A. Fligner, and J.S. Verducci (Eds.).
Probability Models and Data Analysis for Ranking Data.
Lecture Notes in Statistics. New York: Springer-Verlag, (In Press)

A. A. J. Marley[1]

ABSTRACT There are many situations where we wish to combine multiple rank orders or other preference information on a fixed set of options to obtain a combined rank order. Two of the most common applications are determining a social rank order on a set of options from a set of individual rank orders on those options, and predicting (or prescribing) an individual's overall rank order on a set of options from the rank orders on a set of component dimensions of the options. In this paper, I develop solutions to this class of problems when the rank orders can occur probabilistically. I develop aggregation theorems that are motivated by recent theoretical work on the combination of expert opinions and I discuss various models that have the property that the representations are 'of the same form' for both the component and overall rank order probabilities. I also briefly discuss difficulties in actually using such probabilistic ranking models in the social choice situation.

12.1 Introduction

According to Critchlow (1980, p. 97), at an early symposium on ranking methods Sir Maurice Kendall stated that a major outstanding problem was to construct suitable *non-null (probabilistic) ranking models* (Kendall, 1950). By this term, Kendall meant models for the set of $m!$ rank orders on a set of m items where the $m!$ rank orders may occur with unequal probabilities - as opposed to the *uniform model* where all $m!$ rank orders are equally likely to occur. From our current vantage point, it is hard to conceive of the time when this was the pressing research issue in ranking theory; in

[1] Department of Psychology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec, Canada H3A 1B1

contrast, we now have such a diversity of ranking models as to warrant an excellent integrative summary of them being published (Critchlow, Fligner, and Verducci, 1991). The major such models, as described by Critchlow et al., are those based on order statistics, paired comparisons, distances between permutations, and stagewise decompositions of the ranking process. In most developments and applications of these models, it is either implicitly assumed that there is a *single* probability density (or distribution) from some parametric family underlying the sampled rank orders, or else that a single such density (or distribution) of rank orders is to be constructed from the data - which might be, say, rank orders or paired comparisons.

For instance, if the data consist of a sample of rank orders on m items, we can ask when is the sample sum of ranks for each of the m items a sufficient statistic for the distribution generating the data (Martin-Löf, 1973); Buhlmann and Huber (1963) and Huber (1963) (note the early date) solved the parallel problem when the rank orders are constructed from paired comparisons data. However, there are numerous cases where we cannot easily defend the assumption that a *single* common probabilistic ranking distribution from some parametric family generates the data. For instance, suppose that m items are ranked on n criteria, or dimensions, by some individual, and we wish to use these rank order data to construct some overall rank order. If we assume that the rankings are probabilistic on each dimension, then there is *a priori* no reason to assume that the *same* distribution underlies the rank orders on each dimension. It may be more reasonable to assume that the distributions on each dimension belong to some common parametric family, but even this may be too strong an assumption in some situations. Thus, we wish to *aggregate* the data (set of rank orders) and/or the theoretical distributions in some 'sensible' way. The major aims of this paper are to motivate plausible aggregation rules, and also to consider when (in some appropriate sense) the family of distributions is 'closed' under such aggregation; we can then think of the rank order data as coming from a single probabilistic ranking distribution (from some parametric family), with the component distributions belonging to the same family.

After developing the main theoretical results, I briefly discuss various other practical and theoretical issues. For instance, in the above example, I assumed that the data on each dimension are probabilistic, yet frequently we will only have one sample rank order on each dimension - is this sufficient for estimation purposes? (Of course, we might sometimes have multiple samples on a given dimension). Also, the aggregation rules implied above were at the level of the distributions of rank orders, but there are other possible levels for aggregation. For instance, *order statistics models* (Critchlow et al., 1991) assume that these rank order probabilities are determined by the values of a random vector over the set of m items - i.e. associated with each item is some random variable (which may depend on the other random variables), a sample is taken of each random variable, and the items are ranked on a given occasion in the order of the magnitudes of the associated random variables. Now consider the multicomponent case - i.e. each probabilistic ranking distribution on each dimension of a set of n-dimensional items satisfies a (usually distinct) order statistics model. It might be plausible to require that the aggregate rank order probabilities also satisfy an order statistics model, with this latter model functionally determined by the component models. This is a different aggregation model than that suggested in the previous paragraph for aggregating rank order *distributions*, and an interesting theoretical question is what is the relation between these two classes of aggregation models - Alsina (1989) gives partial results on this question.

The above ideas were motivated in terms of combining rank orders defined on the components of multidimensional items. A second important interpretation of the ideas involves *probabilistic social choice.* There are numerous excellent papers on this topic (Intriligator, 1973; Fishburn, 1975, 1984; Fishburn and Gehrlein, 1977; Barbera and Sonnenschein, 1978; Clark, 1992; Gibbard, 1977; Pattanaik and Peleg, 1986; Fishburn, 1990, summarizes various of the earlier papers). Somewhat surprisingly, none of the authors except Clark discusses aggregating probabilistic rank orders - rather, they assume that some more basic data, such as from probabilistic binary choices, have to be combined to give a consensus ranking. I rectify this omission by applying the above ideas and results on multiple probabilistic rank orders to social choice, plus I discuss probabilistic versions of *approval voting* (Brams and Fishburn, 1978). Unfortunately, I can show that voting procedures such as those based on sums of ranks, number of approval votes, etc., are 'optimal' (in a sense to be defined) only when the ranking or choice probability distributions for each voter are the *same* - not only from the same parametric family, but with identical parameters within that family. This limits the applicability of these 'classical' voting procedures in the case of probabilistic social choice, but the result is similar to the known limitations of such aggregation procedures in the nonprobabilistic social choice literature. (e.g. Sen, 1986).

12.2 Notation and Basic Aggregation Theorems

We have a finite m element set X of options, $m \geq 2$, and a finite n element set of dimensions (voters). $R(X)$ denotes the set of rank orders of X (no ties allowed). For $\rho \in R(X), \rho_g$ is the element of X that is ranked in the g-th position of ρ, and $\rho(x)$ denotes the rank order position of $x \in X$ under the ranking ρ. $\rho = \rho_1 \ldots \rho_m, \sigma = \sigma_1 \ldots \sigma_m$, and $\tau = \tau_1 \ldots \tau_m$, are arbitrary elements of $R(X)$. For each i, there is a probability distribution P_i over the $m!$ rank orders of X, and there exists (or is to be developed) an overall probability distribution P over the rank orders of X. I call such a set of rank order probabilities on a set X, denoted (X, P), a *structure of ranking* *probabilities*; for simplicity, I do not refer to n or $P_i, i = 1, \ldots, n$, in the notation. I write $(X, P), (X, Q)$, etc., for such structures, and the *class of ranking probabilities* on X is the family of probability distributions obtained as P and $P_i, i = 1, \ldots, n$, vary. For a set X and $\rho \in R(X)$, I use the notation $P(\rho : X)$ and $P_i(\rho : X), i = 1, \ldots, n$, to denote the relevant probabilities of the rank order ρ occuring. A more accurate notation, and one more in line with that normally used in choice theories, would be $P(\rho : R(X))$ and $P_i(\rho : R(X)), i = 1, \ldots, n$ - however, I do not believe any confusion will arise from using the briefer notation. Also, I tend to talk interchangeably of ranking probability *densitities and distributions*; since all the densities considered in this paper are discrete, this causes no difficulties; however, it is worth noting that most of the aggregation conditions that I present are at the level of the densities. Finally, although most of the results in the paper are stated in terms of a fixed finite set X they can also be applied to arbitrary subsets X of some fixed finite set T, in which case we can discuss whether the rank order distributions are 'consistent' across the subsets X of T; I present one such result (Theorem 2) regarding *order statistics models* (Critchlow et al., 1991) or *random utility models* (Luce and Suppes, 1965) as they are known in psychology.

My general goal is to *motivate aggregation* rules for combining the component rank order distributions $P_i, i = 1, \ldots, n$, to yield the overall rank order distribution P. My assumptions and techniques closely follow those on aggregating expert opinions (reviewed by Genest and Zidek, 1986), and on aggregation in stochastic choice models (Marley, 1991a). Since the majority of the proofs in this paper exactly parallel Marley's (1991a), I do not include them, but simply indicate how to translate the notation of that paper to the notation of the present paper. The end result will be two combination rules, one based on arithmetic means, the other on geometric means. Also, I normally assume at least three elements, i.e. $|X| \geq 3$, since the case of two elements sets, i.e. $|X| = 2$, is of little interest for ranking models, and also has more solutions than arithmetic and geometric means (Marley, 1991a).

I now present and motivate a plausible set of restrictions on the structures of ranking probabilities (X, P) that leads to P being an arithmetic mean of the $P_i, i = 1, \ldots, n$, then I present an alternate set of plausible assumptions that leads to P being a weighted geometric mean of the $P_i, i = 1, \ldots, n$. For the present, X is a fixed finite set.

Assumption M1. (simple marginalization property). There exists a function F_X such that for all structures of ranking probabilities (X, P) and for each $\rho \in R(X)$,

$$P(\rho : X) = F_X[\rho, P_1(\rho : X), \ldots, P_n(\rho : X)].$$

I write $F_X[\rho, \ldots]$ rather than, say $F_X, \rho[\ldots]$ or $F[X, \rho, \ldots]$ to keep the notation similar to that used in the literature on the aggregation of expert opinion (e.g. Genest, 1984) where the function would depend on ρ but its dependence on X would not be explicitly stated. Note that for a given structure (X, P) and $\rho \in R(X)$ Assumption M1 is trivially satisfied by defining $F_X(\rho, \alpha_1, \ldots, \alpha_n) = P(\rho : X)$ for all $\alpha_i \in [0,1], i = 1, \ldots, n$. However, the representation has to hold for *all* structures of ranking probabilities (X, Q) under consideration, not just a particular one (X, P).

I call Assumption M1 simple marginalization because it is a special case of the marginalization property considered in the context of opinion aggregation (McConway, 1981 and Genest, 1984) - see Marley (1991a) for discussion of that property. Clearly, aggregation by simple marginalization is similar to the way univariate or marginal distributions are aggregated to form multivariate distributions; however, it must be emphasized that here the P_i are not necessarily the marginals of P.

The main result of this section is that *provided* $|X| > 2$, the simple marginalization property, plus some regularity and existence conditions (below), implies that F_X is essentially an arithmetic mean with weights that can depend on X. However, when $|X| = 2$, the class of solutions is much larger – see Marley (1991a) for discussion of such solutions.

Assumption M2. For any n-dimensional real vectors $(r_1, \ldots, r_n)$, $(s_1, \ldots, s_n)$ with $r_i, s_i, r_i + s_i \in [0,1], i = 1, \ldots, n$, it is possible to select a structure of ranking probabilities (X, Q) and $\rho, \sigma, \tau \in R(X)$ such that for $i = 1, \ldots, n$,

$$Q_i(\rho : X) = r_i, Q_i(\sigma : X) = s_i, Q_i(\tau : X) = 1 - (r_i + s_i).$$

Note that this condition requires $|X| \geq 3$. It is a technical assumption that allows the application of functional equation results to our problem. This assumption is somewhat 'strange' in that we are dealing with a finite set X, yet we want the total set of available probabilities to be quite dense. Falmagne (1981) discusses the use of conditions similar to Assumption M2 in other situations, and presents weaker versions of those conditions; Aczel and Dhombres (1989, Chapter 6) discuss the related general problem of *conditional* functional equations.

Condition M2 is *not* satisfied by the usual ranking version of Luce's choice model (Luce and Suppes, 1965, and later in this paper). For a three element set $X = \{x, y, z\}$ this ranking model assumes that there are ratio scales $v_i, i = 1, \ldots, n$ such that for each $\rho \in R(X)$,

$$Q_i(\rho : X) = v_i \frac{v_i(\rho_1)}{\sum_{j=1}^{3} v_i(\rho_j)} \cdot \frac{v_i(\rho_2)}{\sum_{k=2}^{3} v_i(\rho_k)}$$

Now consider the case of Assumption M2 on such a three-element set X where none of the $r_i, s_i, r_i + s_i, i = 1, \ldots, n$, equals 0 or 1, i.e. we are required to select $\rho, \sigma, \tau \in R(X)$ and a structure of ranking probabilities (R, Q) such that none of $Q_i(\rho : X), Q_i(\sigma : X), Q_i(\tau : X), i = 1, \ldots, n$ equals 0 or 1, yet

$$Q_i(\rho : X) + Q_i(\sigma : X) + Q_i(\tau : X) = 1,$$

i.e. ρ, σ, τ are the *only* rank orders with nonzero probability. However, it is easy to check that on a three-element set, if the above model holds and three distinct rank orders $\rho, \sigma, \tau \in R(X)$ have nonzero probability, then *all* the scale values are nonzero, and therefore *all* the rank orders in $R(X)$ have nonzero probability, which contradicts the above equation that is required by Assumption M2. It is probably possible to develop weaker versions of Assumption M2 (and stronger versions of Assumption M1) that apply to the ranking version of the choice model; I do not do so here, however, since as we will see later Assumption M1 is not a 'natural' aggregation property for this model.

Assumption M3. (dominance principle). For structures of ranking probabilities $(X, P), (X, Q)$ and for each $\rho \in R(X)$, if $P_i(\rho : X) \leq Q_i(\rho : X)$ for $i = 1, \ldots, n$, then $P(\rho : X) \leq Q(\rho : X)$.

A similar condition has been presented in the combination of expert opinion context by Aczel, Ng, and Wagner (1984), discussed by Genest (1984), and used by Marley (1991a) in aggregating choice probabilities. The condition clearly implies Assumption M1 with F_X satisfying a form of monotonicity.

Assumption M4. (zero preservation property). For a structure of ranking probabilities (X, P), and for each $\rho \in R(X)$, if $P_i(\rho : X) = 0$ for all $i = 1, \ldots, n$, then $P(\rho : X) = 0$.

A similar condition has been presented in the combination of expert opinion context by McConway (1981), generalized by Schmidt (1984), discussed by Genest (1984), and used by Marley (1991a) in aggregating choice probabilities. The interpretation of the condition is again clear.

THEOREM 1. If a class of ranking probabilities on a finite set X satisfies Assumptions M1 and M2, then provided $|X| \geq 3$ there exist weights $w_X(i), i = 1, \ldots, n, w_X(i) \in [-1, 1]$, and a probability measure $\psi(. : X)$ on $R(X)$ such that $\sum_{j=1}^{n} w_X(j) \leq 1$ and for each structure of ranking probabilities (X, P) and $\rho \in R(X)$,

$$P(\rho : X) = \sum_{i=1}^{n} w_X(i) P_i(\rho : X) + \left[1 - \sum_{i=1}^{n} w_X(i)\right] \psi(\rho : X).$$

If Assumptions M2 and M3 hold, then the above representations has $w_X(i) \in [0, 1]$, and if Assumptions M2, M3, and M4 hold, then $w_X(i) \in [0, 1]$ and $\sum_{i=1}^{n} w_X(i) = 1$.

PROOF. The statement of Marley's (1991a) Theorem 1 exactly parallels that of the present Theorem *except* that where we have a structure (X, P) of ranking probabilities, he has a structure (X, P) of choice probabilities - i.e. for each $x \in X$ there is a set of choice probabilities $P_i(x : X), i = 1, \ldots, n$, and $P(x : X)$, where these are the probabilities of choosing the option x from the set X. However, no properties of x, X are used in the proof of Marley's (1991a) Theorem 1 beyond those stated in his Assumptions M1-M4, which parallel the corresponding conditions just given. Thus, if in Marley's (1991a) Theorem 1 we *reinterpret* X as a set of rank orders, i.e. $R(X)$ above, and *reinterpret* $x \in X$ as an element of the set of rank orders, i.e. $\rho \in R(X)$, then we immediately obtain the present result. □

For convenience, I refer to a class of ranking probabilities with the representation of Theorem 1 as satisfying *arithmetic mean combination (of ranking probabilities)*. As in Marley (1991a), it is unclear how to interpret negative weights in this context, although examples with such weights can be constructed that satisfy the probability constraints - see Genest (1984). Thus it is reasonable to add Assumptions M3 and/or M4. The theorem might appear to have no content, in that for a particular structure of ranking probabilities (X, P) one can always set $w_X(i) = 0, i = 1, \ldots, n$ and $\psi(\rho : X) = P(\rho : X)$ for each $\rho \in R(X)$. However, remember that the representation is to hold for *all* structures of ranking probabilities (X, P)satisfying the specified conditions; in particular, the special solution just given says that the overall ranking probabilities have no dependence on the component unidimensional ranking probabilities - this is, of course, a possible solution but not one of much interest for the relevant empirical domains. Also note that the theorem gives no constraint on the form of the (nonnegative) weights $w_X(i)$ and the probability distribution $\psi(. : X)$, only that they exist. Genest and McConway (1990) discuss various interpretations of weights in such *linear opinion pools*, and methods of estimating them with relevant asymptotic properties. In the present context, the obvious interpretation is that $w_X(i)$ is the probability of the person 'attending' to dimension i, and basing the choice on that dimension alone, and that $1 - \sum_{i=1}^{n} w_X(i)$ is the probability of choosing 'randomly' according to the distribution $\psi(. : X)$. It is important to note that, since the weights depend on the context X, such a strategy can lead to quite complex patterns of choices which will not superficially appear to be determined (at any given choice opportunity) by a single dimension.

The above aggregation result was for a *fixed* finite set X; I now discuss the case where the rank orders are defined on *all* the subsets (with at least two elements) of some finite master set T. First, I need to introduce an additional definition.

DEFINITION 1. A *ranking random utility model* is a set of ranking probabilities defined on all the subsets (with at least two elements) of a finite set T for which there is a random vector U on T such that for each $X \subseteq T$ with $|X| \geq 2$, and for each $\rho = \rho_1 \ldots \rho_m \in R(X)$,

$$P(\rho : X) = \Pr[U(\rho_1) > \ldots > U(\rho_m)].$$

If the random vector U consists of components that are independent random variables, then the model is an *independent* random utility model.

Random utility model is the term used in psychology (e.g. Luce and Suppes, 1965); an alternate term is *order statistics model* (e.g. Critchlow et al, 1991). The latter term tends to be used for the case where the random variable representation is only known to hold for a *particular set* X; since *any* rank order distribution on a fixed set X can be given a random variable representation (see the representation for the case $X = T$ at the end of Section 12.4), this case is usually studied in the context of additional assumptions such as that the components of the random vector are independent and/or they have specified distributions such as the normal.

The following result follows easily using the techniques of Theorem 49 of Luce and Suppes (1965).

LEMMA 1. A set of ranking probabilities defined for all subsets (with at least two elements) of a finite set T satisfies a ranking random utility model if and only if there exists a probability distribution P over the set of rank orders of T such that for $X \subseteq T$ with $|X| \geq 2$, and each $\rho \in R(X)$,

$$P(\rho : X) = \sum_{\sigma \in R(\rho, X)} P(\sigma : T)$$

where $R(\rho, X)$ is the set of rank orders on T that agree with the rank order ρ on X.

In order to prove the next theorem, I need to strengthen Assumption M2 which I do by adding a further assumption.

Assumption M5. For any n-dimensional real vector $(r_1, \ldots, r_n)$ with $r_i \in [0, 1]$ and any $\rho \in R(X)$ with $|X| \geq 2$ it is possible to select a structure of ranking probabilities (X, P) such that $P_i(\rho : X) = r_i, i = 1, \ldots, n$.

This is, in fact, a slightly stronger condition than is needed for the proof. Note that it would form part of Assumption M2 if in that condition the rank orders $\rho, \sigma, \tau \in R(X)$ were fixed *before* the structure (X, P) were selected.

In the statement of the next results, *component* ranking probabilities refers to a set $P_i, i = 1, \ldots, n$, and *overall* ranking probabilities refers to P.

THEOREM 2. Assume that each of a set of component and overall ranking probabilities defined for all subsets (with at least two elements) of a finite set T satisfy (usually distinct) ranking random utility models and that Assumption M5 holds for all $X \subseteq T$ with $|X| \geq 2$. Then under the conditions of Theorem 1, its results hold with the additional constraints that for each $X \subseteq T$ with $|X| \geq 3$, and $\rho \in R(X)$, the weights $w_X(i)$ are independent of X and $\psi(\rho : X) = \sum_{\sigma \in R(\rho, X)} \psi(\sigma : T)$.

PROOF. In the Appendix.

When the weights are nonnegative, this result can be interpreted as giving a *probabilistic social choice rule* (Pattanaik and Peleg, 1986). In this interpretation, $P_i, i = 1, \ldots, n$, are the distributions of rank orders for voter $i, i = 1, \ldots, n$, and P is the distribution of rank orders for the society. For a set $X \subseteq T$ with $|X| \geq 3$, the social distribution is determined by that of voter i with probability $w(i)$ and is *imposed* according to the distribution $\psi(. : X)$ with probability $1 - \sum_{j=1}^{n} w(j)$. The special cases where $\sum_{j=1}^{n} w(j) = 1$ are called *random dictactorship* rules. Such rules have been axiomatized and discussed by numerous authors in the case of *deterministic* component rank orders (e.g. Barbera and Sonnenschein, 1978; Clark, 1992; Gibbard, 1977; Pattanaik and Peleg, 1986); in this deterministic case, such random dictatorship rules are the only probabilistic aggregation rules that are neither imposed nor open to strategic manipulation (e.g. Gibbard, 1977). To my knowledge, I am the first author to extend such aggregation rules to the case where the component rank orders can occur probabilistically. Note, however, that if we only have a *single* rank order from each voter, we can use these rules without having to decide whether the voter gives *deterministic* or *probabilistic rank* order data. In particular, for each voter $i, i = 1, \dots, n$, let

$$S_i(\sigma : X) = \begin{array}{ll} 1 & \text{voter } i \text{ produces sample rank order } \sigma \in R(X) \\ & \text{if} \\ 0 & \text{voter } i \text{ does not produce sample rank order } \sigma \in R(X) \end{array}$$

Then we can let the overall rank order probability be given by: for each $\rho \in R(X)$,

$$P(\rho : X) = \sum_{i=1}^{n} w(i) S_i(\rho : X),$$

i.e. we have a random dictatorship rule, and we do not need to know whether repeated samples would give the same component rank order for each voter. Of course, if the voters are 'probabilistic', then a larger sample is desirable, but it is not necessary. This is in stark contrast with *geometric mean aggregation* (discussed next) which does not usually 'work' well either in the deterministic or single sample cases.

The aggregation rules of Theorem 2 satisfy a form of the *independence of irrelevant alternatives condition*: given a feasible set $X \subseteq T$ with $|X| \geq 3$, if the individual (probabilistic) rank orders over X remain the same, then the lottery on the basis of which society makes a choice from X remains the same even though the individual (probabilistic) rank orders over T may have changed otherwise. Note that this independence condition follows from the combination of the ranking random utility model (as represented in Lemma 1) and Assumption M1, neither of which alone is equivalent to the independence condition.

As discussed later, there does not seem to be a natural family (parametric or otherwise) of probabilistic ranking distributions that is 'closed' under such arithmetic combinations. I now turn to an alternate set of assumptions that lead to a weighted geometric mean combination rule under which several standard ranking models are suitably 'closed'.

As above, I present various assumptions, motivating them as I proceed. Again, for the time being, X is a fixed finite set and to avoid excessive technical detail I restrict attention to *nonzero* structures of ranking probabilities, i.e. structures (X, P) such that for each $\rho \in R(X), P(\rho : X) \neq 0$ and $P_i(\rho : X) \neq 0, i = 1, \ldots, n$. For a structure of nonzero ranking probabilities (X, P) and for each $\rho, \sigma, \in R(X)$, let

$$L_X^P(\rho, \sigma) = \frac{P(\rho : X)}{P(\sigma : X)},$$

i.e. $L_X^P(\rho, \sigma)$ is the likelihood or odds ratio of ρ versus occuring according to the measure P. Similar notation is used for the corresponding component odds ratios.

Assumption L1. (likelihood independence property). There exists a function F_X such that all nonzero structures of ranking probabilities (X, P) and for each $\rho, \sigma \in, R(X)$,

$$L_X^P(\rho, \sigma) = F_X\left[L_X^{P_1}(\rho, \sigma), \ldots, L_X^{P_n}(\rho, \sigma)\right].$$

Clearly, this condition states that it does not matter (in calculating likelihood ratios) whether one first calculates them on the individual dimensions, and then combines these ratios over dimensions, or simply calculates likelihood ratios on the multidimensional set.

There are various, roughly equivalent, forms of this assumption, some of which have been used in the aggregation literature. For instance, with $\rho \in \Re \subseteq R(X)$, i.e. $\Re$ is a subset of $R(X)$, let,

$$P(\rho : X|\Re) = \frac{P(\rho : X)}{P(\Re : X)} = \frac{P(\rho : X)}{\sum_{\sigma \in \Re} P(\sigma : X)},$$

i.e. $P(\rho : X|\Re)$ is the conditional probability of ρ occuring given that some element of $\Re$ occurs (with a similar notation for each dimension). Now assume that for $\rho \in \Re \subseteq R(X)$,

$$P(\rho : X|\Re) = G_X[P_1(\rho : X|\Re), \ldots, P_n(\rho : X : \Re)]$$

for some function G_X . This is similar to the *external Bayesian condition* (Genest and Zidek, 1986), and can be used to obtain results similar to Theorem 3 below. However, Assumption L1 leads to that result in a more direct manner.

I now show that when $|X| \geq 3$, the only solutions satisfying Assumption L1 (plus the following two existence and monotonicity conditions) are weighted geometric means.

Assumption L2. For any n-dimensional positive real vectors $(r_1, \ldots, r_n)$, $(s_1, \ldots, s_n)$, and for any X with $|X| \geq 3$, it is possible to select a structure of nonzero ranking probabilities (X, P) and $\rho, \sigma, \tau \in R(X)$, such that for $i = 1, \ldots, n$,

$$L_X^{P_i}(\rho, \sigma) = r_i, L_X^{P_i}(\sigma, \tau) = s_i.$$

Note that this condition requires $|X| \geq 3$. Similar general comments can be made about this condition as were made previously about Assumption M2. However, in contrast to Assumption M2, this condition can be reasonably assumed to hold for the usual ranking version of Luce's choice model. When $|X| = 3$, direct evaluation leads to the desired solution; the case $|X| > 3$ can essentially be reduced to the three-element case by selecting $Y \subset X$ with $|Y| = 3$, and constructing $\rho, \sigma, \tau \in R(X)$ with $\rho = \alpha\eta, \sigma = \beta\eta, \tau = \gamma\eta$, where $\alpha, \beta, \gamma \in R(Y)$ and $\eta \in R(X - Y)$.

Assumption L3. (dominance principle). For structures of nonzero ranking probabilities $(X, P), (X, Q)$, and $\rho, \sigma \in R(X)$,

$$\text{if } L_X^{P_i}(\rho, \sigma) \leq L_X^{Q_i}(\rho, \sigma) \text{ for all } i = 1, \ldots, n,$$

$$\text{then } L_X^{P}(\rho, \sigma) \leq L_X^{Q}(\rho, \sigma).$$

This is again a monotonicity condition for F_X in Assumption L1. It does not imply Assumption L1, although it does imply a condition that can be written as

$$L_X^P(\rho, \sigma) = F_X[\rho, \sigma L_X^{P_1}(\rho, \sigma), \ldots, L_X^{P_n}(\rho, \sigma)],$$

i.e. an explicit dependence on the elements $\rho, \sigma \in R(X)$ is included that may vary as ρ, σ are varied.

THEOREM 3. If a class of nonzero ranking probabilities on a finite set X with $|X| \geq 3$ satisfies Assumptions L1 - L3, then there exist nonnegative constants $w_X(i), i = 1, \ldots, n$, such that for each structure of nonzero ranking probabilities (X, P) and for each $\rho \in R(X)$,

$$P(\rho : X) = \frac{\prod_{i=1}^{n} P_i(\rho : X)^{w_X(i)}}{\sum_{\sigma \in R(X)} \prod_{i=1}^{n} P_i(\sigma : X)^{w_X(i)}}.$$

PROOF. As with Theorem 1, Marley's (1991a) Theorem 2 uses Assumptions similar to L1-L3 to prove a representation on a structure of choice probabilities (X, P). If we reinterpret X as a set of rank orders $R(X)$ and reinterpret $x \in X$ as a rank order $\rho \in R(X)$, then Marley's (1991a) Theorem 2 gives the desired result. □

For convenience, I refer to a class of ranking probabilities with the representation of Theorem 3 as satisfying *weighted geometric mean combination (of ranking probabilities)*. The next section discusses this class of representations, but here I briefly mention a process interpretation of this structure when the model is applied to the ranking of multidimensional options by an individual with $w_X(i) = 1, i = 1, \ldots, n$. In this case, the overall rank order ρ is selected provided ρ is selected simultaneously on all of the dimensions i; otherwise, rank orders are resampled on the dimensions until such an event occurs.

Note that if the component rank orders occur *deterministically* - i.e. for each $i, i = 1, \ldots, n$, there is a unique $\rho^i \in R(X)$ with $P_i(\rho^i : X) = 1$ - then the assumptions of Theorem 3 are not met (we do not have nonzero ranking probabilities). Also, in such cases, the overall rank order probability given by the theorem is well-defined (has nonzero denominator) only if $\rho^i = \rho^j$ for all $i, j \in \{1, \ldots, n\}$, i.e. only if there is *unanimity* among the dimensions (voters) concerning which is the 'best' rank order; put another way, each dimension (voter) can *veto* undesirable rank orders. Similarily, in the case of *probabilistic* rank orders, if we have a single sample rank order for each dimension (voter) with no relations assumed between the rank order distributions across dimensions (voters), then the above arguments from the deterministic case apply. On the other hand, if the component rank order distributions are related in some way - e.g. they are identical - then we can obtain 'sensible' results for the combined rank order probabilities in this single sample version of weighted geometric mean aggregation; such cases are presented in the next section.

Although ranking is the major focus of this paper, the discussion of *L(uce)-decomposability* (the following section and Critchlow et al., 1991) requires the introduction of choice probabilities $P(x : Y), x \in Y \subseteq X$ (respectively, $P_i(x : Y), i = 1, \ldots, n$), where $P(x : Y)$ (respectively, $P_i(x : Y)$) can be interpreted as the probability that the element x is selected as the 'best' overall (respectively, on dimension i) with respect to the elements in the set X. Then *weighted geometric mean combination (of choice probabilities)* (Marley, 1991a, Theorem 2) means that there are nonnegative constants $w_Y(i), i = 1, \ldots, n$, such that for every $x \in Y \subseteq X$,

$$P(x : Y) = \frac{\prod_{i=1}^{n} P_i(x : Y)^{w_Y(i)}}{\sum_{y \in Y} \prod_{i=1}^{n} P_i(y : Y)^{w_Y(i)}}.$$

12.3 Specific Multidimensional Ranking and Subset Selection Models and Their Properties

As mentioned previously, Critchlow et al (1991) describe numerous classes of ranking models and their properties. Specifically, they describe *order statistics models, ranking models induced by paired comparisons, ranking models based on distances between permutations*, and *multistage ranking models.* They also study various properties satisfied by some or all of the models, namely *label invariance, strong unimodality, complete consensus*, and *L(uce)-decomposability.* I now consider when a given model or condition is *closed* under the above weighted geometric mean model - i.e. if the given condition holds for each of the ranking structures on the dimensions, does it hold on the overall ranking structure? I will not study all of the models and the properties here, but simply give preliminary results on models satisfying strong unimodality or L-decomposability and on models induced by paired comparisons.

Strong unimodality holds if the structure of ranking probabilities (X, P) has a modal ranking π_0 such that the probability $P(\rho : X), \rho \in R(X)$, is nonincreasing as ρ moves farther away from π_0 by interchanging adjacent items in ρ on which ρ and π_0 agree in rank order (see Critchlow et al., 1991, for the precise specification). One can check that strong unimodality is closed under weighted geometric mean combination provided the modal ranking is the same on all dimensions - not a plausible condition - but it is not in general invariant otherwise.

A structure of ranking probabilities (X, P) satisfies *L(uce)-decomposability* provided there exist choice probabilities $P(x : Y), x \in Y \subseteq X$, such that for every $\rho = \rho_1 \dots \rho_m \in R(X)$,

$$P(\rho : X) = P(\rho_1 : \{\rho_1, \dots, \rho_m\})\ P(\rho_2 : \{\rho_2, \dots, \rho_m\}) \dots$$
$$P(\rho_{m-1} : \{\rho_{m-1}, \rho_m\})$$

(see Critchlow et al, 1991, for discussion and interpretation of such representations). [This is the one case in the paper where my notation $P(\rho : X)$ for the probability of the *rank order* ρ is open to misinterpretation. Note that in the above expression, the term on the left hand side is a *rank order* probability, whereas each term on the right hand side is a *choice* probability].

Bringing together the various conditions that I have introduced, we have: weighted geometric mean combination of ranking probabilities; weighted geometric mean combination of choice probabilities; and L-decomposability. The obvious next topic is which ranking and choice models satisfy some or all of these properties. This leads to a number of fascinating questions and characterizations on which I have numerous results (Marley, 1991a, 1991b). As illustration, I now present results on a particular class of models where the ranking probabilities are induced from paired comparisons. I first discuss the models from the viewpoint of possibly descriptive probabilistic models of ranking of multidimensional options, then in terms of prescriptive probabilistic models of social choice.

For distinct $x, y \in X, p(x, y)$ denotes the probability that item x is preferred to item y in a paired comparison of these two items. Assume that every distinct pair of items can be compared, and that if the results are consistent with a rank order then that rank order is selected; otherwise, the entire process is repeated until a ranking is obtained. This process yields, for $\rho = \rho_1 \dots \rho_m \in R(X)$,

$$P(\rho : X) = \frac{\prod_{1 \leq g < h \leq m} p(\rho_g, \rho_h)}{\sum_{\sigma \in R(X)} \prod_{1 \leq j < k \leq m} p(\sigma_j, \sigma_k)}.$$

For notational simplicity I write this as

$$P(\rho : X) \sim \prod_{1 \leq g < h \leq m} \rho(\rho_g, \rho_h)$$

(with similar notation for all such 'normalized' representations). Clearly, if the above representation holds for each dimension $i, i = 1, \ldots, n$, and if weighted geometric mean combination holds with weights that do not depend on X, then a representation of the same form holds for the overall ranking probabilities $P(\rho : X)$ - i.e.

$$P(\rho : X) \sim \prod_{1 \leq g < h \leq m} \rho(\rho_g, \rho_h)$$

$$\text{where} \qquad p(\rho_g, \rho_h) \sim \prod_{i=1}^{n} p_i(\rho_g, \rho_h)^{w(i)}.$$

The unidimensional version of this model is due to Mallows (1957), and satisfies L-decomposability (Marley, 1968; Critchlow et al., 1991). Since the multidimensional version has the same form, it also satisfies L-decomposability. Thus, I have presented an example of a ranking model that satisfies geometric mean combination of ranking probabilities with L-decomposability holding on each dimension and overall. However, the choice probabilities of the decomposition (explicitly stated in Marley, 1968) do not combine according to weighted geometric mean combination when $|X| \geq 3$. (I have not proved this result, which seems obviously correct). Therefore, an open problem in this area concerns whether or not there is a ranking model that satisfies L-decomposability on each dimension and overall, with both the ranking probabilities and the induced choice probabilities satisfying weighted geometric mean combination.

Continuing with the development of the model, if the binary choice probabilities have the Bradley-Terry-Luce form, i.e. there are ratio scales $v_i, i = 1, \ldots, n$, such that for $x, y \in X$,

$$p_i(x, y) = \frac{v_i(x)}{v_i(x) + v_i(y)},$$

then we obtain

$$P(\rho : X) \sim \prod_{l=1}^{m} v(\rho_l)^{m-1}$$

$$\text{where} \qquad v(\rho_l) = \prod_{i=1}^{n} v_i(\rho_l)^{w(i)}.$$

If a set of rank orders is probabilistically generated by this model with v_i and $w(i)$ independent of i for all i (or, alternatively, one only considers the model for a fixed i), then a sufficient statistic for the above ranking model is the vector whose k-th component is the sample sum of the ranks associated with the k-th item, $k = 1, \ldots, m$ (Martin-Löf, 1973). It is probably possible to prove the converse of this result, i.e. if the overall rank order probabilities are generated by the above model, then the sample sum of ranks is a sufficient statistic only if the v_i and $w(i)$ are independent of i. (See Buhlmann and Huber, 1963, and Huber, 1963, for relevant techniques, and also the discussion in the next paragraphs).

Turning to the social choice interpretation, assume v_i is independent of i and $w(i) = 1/n$ for all i, giving that one can estimate the (common) $p(x, y), x, y \in X$, from the collective choices of a group of n voters. Now suppose that each of the n voters compares each distinct $x, y \in X$, and let

$$a(x,y) = \begin{array}{ccc} k - \frac{n}{2} & & k \text{ of the } n \text{ voters select } x \text{ over } y \\ & \text{if} & \\ \frac{n}{2} - k & & n - k \text{ of the } n \text{ voters select } y \text{ over } x \end{array}$$

(i.e. the row scores are transformed so that for each distinct $x, y \in X, a(x, y) + a(y, x) = 0$). Now suppose that for each item x we calculate the score $\sum_{z \in X - \{x\}} a(x, z)$, and rank the items in descending order of these scores. We can then ask for what class of 'underlying probability structures is such a ranking by scores 'optimal' (with these terms precisely defined in Buhlmann and Huber, 1963, and Huber, 1963). It turns out that such is the case for precisely the above special case probabilistic ranking model just described. Note the unfortunate assumption that the binary choice probabilities of *all* the voters are based on the *same* scale values for this ranking by scores to be optimal (as is the case in the Martin-Löf result). This is a very restrictive condition making it difficult to vigorously defend such ranking by total scores methods.

So far I have used aggregation techniques to study the combination of probabilities for the selection of the 'best' element from some available set (Marley, 1991a, 1991b), and to study the combination of ranking probabilities on some set (this paper). I now apply similar techniques to probabilistic models of subset selection - first, as a descriptive model for choice amongst multidimensional options; second, as a probabilistic version of *approval voting* (Brams and Fishburn, 1978; Fishburn, 1978; Fishburn and Brams, 1981).

A person is presented with a set X of (multidimensional) items, from which he/she may select any subset $Y, \phi \subseteq Y \subseteq X$; in the approval voting case these would be the 'acceptable' items. Consider the following process that might be descriptive of the selection process on the i^{th} dimension (later, of the i^{th} voter). The person considers each item x in turn, and independently accepts (respectively, rejects) that item with a probability $v_i(x)$ (respectively, $1 - v_i(x)$); for simplicity in the following, I assume $v_i(x) \neq 0, 1$ for any i or x. Then with X the set of items, Y a subset of X, and $P_i(Y : X)$ the probability that the person approves of subset $Y, \phi \subseteq Y \subseteq X$, on dimension i, we have

$$P_i(Y : X) = \prod_{y \in Y} v_i(y) \prod_{z \in X-Y} (1 - v_i(z)).$$

Note that since

$$\sum_{\phi \subseteq Y \subseteq X} P_i(Y : X) = 1,$$

i.e. some subset of X must be selected by this process, we can divide the expression for $P_i(Y : X), \Phi \subseteq Y \subseteq X$, by this quantity to obtain

$$P_i(Y : X) = \frac{\prod_{y \in Y} v_i(y) \prod_{z \in X-Y} (1 - v_i(z))}{\sum_{\theta \subseteq S \subseteq X} \prod_{s \in S} v_i(s) \prod_{t \in X-S} (1 - v_i(t))},$$

which, assuming $v_i(y) \neq 0, 1$ for any $y \in X$, and letting

$$u_i(y) = \frac{v_i(y)}{1 - v_i(y)}$$

becomes

$$P_i(Y : X) = \frac{\prod_{y \in Y} u_i(y)}{\sum_{\theta \subseteq S \subseteq X} \prod_{s \in S} u_i(s)},$$

i.e. $$P_i(Y : X) \sim \prod_{y \in Y} u_i(y).$$

Now assuming that these subset selection probabilities are aggregated as in Theorem 3 with weights that do not depend on the set Y (again, the axioms leading to that theorem are easily reinterpreted for the present context), we obtain that

$$P(Y : X) \sim \prod_{i=1}^{n} P_i(Y : X)^{w(i)}$$

from which it follows by simple substitution and cancellation of common terms that

$$P(Y : X) \sim \prod_{i=1}^{n} \left(\prod_{y \in Y} u_i(y) \right)^{w(i),}$$

i.e. $$P(Y : X) \sim \prod_{y \in Y} u(y)$$

where $$u(y) = \prod_{i=1}^{n} u_i(y)^{w(i)}.$$

In particular, the component and overall subset selection probabilities are of the 'same form'. In fact, if we let

$$v(y) = \frac{u(y)}{1 + u(y)}$$

and then reexpresses the $u(y)$ in terms of the $u_i(y)$ and hence the $v_i(y)$, we obtain that

$$v(y) = \frac{\prod_{i=1}^{n} v_i(y)^{w(i)}}{\prod_{i=1}^{n} v_i(y)^{w(i)} + \prod_{i=1}^{n}(1 - v_i(y))^{w(i)}}$$

and

$$P(Y : X) \sim \prod_{y \in Y} v(y) \prod_{z \in X - Y} (1 - v(y)).$$

In particular, when $w(i) = 1$ for all i, $v(y)$ has the following interpretation: accept y if it is acceptable on all dimensions, reject y if it is unacceptable on all dimensions, otherwise resample. Then $v(y)$ (respectively, $1 - v(y)$) is the probability that y is accepted (respectively, rejected) according to the aggregate rule.

Turning to the approval voting interpretation, assume $w(i) = 1$ for all i, and let Y_i be the subset selected by voter $i, i = 1, \ldots, n$. Then the probability of the selection vector $(Y_1, \ldots, Y_n)$ across voters is given by

$$\prod_{i=1}^{n} P_i(Y_i : X),$$

which has the form

$$\eta(X) \prod_{i=1}^{n} \left(\prod_{y \in Y_i} u_i(y) \right)$$

where

$$\eta(X) = \prod_{i=1}^{n} \left(\sum_{\phi \subseteq S \subseteq X} \prod_{s \in S} u_i(s) \right)^{-1}$$

is independent of the data. Therefore the probability of the selection vector $(Y_1, \ldots, Y_n)$ is

$$\eta(X) \prod_{i=1}^{n} \prod_{y \in Y} u_i(y)^{s\,i^{(y)}}$$

where

$$s_i(y) = \begin{matrix} 1 & & \text{voter } i \text{ accepts option } y \\ & \text{if} & \\ 0 & & \text{voter } i \text{ rejects option } y. \end{matrix}$$

In particular, if $u_i(y)$ is independent of i, then the vector of vote sums $\sum_{y \in Y} s_i(y)$, the score used in approval voting, is sufficient for the distribution of the selection vector. The techniques of Buhlmann and Huber (1963) and Huber (1963) should now be used to prove (if it is so) that ranking in descending order of these scores is 'optimal' within the class of models for probabilistic subset selection if and only if the probabilistic subset model has the above form (with u_i and $w(i)$ independent of i). As with the previous social choice interpretation, this result (if true) is disappointing as it says that the usual method of social choice ranking based on approval scores is optimal only if all the voters have the same scale values for all the options.

12.4 Multidimensional Random Variable Models

The above probabilistic models of ranking, choice, and subset selection amongst multidimensional options assume that the probabilities for the overall choices can be written as an aggregate function of the corespond-ing probabilities on the component dimensions. For *random utility models* (Section 12.2), also known as order statistics models (Critchlow et al., 1991), an alternative (in general incompatible - see below) combination process is plausible. Remember, in a random utility model (generalized to multiple dimensions), for each x in the master set T and for each dimension $i, i = 1, \ldots, n$, there is a random variable $\boldsymbol{t}_i(x)$ such that for $\rho = \rho_1 \ldots \rho_m \in R(X), X \subseteq T, |X| \geq 2$,

$$P_i(\rho : X) = \Pr[\boldsymbol{t}_i(\rho_1) > \ldots \boldsymbol{t}_i(\rho_m)],$$

and also there are overall random variables $\boldsymbol{t}(x), x \in T$, such that for $\rho = \rho_1 \ldots \rho_m \in R(X), X \subseteq T, |X| \geq 2$,

$$P(\rho : X) = \Pr[\boldsymbol{t}(\rho_1) > \ldots > \boldsymbol{t}(\rho_m)].$$

With these forms in front of us, there is an obvious plausible combination model: assume that there is a function H such that for each $x \in T$,

$$\boldsymbol{t}(x) = H[\boldsymbol{t}_1(x), \ldots, \boldsymbol{t}_n(x)].$$

Clearly, H has to be constrained in such a way that $\boldsymbol{t}$ is a random variable (Alsina, 1989). An alternative approach to the aggregation at the random variable level is to consider the cumulative distributions associated with the random variables, i.e. for real t, let

$$F_{x,i}(t) = \Pr[\boldsymbol{t}_i(x) < t] \quad , i = 1, \ldots, n,$$

$$F_x(t) = \Pr[\boldsymbol{t}(x) < t],$$

and assume that there is a function f such that for real t,

$$F_x(t) = f[F_{x,1}(t), \ldots, F_{x,n}(t)],$$

where f is such that F_x is a cumulative distribution (Alsina, 1989). In fact, these two approaches are often incompatible (Alsina, 1989), and neither approach is in general compatible with the earlier approach of aggregation at the level of the ranking probabilities - see Theorem 2 and the discussion below. I do not currently have general results on the class of multidimensional order statistics models, so I briefly present one result regarding a multidimensional version of Luce's choice model, and relate it to the previous aggregation results. (Clark, 1992, applies the random variable aggregation approach to the social choice interpretation).

Consider the following independent random variables: for $x \in T, t \geq 0, i = 1, \ldots, n$,

$$\Pr(\boldsymbol{t}(x) \leq t) = \exp - \frac{v_i(x)}{t}$$

for some (ratio) scales $v_i, i = 1, \ldots, n$, on X, and let

$$\boldsymbol{t}(x) = \overset{\max}{i} \quad \boldsymbol{t}_i(x).$$

Then clearly

$$\Pr(\boldsymbol{t}(x) \leq t) = \exp - \frac{v(x)}{t}$$

where

$$v(x) = \sum_{i=1}^{n} v_i(x),$$

and for $\rho \in R(X)$ we obtain (see Robertson and Strauss, 1981)

$$P_i(\rho : X) = \Pr[\boldsymbol{t}_i(\rho_1) > \ldots > \boldsymbol{t}_i(\rho_m)]$$

$$= \prod_{h=1}^{m-1} \frac{v_i(\rho_h)}{\sum_{k=h}^{m} v_i(\rho_k)},$$

and

$$P(\rho : X) = \Pr[\boldsymbol{t}(\rho_1) > \ldots > \boldsymbol{t}(\rho_m)]$$

$$= \prod_{h=1}^{m-1} \frac{v(\rho_h)}{\sum_{k=h}^{m} v(\rho_k)}$$

with, for $x \in X$,

$$v(x) = \sum_{i=1}^{n} v_i(x).$$

There are several interesting observations to be made about these representations, which suggest the need for further work. First, the ranking probabilities (both componentwise and overall) obviously satisfy L-decomposability with the choice probabilities in the decompositions satisfing Luce's choice model, i.e. for $x \in Y \subseteq X$,

$$P_i(x : Y) = \frac{v_i(x)}{\sum_{y \in Y} \ v_i(y)}, i = 1, \ldots, n,$$

$$\text{and} \qquad P(x : Y) = \frac{v(x)}{\sum_{y \in Y} \ v(y)}.$$

$$\text{where } v(y) = \sum_{i=1}^{n} \ v_i(y).$$

In fact, the unidimensional version of the above ranking model motivated Critchlow et al's (1991) definition of L-(or Luce-) decomposability. However, neither the ranking or the choice probabilities combine according to weighted geometric mean combination - in fact, Marley (1991a,1991b) shows that such component and overall choice probabilities satisfying Luce's choice model (with ratio scales $v_i, i = 1, \ldots, n$, and v) satisfy geometric mean combination provided $v(x) = \prod_{i=1}^{n} v_i(x)$; he was unable to construct an (aggregate) order statistics model with this latter form (on the choice probabilities).

Continuing, the earlier Theorem 2 shows that under its aggregation condition and its reasonable 'technical' conditions, *no* aggregate random utility model satisfies weighted geometric mean combination. That Theorem shows that under its conditions, if a multicomponent random utility model satisfies an aggregation rule at the level of the *rank order probabilities*, then the aggregation rule is an arithmetic mean. Note that there are also random variable representations of this arithmetic mean aggregation rule at the level of *(random) rank orders*: with the notation as in Theorems 1, 2 for $\rho = \rho_1 \ldots \rho_m \in R(T)$, and $i = 1, \ldots, n$, let

$$\boldsymbol{t}(\rho) = \begin{array}{ccc} \boldsymbol{t}_i(\rho) & & w(i) \\ & \text{with probability} & \\ \boldsymbol{Q}(\rho : X) & & 1 - \sum_{i=1}^{n} \ w(i) \end{array}$$

where $\boldsymbol{Q} = (\boldsymbol{Q}(1), \ldots, \boldsymbol{Q}(m))$ (constructed below) is a random rank order such that for each $\rho \in R(T), P(\boldsymbol{Q}(\rho_1) < \ldots < \boldsymbol{Q}(\rho_m)) = \psi(\rho : T)$. Note that I have written the random utility model in the reverse of the usual order to take advantage of the definition of ρ_i as the object with the i-th rank; and we can then select $\boldsymbol{Q}(x)$ as the (random) rank of option x, i.e. $\boldsymbol{Q}(\rho_i)$ will equal i. So now with T the ordered set $(1, \ldots, m)$, and $\boldsymbol{\mathcal{T}}$ the set of permutations of T, for a permutation π in $\mathcal{T}$ let $\pi(i), i = 1, \ldots, m$, be the rank of option i, and let $\pi^{-1}(j), j = 1, \ldots, m$, be the item that is assigned rank j. Now define the random vector $\boldsymbol{Q} = (\boldsymbol{Q}(1), \ldots, \boldsymbol{Q}(m))$ (where $\boldsymbol{Q}(i)$ is the random rank of option i) by $\Pr(\boldsymbol{Q} = \pi) = \psi(\pi^{-1}(1) \ldots \pi^{-1}(m) : T)$; this gives the desired form of the random vector $\boldsymbol{Q}$.

Note that the above represents the overall random utility model as a *mixture* of the component random utility models with the mixture being at the level of the rank orders. Thus, this representation is not immeditely of the form discussed above where the aggregation was at the level of the random variables associated with individual elements of T.

12.5 Conclusion

I have used assumptions and techniques developed in the study of aggregation of expert opinions to motivate probabilistic models for choice, ranking and subset selection on multidimensional options. There remain a large number of open questions regarding relations between various models and properties for such probabilistic choice, ranking, and subset selection data; for instance, we do not currently know whether there is a probabilistic ranking model that satisfies L-decomposability on each component and overall, with both the ranking probabilities and the induced choice probabilities satisfying geometric mean combination. Reinterpretations of the presented framework give parallel probabilistic models for the social choice problem. In particular, I have shown that arithmetic mean combination of ranking probabilities can be applied equally well to the case of deterministic component rank orders and to the case of a single sample from each of a set of nondeterministic (probabilistic) rank order distributions. However, no standard probabilistic ranking (or choice) model is 'closed' under arithmetic mean combination, in contrast to numerous such 'closed' models in the case of weighted geometric mean combination. Unfortunately, weighted geometric mean combination is generally inapplicable to the case of deterministic component rank orders, and is only applicable to the case of a single sample from each of a set of nondeterministic (probabilistic) rank order distributions when these component rank order distributions are related in some way - e.g. when they are identical. These latter results also show that for the particular probabilistic models studied, classical social choice decision rules based on sums of ranks, sums of binary superiority, etc., are only 'optimal' when each voter has the same probabilistic (choice, ranking, subset) distribution from a particular limited family of distributions - note well that it is required that all voters have the *same* distribution (parameters included), not simply distributions with possibly different parameters all from the same limited family. These results are disappointing, but perhaps not surprising given the known difficulties in the deterministic case of obtaining a satisfactory social choice, or aggregation, function (Sen, 1986).

Acknowledgements:

This research was supported in part by a grant from the Natural Science and Engineering Research Council of Canada. I thank the referees for their critical comments which lead me to attempt to clarify several points.

APPENDIX

Proof of Theorem 2. Under the assumed conditions, the results of both Theorem 1 and Lemma 1 are valid. Therefore, for $X \subseteq T$ with $|X| \geq 3$ and for arbitrary $\rho \in R(X)$,

$$\begin{aligned} P(\rho : X) &= \sum_{\sigma \in R(\rho,X)} P(\sigma : T) \\ &= \sum_{\sigma \in R(\rho,X)} \left\{ \sum_{i=1}^{n} w_T(i) P_i(\sigma : T) + \left(1 - \sum_{i=1}^{n} w_T(i)\right) \psi(\sigma : T) \right\} \\ &= \sum_{i=1}^{n} w_T(i) \sum_{\sigma \in R(\rho,X)} P_i(\sigma : T) + \left(1 - \sum_{i=1}^{n} w_T(i)\right) \\ &\quad \sum_{\sigma \in R(\rho,X)} \psi(\sigma : T) \\ &= \sum_{i=1}^{n} w_T(i)\ P_i(\rho : X) + \left(1 - \sum_{i=1}^{n} w_T(i)\right) \sum_{\sigma \in R(\rho,X)} \psi(\sigma : T). \end{aligned}$$

But since $\rho \in R(X)$, we also have

$$P(\rho : X) = \sum_{i=1}^{n} w_X(i)\ P_i(\rho : X) + \left(1 - \sum_{i=1}^{n} w_X(i)\right) \psi(\rho : X).$$

Combining these two equations gives

$$\sum_{i=1}^{n} (w_T(i) - w_X(i))\ P_i(\rho : X) + \left(1 - \sum_{i=1}^{n} w_T(i)\right)$$
$$\sum_{\sigma \in R(\rho,X)} \psi(\sigma : T) - \left(1 - \sum_{i=1}^{n} w_X(i)\right) \psi(\rho : X) = 0. \qquad (A1)$$

Now using Assumption M5 to select$P_i, i = 1, \ldots, n$, with $P_i(\rho : X) = 0$, this reduces to

$$\left(1 - \sum_{i=1}^{n} w_T(i)\right) \sum_{\sigma \in R(\rho,X)} \psi(\sigma : T) = \left(1 - \sum_{i=1}^{n} w_X(i)\right) \psi(\rho : X). \quad (A2)$$

Therefore Equation A1 reduces to

$$\sum_{i=1}^{n}(w_T(i) - w_X(i))P_i(\rho : X) = 0.$$

Now using Assumption M5 to select $P_i, i = 1, \ldots, n$, with $P_i(\rho : X) = 0, i \neq k$, and $P_k(\rho : X) > 0$, this reduces to $(w_T(k) - w_X(k))P_k(\rho : X) = 0$ with $P_k(\rho : X) > 0$ - i.e. we must have $w_T(k) = w_X(k)$. But both k and X were selected arbitrarily, therefore $w_X(i) = w_T(i)$ for $i = 1, \ldots, n, X \subseteq T$ with $|X| \geq 3$. In particular $\sum_{i=1}^{n} w_T(i) = \sum_{i=1}^{n} w_X(i)$ for each $X \subseteq T$ with $|X| \geq 3$, and so returning to Equation A2, if $\sum_{i=1}^{n} w_T(i) \neq 1$ then we obtain

$$\psi(\rho : X) = \sum_{\sigma \in R(\rho, X)} \psi(\sigma : T),$$

and if $\sum_{i=1}^{n} w_T(i) = 1$ then we can impose the previous equality as it has no effect on the representation. □

12.6 References

[1] Aczel, J., and Dhombres, J. (1989). *Functional equations in several variables.* New York: Cambridge University Press.

[2] Aczel, J., Ng, C.T., and Wagner, C. (1984). Aggregation theorems for allocation problems. *SIAM Journal Algebraic Discrete Methods* **5**, 1-8.

[3] Alsina, C. (1989). Synthesizing judgements given by probability distribution functions. Manuscript, Departament Mathematiques Universidad Politecnica de Catalunya.

[4] Barbera, S. and Sonnenschein, H. (1978). Preference aggregation with randomized social orderings. *Journal of Economic Theory*, **18**, 244-254.

[5] Brams, S.J., and Fishburn, P.C. (1978). Approval voting. *American Political Science Review*, **72**, 831-847.

[6] Buhlmann, H., and Huber, P.J. (1963). Pairwise comparison and ranking in tournaments. *The Annals of Mathematical Statistics*, **34**, 501-510.

[7] Clark, S.A. (1992). The representative agent model of probabilistic social choice. *Social Sciences*, **23** (in press).

[8] Critchlow, D.E. (1980). Metric methods for analyzing partially ranked data. Lecture Notes in Statistics, #34, D. Brilliinger, S. Fienberg, J. Gani, J. Hartigan, and K. Krickerberg (Eds.). Berlin: Springer-Verlag.

[9] Critchlow, D.E., Fligner, M.A., and Verducci, J.S. (1991). Probability models on ranking. *Journal of Mathematical Psychology*, **35**, 294-318.

[10] Falmagne, J.C. (1981). On a recurrent misuse of a classical functional equation result. *Journal of Mathematical Psychology*, **23**, 190-193.

[11] Fishburn, P.C. (1975). A probabilistic model of social choice: Comment. *The Review of Economic Studies*, **42**, 297-301.

[12] Fishburn, P.C. (1978). Axioms of approval voting: direct proof. *Journal of Economic Theory*, **19**, 180-185.

[13] Fishburn, P.C. (1984). Probabilistic social choice based on simple voting comparisons. *Review of Economic Studies*, **51**, 683-692.

[14] Fishburn, P.C. (1990). Multiperson decision making: A selective review. In Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory. J. Kacprzyk and M. Fedrizzi (Eds.). Dordrecht: Kluwer. To Appear.

[15] Fishburn, P.C. and Brams, S.J. (1981). Expected utility and approval voting. *Behavioral Science*, **26**, 136-142.

[16] Fishburn, P.C. and Gehrlein, W.V. (1977). Towards a theory of elections with probabilistic preferences. *Econometrica*, **45**, 1907-1924.

[17] Genest, C. (1984). Pooling operators with the marginalization property. *The Canadian Journal of Statistics*, **12**, 153-163.

[18] Genest, C., and McConway, K.J. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting*, **9**, 53-73.

[19] Genest, C., and Zidek, J.V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, **1**, 114-148.

[20] Gibbard, A. (1977). Manipulation of schemes that mix voting with chance. *Econometrica*, **45**, 665-681.

[21] Huber, P. J. (1963). Pairwise comparison and ranking: Optimum properties of the row sum procedure. *The Annals of Mathematical Statistics*, **34**, 511-520.

[22] Intriligator, M.D. (1973). A probabilistic model of social choice. *Review of Economic Studies*, **40**, 553-560.

[23] Kendall, M.G. (1950). Discussion on Symposium on Ranking Models. *Journal of the Royal Statistical Society, Series B*, **12**, 189.

[24] Luce, R.D. (1959). *Individual Choice Behavior.* New York: Wiley.

[25] Luce, R.D. and Suppes, P. (1965). Preference, utility, and subjective probability. In R.D. Luce, R.R. Bush, and E. Galanter (Eds.), Handbook of Mathematical Psychology, III. New York: Wiley. pp. 230-270.

[26] Mallows, C.L. (1957). Non-null ranking models. I. *Biometrika*, **44**, 114-130.

[27] Marley, A.A.J. (1968). Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology*, **5**, 311-322.

[28] Marley, A.A.J. (1991a). Aggregation theorems and multidimensional choice models. *Theory & Decision*, **30**, 245-272.

[29] Marley, A.A.J. (1991b). Context-dependent probabilistic choice models based on measures of binary advantage. *Mathematical Social Sciences*, **21**, 201-231.

[30] Martin-Löf, P. (1973). Statistika Modeller. Lecture notes in Swedish, compiled by Rolf Sundberg,

Stockholm University.

[31] McConway, K.J. (1981). Marginalization and linear opinion pools. *Journal of the American Statistical Association*, **76**, 410-414.

[32] Pattanaik, P.R. and Peleg, B(1986). Distribution of power under stochastic social choice rules. *Econometrica*, **54**, 909-921.

[33] Robertson, C.A. and Strauss, D.J. (1981). A characterization theorem for random utility variables. *Journal of Mathematical Psychology*, **23**, 184-189.

[34] Schmidt, F.F. (1984). Consensus, respect, and weighted averaging. *Synthese*, **62**, 25-46.

[35] Sen, A. (1986). Social Choice Theory. In K.J. Arrow and M.D. Intriligator (Eds.). Handbook of Mathematical Economics, Vol. III. Chapter 22, 1073-1181.

13

A Nonparametric Distance Model for Unidimensional Unfolding

Rian van Blokland-Vogelesang[1]

ABSTRACT
The unidimensional unfolding model is placed in the wider context of social choice theory, median procedures and strictly unimodal distance models for rankings. Social choice theory is used to construct a framework for the unfolding model; for example, given single-peaked preference functions for individuals, Simple Majority Rule yields the median ordering as a group consensus ordering. We generalize Coombs' and Goodman's (1954) theorems: if the data follow a strictly unimodal distance model, the median ordering is an admissible ordering of the J scale that has the highest probability. This is because the maximum likelihood and the minimum-number-of-inversions criterion yield the same ordering in a strictly unimodal distance model: the mean/modal/median ordering. We prove that the group consensus ordering is transitive and is the modal or median ordering. Also, we prove that the social preference function is unimodal on the J scale in this case.

Key words and phrases: unfolding, median ordering, group consensus ranking, Kemeny distance, unimodal distance model for rankings, maximum likelihood, minimum number of inversions.

13.1 Introduction

We start with an overview of the unidimensional unfolding model as devised by Coombs (1964). Consider an experimenter asking individuals to judge which of a number of policies is more beneficial to society; he or she might use such judgments to study the *statements* (more generally, *options*) or to study the *individuals*. If the experimenter decides to interpret the behavior observed as relations between an individual *and* an option, this will lead to different conclusions than if the observed behavior is interpreted as relations between options only. Relations between individuals and options are called *preference data*. An *unfolding technique* is an algorithm for constructing

[1] University of Leiden, FSW, Department of Datatheory, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

a psychological space from such data. If the psychological space consists of one dimension, it is called a *joint scale* or J scale, and the unfolding technique is called *unidimensional unfolding.* The dimension found can be seen as the *latent structure*, a common reference frame, in a certain field of research.

Coombs' (1950, 1954, 1964) unidimensional unfolding model was devised for the analysis of complete orderings of preference, where n individuals rank k options $O_1, O_2, \ldots, O_k$, from most to least preferable. In the unfolding model, each individual and each option is represented on a single dimension, called the J scale. The locations associated with individuals are called *ideal points*, and represent the best possible option from the individual's point of view. Admissible orderings on the unidimensional unfolding scale are *Single-Peaked Preference Functions* (SPF's). An SPF is defined as follows: if the J scale is $O_1, O_2, O_3, \ldots, O_k$, then for each triple of ordered options O_j, O_{j+1}, O_{j+2}, one of the three following relations should hold: (a) $O_j >_p O_{j+1} >_p O_{j+2}$, (b) $O_j <_p O_{j+1}$ and $O_{j+1} >_p O_{j+2}$, or (c) $O_{j+2} >_p O_{j+1} >_p O_j$, where $j = 1, 2, \ldots, k-2$, and '$>_p$' denotes 'is preferred to' . This means that in passing from one option to the next, each individual's preference function monotonically rises to a peak, and then drops off monotonically. Each individual's preference ranking of options is then given by the rank order of the distances of option points from the ideal point, with nearer options being most preferred. At this point, we make a distinction between an admissible *ordering* (a J scale ordering) and an individual *ranking* (an ordering stated by an individual). Both include the options from most preferable to least preferable in that order.

Let the J scale for k options denote a k-scale, and let $A, B, C, D, \ldots$ denote the successive options on the scale. Midpoints are represented in lowercase, e.g., ab is the midpoint between options A and B. J scales are named according to the first admissible ordering on the J scale, which corresponds to the order of the options along the J scale. Possible rankings of preference correspond to segments of the J scale (see Figure 1). The Figure shows a 3-scale with options A, B, C. Between the options A and B is the midpoint ab; to the left of the midpoint ab is the segment ABC; this represents the set of ideal points of individuals with $A >_p B, A >_p C$ and $B >_p C$. (thus, nearer to A than to B or C) To the right of the midpoint ab is the segment BAC: the set of ideal points of individuals with $B >_p A, B >_p C, A >_p C$ (thus, nearer to B than to A or C), and so forth. Four out of six (=3!) rankings are *admissible orderings*, two are not represented on the J scale and, hence, are *inadmissible* orderings for this scale.

With four options, two distinct 4-scales arise, depending on the order of the midpoints ad and bc (see Figure 2). In Figure 2a their order is ad, bc; in Figure 2b this is bc, ad. So there are two distinct midpoint orders or *quantitative* 4-scales. These 4-scales differ only in the admissible orderings in the midst of the J scale ($BCDA$ and $CBAD$, respectively). The quantitative

J scale is defined as a *strict* order of options and of the midpoints between the options. If we disregard the particular order of the midpoints, the *qualitative J* scale arises. The qualitative J scale is defined as a strict order of the options only, this is denoted the J *order* in the following. Midpoints are not strictly ordered on the qualitative J order: they are *partially* ordered, and admissible orderings cannot be represented on a unidimensional continuum, they can be represented in a lattice of orderings (see Van Blokland, 1991). For $k = 3$, there is only one quantitative 3-scale, for $k = 4$ two (see Figure 2), for $k = 5$ twelve. With larger numbers of options, the number of quantitative J scales that can be derived from one qualitative J order, increases very quickly (for $k = 9$, this is 4,451,496,278). In considering that there are $k!/2$ distinct qualitative J orders, it should be clear that the total number of possible quantitative J scales is very large.

Unfolding can be defined as follows: from the set of individual rankings, we wish to determine the J scale on which individuals as well as options are ordered. Individual rankings then correspond to so-called *folded J* scales. This can be explained as follows. When the J scale is picked up and folded in any (ideal) point, a folded J scale arises: the options project on the folded J scale in order of increasing distance from the folding point, and the first option corresponds to the ideal option (see Figure 3). The number of admissible orderings on a qualitative J scale is 2^{k-1}, and is equal to the number of ways folded J scales can be constructed (Coombs, 1964; Davison, 1979). For a quantitative J scale this number is $\binom{k}{2} + 1$ (i.e., the number of midpoints plus one).

For a variety of reasons, however, individuals generally do not all produce rankings consistent with one underlying qualitative or quantitative J scale, and a variety of methods have been developed for the unfolding of imperfect data. Many start from a distance model and a metric or non-metric loss function (e.g., Roskam, 1968; Carroll, 1972; Heiser, 1981; 1987); others rely on parametric functions to describe the choice probabilities (e.g., Sixtl, 1973; Jansen, 1983; Desarbo and Hoffman, 1986; Andrich, 1988, 1989; and Formann, 1988). Almost invariably, the latter techniques are suited for dichotomous data only.

Our approach differs from existent procedures since neither a parametric distance model nor a parametric function is used to describe individuals' preferences, and instead, a minimum-number-of-inversions criterion is used. This criterion has not yet been used in unfolding analysis. The procedure begins with a nonparametric distance measure related to Kendall's (1970) τ: the number of inversions between an individual's ranking and the qualitative or quantitative J scale. For each distinct J scale, the number of inversions between each individual ranking and each admissible ordering of the J scale is assessed; each individual ranking is assigned the admissible ordering from which it has a minimum number of inversions. Thus, for *each* individual ranking the number of inversions from *each* admissible ordering on *each* qualitative or quantitative J scale has to be assessed. The best J

scale is the scale for which the total number of inversions from individuals' rankings is a minimum. We shall not go into the specific combinatorial optimization strategy that is used to find the best *J* scale. This is discussed in Van Blokland (1991).

In the following, we show that the unidimensional unfolding model can be placed in the context of social choice theory (section 13.2), distance measures for rankings (section 13.3), and probability models on rankings (section 13.4). In the original Coombsian unfolding model, single-peakedness is assumed for all individual preference functions. In this case, the median ordering is a group consensus ordering and is an admissible ordering of the qualitative or the quantitative *J* scale. We prove that the requirement of exclusively SPF's is unnecessary strong, and that the same holds without this restriction. This is shown to hold at least if rankings follow a strongly unimodal distance model for rankings (Section 13.5). For this case, we also prove that the best quantitative *J* scale according to the criteria of Maximum Likelihood (ML) and Minimum-Number-of-Inversions (MNI) are the same (Section 13.6). Section 13.7 presents two illustrations of the theoretical results, Section 13.8 closes with a discussion.

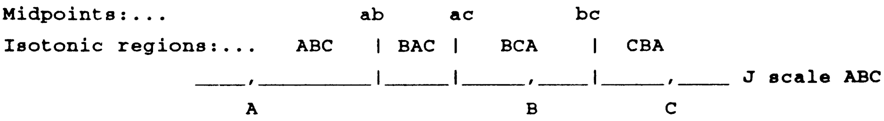


Figure 1 *The 3-scale ABC*

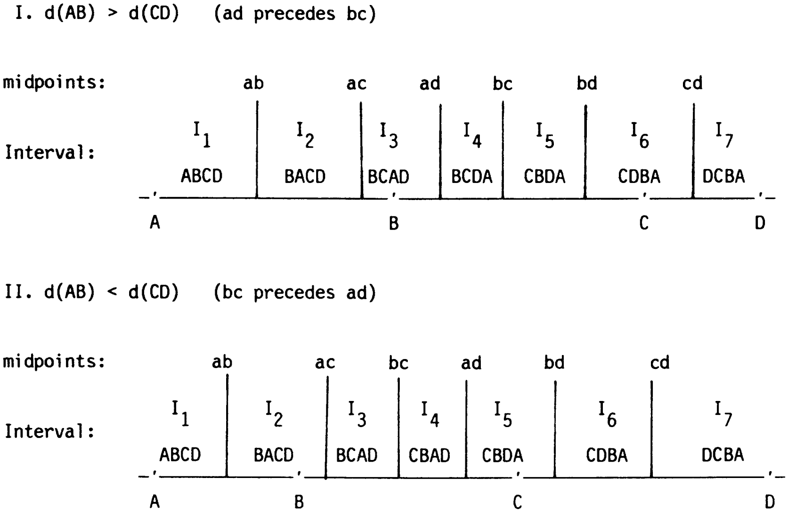


Figure 2 *The two possible midpoint orders for 4-scale ABCD; ad precedes bc (top) and bc precedes ad (bottom)*

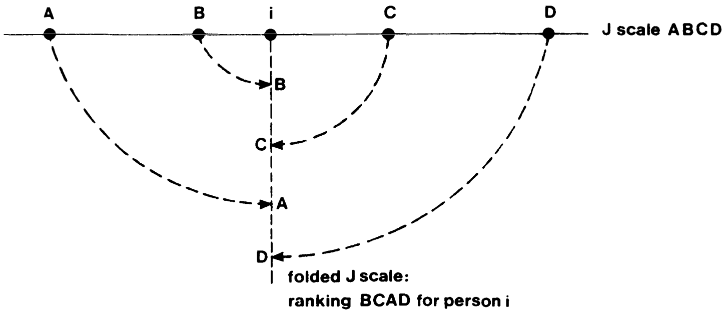


Figure 3 *J scale ABCD and folded J scale BCAD for individual i*

13.2 Social Choice Theory

A main issue in the field of welfare economics has been the construction of a social preference out of a variety of individual preferences. The problem is to establish a fair procedure to combine the individual rankings to reach a group consensus ordering. This problem is related to the unfolding problem, since there is a strong connection between the existence of a unidimensional unfolding scale, the conditions under which group consensus orderings are transitive orderings, and uniqueness of median, modal, and mean orderings. These traditions and its ramifications are amply described in Luce and Raiffa (1957); Sen (1982); Fishburn (1972); Riker and Ordeshook (1973); Vickrey (1960); Roberts (1976); Fligner and Verducci (1986); Critchlow, Fligner and Verducci (1989), and many others. These subjects are discussed in this section, as far as they relate to unfolding. First, we present some definitions and notation; after this we proceed with an overview of the main results from social choice theory, namely, the conditions a group consensus function should satisfy and the specific function that best satisfies these conditions. It turns out that the optimal group consensus function is Simple Majority Rule which yields the median ordering on the quantitative J scale as a group consensus ordering, given single-peakedness of all individual preference functions. The relaxation of this restriction is the subject of the next section.

Definitions:

- O is the set of options (alternatives),
- $O = \{a, b, c, d \ldots\}$; when referring to unfolding, $O = \{A, B, C, D, \ldots\}$ is used.
- S is the set of individuals, to be labelled $1, 2, \ldots, i, \ldots n$.
- P_i is the ranking of the ith individual $(1 \leq i \leq n)$.

- aP_ib abbreviates the statement that i prefers a to b, where $i \in S$ and $a, b \in O$.
- Given a group, the rankings $(P_1, P_2, \ldots, P_n)$ of the members of the group define the group's *profile.*
- aI_ib denotes the statement that i is indifferent between a and b.
- Weak preference, the 'at least as good as', $aPb \cup aIb$, is a weak order.
- A *strict weak order* is one in which aPb and aIb cannot both occur.
- A relation is *transitive* if $(a, b) \in P$ and (b, c) then $(a, c) \in P$, for $a, b, c, \in O$.
- The relation "P or I" is a transitive relation and is called a *partial order.*
- A strict partial order is irreflexive and transitive.
- a and b are *preference-comparable* (comparable or connected), iff aPb or bPa (or both), that is, if and only if $(a, b) \in P \cup P^{-1}$.
- a and b are preferentially *incomparable* if $(a, b) \notin P \cup P^{-1} \Leftrightarrow (a, b) \in \neg(P \cup P^{-1})$, or: (a, b) belongs to the complement of the relation $P \cup P^{-1}$.
- A linear ordering P of O is a partial ordering in which each pair of options is comparable, i.e., aPb or bPa for all $a, b \in O$.
- A relation is *asymmetric* if $(a, b) \in P$, then $(b, a) \notin P$, for $a, b \in O$.
- The relation of strict preference $a >_p b$ is transitive and asymmetric.
- A *complete ranking* is a strict weak ordering without ties.

Group Consensus Orderings:
Given a profile of rankings $P_1, \ldots, P_n$, the problem is to find a ranking P on O that represents the *group consensus ordering.* A rule for determining the group consensus ordering from a group profile is called a *group consensus function.* Examples are Simple Majority Rule (Condorcet, 1785), Borda Rule (Borda,1781), Plurality Rule (Malkevitch et al., 1974) and Lexicographical Rule (Tversky, 1969). The most important rule is Simple Majority Rule: given a profile $P_1, \ldots, P_n$, let the group rank option a over b if and only if a majority (more than half) of the individuals ranks a over b. Simple Majority Rule, however, sometimes fails to yield a transitive ordering. This phenomenon is called the *voting paradox*, and goes back as

far as Borda and De Condorcet. For example, let $O = \{a, b, c\}, n = 3$ and the group profile is $P_1 = abc, P_2 = bca$, and $P_3 = cab$. The group ranking P would then have aPb, bPc, and cPa. Since a ranking is asymmetric by definition (see above), such a group ranking is impossible, it would not be transitive.

Arrow (1951) stated five conditions a 'fair' group consensus ordering should satisfy: (1) Unrestricted Domain for Preferences (all possible preference rankings are permissible); (2) Positive Association of Social and Individual Values (if aP_ib for every i, then the social outcome is aPb); (3) Independence of Irrelevant Alternatives (the social outcome remains the same if an option is deleted); (4) Citizen's Sovereignty (the social outcome is not imposed by some kind of government); (5) Non- Dictatorship (the social outcome is not determined by a single individual). Simple Majority Rule is the only rule that satisfies all of Arrows' conditions, except the implicit requirement of being a unique, transitive ordering; this may be violated. In posing restrictions on the domain of rankings (thereby violating Arrow's first condition, the unrestrictedness of preferences), Black (1948a,b) and Arrow (1951) proved the following important result. Under the condition that the individual preference rankings be SPF's on a unidimensional continuum, it holds that the top choice of the median (middle) individual yields the social ideal, the option that is most preferred by the group as a whole: the *group consensus*.

This group decision process applies if the options are strictly ordered, that is, for a *qualitative* J scale. Coombs (1954) and Goodman (1954) showed that an analogous, but stronger, assertion holds. They proved that the ranking of the median individual equals the group consensus ordering according to Simple Majority Rule iff the individual preference rankings are restricted to be SPF's on a *quantitative* J scale. If the J scale is folded downwards in the ideal point of the median individual (see section 13.1), the preference ranking of the median person arises as a folded J scale. The options project onto this folded J scale in order of increasing distance from the median individual's ideal option.

The epitome of Black's, Arrow's, Coombs', and Goodman's conditions is, that social preferences are strictly ordered if some sort of inner harmony exists among choosers. The existence of single- peaked functions reflects a common cultural uniformity about the standard of judgment, even though people differ about what ought to be chosen under that standard (Riker & Ordeshook, 1973, p. 105; Coombs, 1964, p. 397). The generalization of Coombs' and Goodman's conditions to preferences that are not necessarily single-peaked functions is the main purpose of this paper. To this end, we first show that the median ordering is a group consensus ordering in general, not only for rankings that are restricted to be SPF's on a quantitative J scale. Next we show that it holds quite generally that the median ordering is a folded J scale if the rankings satisfy a strongly unimodal distance model for rankings. The unimodal distance model for rankings can be seen as a nonparametric analogue of the normal distribution for real numbers on a line. In doing so, we wish to create a frame of reference to interpret the results of the unidimensional unfolding model in the context of social choice theory. This is the subject of the next sections.

13.3 Distance Measures for Rankings

Kemeny (1959) and Kemeny and Snell (1972) presented an axiomatic approach to arrive at a unique distance measure for rankings and to define a group consensus ordering in terms of this distance. In their approach, rankings are strict weak orderings, and are represented as points in geometrical space. Rankings that differ in the order of only two options are connected by a line; the ranking that has both options tied lies on the line between them.

Kemeny's distance function is based on May's (1952) paired comparisons distance and is defined as follows. Suppose P and Q are rankings and a and b are options in O. Define $D_P(a,b) = 1, 0, -1$, according as aPb, aIb, or bPa. Then for all a, b in O

$$D_{P,Q}(a,b) = |D_P(a,b) - D_Q(a,b)|,$$

and the function $d(P,Q)$ defined as

$$d(P,Q) = \sum_{a,b} D_{P,Q}(a,b)$$

provides the unique distance measure d that satisfies all axioms. This distance is called the *Kemeny distance.* For strict linear orders, this distance is equal to the Kendall (1970, Ch. 2) τ-distance, which was used to define the correlation coefficient τ (see Bogart, 1973). Kemeny and Snell defined two group consensus orderings: the *median ordering* and the *mean ordering.* The median ('Med') is defined as the ordering for which the sum of all distances $d(\text{Med}, Q_i)$ $(i = 1, \ldots, n)$ is a minimum and the mean ('Mean') is defined as the ordering for which the sum of all squared distances $d(\text{Mean}, Q_i)^2$ is a minimum:

$$\text{Median:} \qquad \sum_i d(\text{Med}, Q_i) \qquad \text{is minimal}$$

$$\text{Mean:} \qquad \sum_i d(\text{Mean}, Q_i)^2 \qquad \text{is minimal}$$

These definitions for rankings are analogous to those from classical statistics for real numbers on a line: the median ordering is the mid-ordering on a scale and the mean ordering represents the center of gravity for all rankings. However, Kemeny and Snell did not show why medians or means should be taken as group consensus orderings (see Roberts, 1976). Moreover, in the situation in which the voter's paradox occurs, neither the median nor the mean yields a unique ordering.

The results of Bogart (1973; 1975) demonstrate the benefits of the median ordering as a consensus ordering: if Simple Majority Rule applied to (partial, weak, or linear) orderings gives rise to an ordering, then this ordering is the unique median with respect to the city-block metric unless for some options a and b the number of individuals preferring a to b is equal to the number preferring b to a. A similar result was obtained for the mean ordering with respect to the Euclidean metric. Essentially, Bogart showed that medians and means are unique precisely when there is a majority winner for each pair of options, that is, precisely when the situation of the voting paradox does not occur.

13.4 Strongly Unimodal Distance Models for Rankings

Some widely used examples of distance models for rankings based on paired comparisons are the Thurstone (1927) Case V model (see Torgerson, 1958, Ch. 9), and Mallows' (1957) model. The latter model includes two submodels; one is popularized by Schulman (1979) and is based on Spearman's ρ-distance. Another submodel is Mallows' ϕ-model which is based on Kendall's (1970) τ-distance. This model is completely worked out by Feigin and Cohen (1978), who also provide maximum likelihood estimates for model parameters and furnished tables for the distribution of the number of inversions from the central ordering. Therefore, we shall refer to this model as Feigin and Cohen's model and we use their notation.

Feigin and Cohen's model assigns probability to a ranking in inverse relation to its Kendall τ-distance from some central ordering. The model is a nonparametric one: it does not depend on parametric assumptions, and is based on a simplification of the Bradley-Terry (1952) model by the following assumption: rankings ω with the same Kemeny or Kendall distance from the central ranking ω_0, have the same probability. For an overview of distance based ranking models in the paired comparisons tradition, see Fligner and Verducci (1988), Critchlow, Fligner, and Verducci (1988). In the following paragraphs, Feigin and Cohen's model is overviewed.

The distribution in Feigin and Cohen's model is based on the number of inversions between rankings. An individual ranking is denoted $\omega = (\omega_1, \ldots, \omega_j, \ldots, \omega_k)$ and is a permutation of the numbers $(1, \ldots, j, \ldots, k)$, in which $\omega_j (j = 1, \ldots, k)$ denotes the jth option. The distribution of ω depends on one discrete and one continuous parameter, denoted ω_0 and θ, respectively. The first is a permutation that acts as a location: $\omega_0 = (\omega_{01}, \ldots, \omega_{0j}, \ldots, \omega_{0k})$ and is denoted the *central ordering* in which

ω_{oj} denotes the jth option. The second parameter θ is a non-negative dispersion parameter ($0 \leq \theta \leq 1$). The number of inversions between a ranking ω and the central ordering ω_0 is denoted $X(\omega_0, \omega) = x$. The *probability distribution* of a ranking ω is:

$$P_{\omega_{0,\theta}}(\omega) = (f(\theta))^{-1}\theta^x, \qquad 0 \leq \theta \leq 1 \tag{1}$$

where $f(\theta) = \sum_\omega \theta^x$, a normalizing constant so that the probabilities sum to 1.

There are a^k_x permutations with x inversions for fixed k (Kendall, 1970, 5.2, where a^k_x is termed $u(n,s)$). Consequently, the probability distribution of the number of inversions $X(\omega_0, \omega) = x$ is:

$$P_\theta(X = x) = (f(\theta))^{-1} a^k_x \theta^x, \qquad x = 0, 1, \ldots, \binom{k}{2} \tag{2}$$

where $\binom{k}{2}$ = maximum number of inversions,
a^k_x = number of orderings with x inversions from ω_0 ,
$f(\theta) = \sum_x a^k_x \theta^x = \sum_\omega \theta^x$, a normalizing constant.

With increasing values of θ, the concordance decreases; $\theta = 0$ corresponds to complete concordance and $\theta = 1$ corresponds to random selection of rankings.

Strong Unimodality

Let $\omega = \omega_1, \omega_2, \ldots, \omega_k$ denote an arbitrary ranking of k options. Let the probability function $P(\omega)$ represent a probability model on rankings, in particular, let $P_\theta(\omega)$ stand for the Feigin and Cohen model. A ranking ω_m is a *modal ordering* if it uniquely maximizes $P(\omega)$. A probability model on ranking data is *strongly unimodal* if it has a modal ordering ω_m, and the probability $P(\omega)$ is nonincreasing as ω moves farther from ω_m along a certain type of path (see Critchlow, Fligner, and Verducci, 1988). This is an analogy with the usual definition of strong unimodality for univariate probability distributions. Feigin and Cohen's model is strongly unimodal, since the probability $P_\theta(\omega)$ decreases according to increasing Kendall or Kemeny distance from the central ordering $\omega_0 (0 \leq \theta \leq 1)$, hence, the median and modal ordering coincide here.

The maximum likelihood (ML) estimator for $\omega_0, \hat{\omega}_0$, is the ordering for which the total number of inversions with respect to all rankings, $\sum_i X$ (ω_0, ω_i), is a minimum. Thus, ω_0 is a central ordering according to both the ML and the MNI criteria .Therefore, estimates based on the ML criterion and on the MNI criterion yield equal results under Feigin and Cohen's model. For the simplest case of three options, A, B, and C, and $\hat{\omega}_0 = ABC$, this equality of results is represented in Table 1. The ML estimate for θ given $\hat{\omega}_0$ is the value of θ for which the mean number of inversions $\bar{x}$ equals $E_\theta(X)$ (see Feigin and Cohen, 1978).

Table 1.

Probabilities of all rankings of three options A, B, and C as a function of the number of inversions X from the median ordering ABC and selected values of the dispersion parameter θ from the Feigin and Cohen model.

Ranking	X	π_x	$\theta = .10$	$\theta = .20$	$\theta = .50$	$\theta = .90$
ABC	0	π_0	.819	.672	.381	.194
ACB	1	π_1	.082	.135	.190	.175
BAC	1	π_1	.082	.135	.190	.175
BCA	2	π_2	.008	.027	.095	.157
CAB	2	π_2	.008	.027	.095	.157
CBA	3	π_3	.001	.005	.048	.142
Total			1.000	1.001	.999	1.000

Table 2.

Admissible orderings for three possible 3-scales with options A, B, and C. The median ordering ABC is marked with '$\star$'. The last column gives the probabilities of the admissible orderings given Feigin and Cohen's model with $\theta = .20$.

Ranking		J Scale		Probability
ω	ABC	BAC	ACB	$\theta = .20$
ABC	$\star ABC$	$\star ABC$	-	$\pi_0 = .672$
BAC	BAC	BAC	-	$\pi_1 = .135$
BCA	BCA	-	BCA	$\pi_2 = .027$
CBA	CBA	-	CBA	$\pi_3 = .005$
ACB	-	ACB	ACB	$\pi_1 = .135$
CAB	-	CAB	CAB	$\pi_2 = .027$
Total	.839	.969	.194	1.001

13.5 Generalization of Coombs' and Goodman's Conditions

In the following paragraphs, we show that the median ordering is a folded J scale under rather general conditions. In particular, we show that the group consensus ordering is a folded J scale given that the rankings satisfy a strongly unimodal probability model for rankings. This constitutes the generalization of Coombs' and Goodman's conditions.

Niemi (1969) defined the *dominant J scale* as the scale that satisfies the largest proportion of individual rankings, and proved that, for the case of

3-scales containing at least 2/3 of the individual rankings, the probability that the group consensus ordering is a folded J scale, approaches one if the number of individuals is sufficiently large. Niemi proved that the probability of an intransitive group consensus ordering decreases monotonically as the proportion of preference rankings satisfying a common 3-scale increases. In Niemi's approach, the data have to satisfy an *equally probable* condition: all *admissible* orderings on the one hand and all *inadmissible* orderings on the other are assumed to have the same probability. Even when individuals share a common reference frame, the condition that all possible admissible orderings have equal probabilities seems unrealistic. This would correspond to a uniform distribution of individuals on the J scale, no matter how large the proportion of individuals who state single-peaked preferences. In fact, Niemi's (1969) conditions seem to imply random choice, within two disjoint sets of individuals: those who choose admissible orderings and those who choose inadmissible orderings.

A more general approach could start from a *non-null concordance* condition and a nonparametric distance measure that is based on the number of inversions between rankings. By using a unimodal distance model for ranking data, it can be proved that the median ordering is a folded J scale in general. Any ranking model for which probabilities of rankings strictly decrease with increasing numbers of inversions from the median ordering can be used. Feigin and Cohen's model is suited for this purpose. In the following, we prove that the median ordering is a folded J scale if the data follow the distribution specified by Feigin and Cohen's model. First, we introduce some notation and definitions.

Let a *median-compatible J scale* denote a J scale that includes the median ordering as an admissible ordering and let a *non-compatible J scale* denote a J scale that does *not* include the median ordering as an admissible ordering. Without loss of generality, in the following paragraphs it is assumed that $ABCD...$ is the median or modal ordering ω_0.

The probability of an arbitrary ranking ω with x inversions from the median ordering ω_0 is given by (1) and geometrically decreases with the number of inversions from ω_0. Let π_x denote the probability of a ranking with x inversions from $\omega_0 : \pi_x = \theta^x / f(\theta)$, from (1), and can be found from the Appendix using the relation $\pi_x = P_\theta(X = x)/a_x^k$, from (2). The probabilities of the six possible rankings for three options given selected values of θ were presented in Table 1. Since ABC is the median ordering, it has the largest probability (π_0). The two rankings with one inversion from ABC : ACB and BAC, have the second-largest probabilities (π_1). Rankings with two inversions from ABC : BCA and CAB, have the next largest probabilities (π_2). Lastly, CBA has three inversions from the median ordering ABC and has the smallest probability (π_3).

Now, let's introduce possible J scale orderings. As is shown in Table 2, there are three distinct J scales for three options (all other ones are mirror images of these): one with B in the middle, one with A in the middle, and one with C (the least preferred option) in the middle. From Table 2, it can be verified that if the median ordering is ABC and the data satisfy Feigin and Cohen's model, the admissible orderings of the median-compatible 3-scale BAC have the largest probability: $\pi_0 + 2\pi_1 + \pi_2$, which is $.672 + 2(.135) + .027 = .969$ if $\theta = .20$. Hence, 3-scale BAC is called the **dominant** J scale, following Niemi (1969). This result does not only hold for $\theta = .20$. From Figure 4, it can be seen that for $0 < \theta < 1$, and given the median ordering ABC, 3-scale BAC has the largest probability, while the non-compatible 3-scale ACB has the smallest probability.

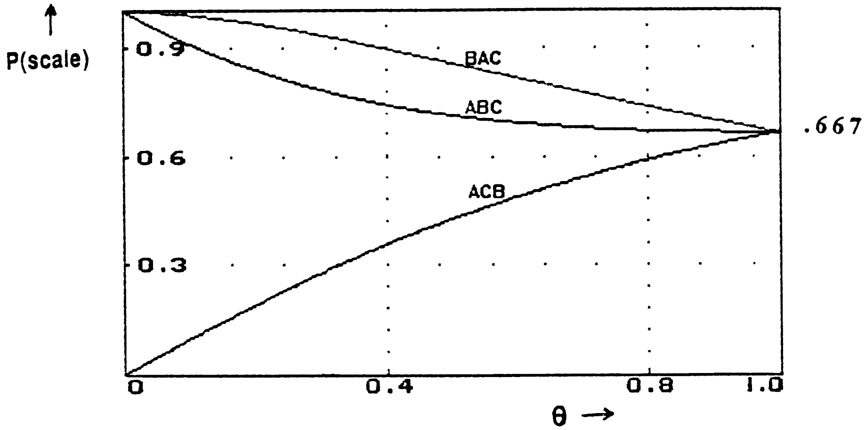


Figure 4 *Proportions of individuals who satisfy each of the three possible 3-scales, given the F & C model* $(0 \leq \theta \leq 1)$ *and median ordering* ABC

From the above example, we conclude three things. If the conditions of Feigin and Cohen's model are met and if the median ordering is ABC, the median-compatible 3-scale BAC is the *dominant* or ML 3-scale in the first place. If $\theta = 1$, we cannot discriminate between the 3-scales; they have the same probability and we expect each option A, B, and C to occur in all positions (first, second, and last) with equal probability, which is precisely the standard example of the voting paradox! Therefore, the median ordering is not unique, it can be any of the three possible rankings; or rather, the median ordering is a three-way tie in this degenerate case, except for random differences in the rankings. Secondly, the *social* preference function is single-peaked on the J scale, since the probabilities of the admissible orderings of 3-scale BAC decrease away from $\omega_0 = ABC$, on both sides (the admissible orderings are BAC, ABC, ACB, and CAB, having probabilities $\pi_1, \pi_0, \pi_1, \pi_2$, resp.). Third, the admissible orderings of the ML 3-scale have, in total, the largest probability, since the probability of a ranking decreases monotonically with increasing numbers of inversions away from the median ordering. Therefore, the ML 3-scale is at the same time the MNI 3-scale. This can be proved in general, using the strong unimodality property of Feigin and Cohen's model. To prove this, we need only the strong unimodality property of the Feigin and Cohen model, therefore, these results can be generalized to any strictly unimodal distance model for rankings for any number of options. This is the subject of the next section.

13.6 Equal Results for ML or MNI Criterion

In this section, we show that Coombs' (1954), and Goodman's (1954) results can be generalized to the case of k options under rather general probabilistic conditions: a strongly unimodal distance model for rankings. In particular, we wish to show that the median or modal ordering is a folded qualitative or quantitative J scale when individuals' preference rankings are not all single-peaked functions.

First, we need some preliminary theoretical results about the midpoint order, the distances of admissible orderings with respect to the median ordering on a non-compatible scale, and the single-peakedness of the social preference function on the J scale. After having established these, we proceed to prove the proposition for k options.

Definitions

The admissible orderings of a quantitative J scale constitute a subset of those of the qualitative J order. Thus, if the median ordering proves to be a folded quantitative J scale, it is a folded J order. Therefore, we consider only *quantitative* J scales, and a 'J' scale will stand for a quantitative J scale in this section. Let σ_{st} denote the inversion of two adjacent options S and T in any permutation, then $\sigma_{st} \cdot P$ denotes that σ_{st} is applied to ranking P to produce the inversion of S and T in P. Let $p_i (i = 1, 2, \ldots, \binom{k}{2} + 1)$ denote an admissible ordering on the quantitative J scale, and let $X(\omega_0, p_i)$ denote the number of inversions (the *distance*) between the ith admissible ordering and the median ordering; also, let $P_\theta(p_i)$ denote the probability of p_i given Feigin and Cohen's model. A $w - non - compatible J$ scale is a non-compatible J scale for which the admissible orderings have at least w inversions from the median ordering ω_0, and $w = \min_i X(\omega_0, p_i)$, where $i = 1, 2, \ldots, \binom{k}{2} + 1$. A non-compatible J scale denotes a 1-non-compatible J scale. Without loss of generality, it is assumed that $ABCD\ldots$ is the median ordering ω_0.

Midpoint Order in a k-Scale

The 3-scale is the smallest scale that imposes restrictions on the rankings: two out of six rankings are inadmissible for each possible 3-scale (see Table 2). Therefore, the 3-scale is taken as the smallest J scale.

Each quantitative scale is a path along admissible orderings. A step along such a path moves from an arbitrary admissible ordering p_i to $p_{i+1} = \sigma_{st} \cdot p_i$, where options S and T are adjacent in p_i. The step is *away from* ω_0 if p_i and ω_0 initially agree on the order of the options S and T, and *towards* ω_0 otherwise. Since each option pair may be interchanged only once (there is one midpoint between two options), the interchanges are disjoint. Thus, each successive interchange yields one more inversion from p_1, the first admissible ordering on the J scale. The order of interchanging options on the J scale is from left to right, and corresponds to the midpoint order on the scale. This was illustrated in the 3-scale ABC in section 13.5. If the J order is ABC, the midpoints are necessarily ordered $ab \longrightarrow ac \longrightarrow bc$, where '$\longrightarrow$' denotes 'precedes'. This follows from the positions of the options on the J order, and can be shown as follows. If A is the first option on the scale, and B precedes C on the scale, the midpoint between A and $B(ab)$ must precede the midpoint between A and $C(ac)$. In general, with a J order $ABCD...YZ$, the midpoint order must satisfy the restrictions:

$$ab \longrightarrow ac \longrightarrow ad \longrightarrow \ldots \longrightarrow ay \longrightarrow az;\ bc \longrightarrow bd \longrightarrow \cdots \longrightarrow by \longrightarrow bz, \ldots \qquad (3)$$

In the same way, the midpoint order must satisfy analogous restrictions when keeping the last option on the J scale fixed:

$$ad \longrightarrow bd \longrightarrow cd; az \longrightarrow bz \longrightarrow cz \longrightarrow \ldots \longrightarrow yz. \qquad (4)$$

w-Non-Compatible Scales

It takes three options to build a non-compatible scale: a non-compatible 3-scale has the *least* preferred option in the *middle* position (this follows from the definition of an SPF, see section 13.1). Therefore, violations of the unfolding model can be detected by inspection of triples of options (cf. Dijkstra, 1978; Van Schuur, 1984). Using the restrictions on the order of the midpoints (from (3) and (4)), we can derive the successive distances $X(\omega_0, p_i)(i = 1, \ldots, \binom{k}{2} + 1)$ on a non-compatible J scale. This is illustrated for a small selection of cases. To simplify matters, we consider only non-compatible scales that comprise the non-compatible 3-scale ACB. The results are easily generalized to cases in which not B and C, but another pair of options is in the wrong order on the J scale.

At the moment, we are concerned with non-compatible J scales only. Therefore, additional options must be adjoined to the non-compatible 3-scale ACB in order of decreasing preference away from the most preferred option A, on both sides. To show this, we prove the following propositions.

Proposition I

The options on a median-compatible J scale are in order of increasing social preference towards the socially ideal option A and in order of decreasing social preference away from it. This implies that the social preference function is unimodal on the J scale.

Proof

If the scale is median-compatible, the median ordering ABC is an admissible ordering, and is thus a folded J scale. Since the median ordering is a group consensus ordering, it holds the options in order of decreasing social preference. In folding the J scale in A, the options to the right of A project on the folded J scale in order of increasing distance from A, hence, they must be in order of decreasing social preference away from A on the J scale. An analogous reasoning holds for options to the left of A. Hence, the social preference function is unimodal on the J scale. □

Proposition II

The options on a w-non-compatible J scale are in order of increasing social preference towards the socially ideal option A and in order of decreasing social preference away from it, apart from w inversions.

Proof

A non-compatible J scale arises if there is a triple of options where the least preferred option holds the middle position. Thus, the interchange of A and B in ω_0 does not give rise to a non-compatible scale, nor can it contribute to the non-compatibility of the J scale: only the placement of option C between A and B yields a non-compatible 3-scale. If the k-scale is w-non-compatible, then, by definition, there is an admissible ordering Q such that $X(\omega_0, Q) = w$, and Q is a folded J scale. Thus, there are w inversions $\sigma_{s_j t_j}$ of adjacent options S_j and $T_j (j = 1, \ldots, w)$, such that

$$Q = \sigma_{s_w t_w} \bullet \ldots \bullet \sigma_{s_2 t_2} \bullet \sigma_{s_1 t_1} \bullet \omega_0.$$

These inversions are away from ω_0, but, as above, cannot concern A and B. Therefore, the interchanges concern pairs of adjacent options either to the right or to the left of A on the J scale. In folding the J scale in A, the options project on the folded scale in order of increasing distance from A. Hence, if two options are ordered away from ω_0 in Q, they must be so on the J scale; if their order in Q agrees with that in ω_0, they must be in order of decreasing preference on the J scale, away from the social ideal A (from Proposition I). □

From Proposition II, it follows that we can construct non-compatible J scales out of the 3-scale ACB by adjoining options in order of increasing preference towards ACB and in order of decreasing preference away from it. This can be illustrated in the following 5-scales: J scales ED**ACB** and **ACB**DE are 1-non-compatible, J scale DE**ACB** is 2-non-compatible. This can be verified by folding the J scale at A.

Distances from ω_0 : Non-Compatible J Scales

Using the above results, we can now derive the successive distances $X(\omega_0, p_i)(i = 1, \ldots, \binom{k}{2}+1)$ on a non-compatible J scale for different placements of the non-compatible triple ACB on the J scale. This is illustrated for $k = 3, 4, 5$.

Case k=3

If the median ordering is ABC, only the placement of option C between A and B yields a non-compatible 3-scale. In the following, only 3-scale ACB is considered; analogous results are obtained in considering its mirror image: 3-scale BCA. The admissible orderings of 3-scale ACB are ACB, CAB, CBA, BCA, and have $X = 1, 2, 3, 2$ inversions, resp., from $\omega_0 = ABC$ (see Table 1). From (3) and (4), it follows that the midpoint order on 3-scale **ACB** is:

$$ac \longrightarrow ab \longrightarrow bc, \tag{5}$$

thus, options C and B cannot be interchanged directly. Each step along the 3-scale ACB involves the interchange of two adjacent options and moves away from p_1 since the interchanges are disjoint. Let $Q = \sigma_{bc} \bullet \omega_0$, thus $X(\omega_0, Q) = 1$. Here, $Q = p_1$, and each step moves away from Q. The first two steps move away from ω_0 too, but in p_3, B and C are adjacent and the third step involves their interchange (from (5)); therefore, since the order is CB in p_3 and BC in ω_0, the third step is *towards* ω_0. The distances $X(\omega_0, p_i), i = 1, 2, 3, 4$, are given below:

$$[1\ 2\ 3\ 2], \tag{6}$$

and $X(\omega_0, p_i) > 1$ if $p_i \neq Q$.

Case k=4

If the J scale is **ACB**D , then Q is the first admissible ordering on the scale. To obtain remaining admissible orderings, options A, C and B have to be interchanged as in 3-scale ACB (from (5)), and it takes three inversions $(D, A; D, B; D, C)$ to interchange D with remaining options. Hence, the distances $X(\omega_0, p_i), i = 1, 2, \ldots, 7$, are resp.:

$$[1\ 2\ 3\ 2]\ 3\ 4\ 5.$$

If the J scale is D**ACB**, it takes three inversions $(D, A; D, C; D, B)$ to reach $Q = \sigma_{bc} \bullet \omega_0$; after this, A, B, and C have to be interchanged as in 3-scale ACB (from (5)); hence, the distances $X(\omega_0, p_i), i = 1, 2, \ldots, 7$, are resp.:

$$4\ 3\ 2\ [1\ 2\ 3\ 2].$$

The numbers in the brackets correspond to those of the constituting non-compatible 3-scale ACB, (from (6)).

Case k=5

If the J scale is ED**ACB**, it takes four inversions for E and three inversions for D to reach $Q = \sigma_{bc} \bullet \omega_0$; after this, A, B, and C have to be interchanged (as before). Hence, the distances $X(\omega_0, p_i), i = 1, 2, \ldots, 11$, are resp.:

$$8\ 7\ 6\ 5\ 4\ 3\ 2\ [1\ 2\ 3\ 2].$$

In the same way, these distances $X(\omega_0, p_i), i = 1, 2, \ldots, 11$, are

$$[1\ 2\ 3\ 2]\ 3\ 4\ 5\ 6\ 7\ 8\ 9,$$

for the J scale **ACB**DE. As examples, the admissible orderings of the median-compatible 4-scale $CBAD$ and the non-compatible 4-scale $DACB$ are shown below with the number of inversions from the median ordering:

median-compatible 4-scale:
CBAD BCAD BACD ABCD ABDC ADCB DACB distance: 3 2 1 0 1 2 3

non-compatible 4-scale:
DACB ADCB ACDB ACBD CABD CBAD BCAD distance: 4 3 2 1 2 3 2

Distances from ω_0 : *w-Non-Compatible J Scales*

The above derivations and examples concerned 1-non-compatible J scales. For w-non-compatible scales, the same rules (from (3) and (4)), and rules analogous to (5) and (6) apply, the main difference being now that we start with a minimum distance of w inversions from the median ordering. Using Proposition II, it follows that w pairs of options are ordered away from ω_0 on the J scale. There is an admissible ordering p_i, say Q, such that $X(\omega_0, Q) = w$, and each step on the J scale from p_i to p_{i+1} (to the right of Q) or to p_{i-1} (to the left of Q) moves farther away from Q. This step moves away from ω_0 too, if p_i and ω_0 initially agree on the order of the options concerned, but *towards* ω_0 otherwise. Therefore, there are w steps *towards* ω_0, In other words, w times a pair of options is interchanged whose inversion is now **towards** ω_0. As an illustration of a 2-non-compatible scale, 4-scale **BDA**C is given below:

2-non-compatible 4-scale:
BDAC DBAC DABC ADBC ADCB ACDB CADB distance: 3 4 3 2 3 2 3

Stochastic Ordering of Permutations

Let D be a ranking in which the options are represented by the alphabetical sequence of letters $a, b, c, \ldots, z$, and let $E, F, \ldots, T$ denote different permutations of these letters. Let D and E differ only in the inversion of two letters and let D have a greater probability than E , then we write $D > E$. If there is a sequence of permutations $D, E, F, \ldots, T$ in which we proceed from one permutation to the next by interchanging two letters which were in alphabetical order, then $D > E > F > \ldots > T$, and we say that D and T are *stochastically ordered* (see Henery, 1981). As an example, 4-scale $CBAD$ is given below, with numbers of inversions from the median ordering $ABCD$ and the probabilities of its admissible orderings. These orderings are stochastically ordered, away from ω_0, on both sides of ω_0.

Ordering:	CBAD	BCAD	BACD	ABCD	ABDC	ADBC	DABC
Inversions:	3	2	1	0	1	2	3
Probability:	π_3	π_2	π_1	π_0	π_1	π_2	π_3

Dominance of the Median-Compatible J Scale

Theorem

The dominant J scale is median-compatible if the data follow the distribution specified by Feigin and Cohen's model.

Proof:

Suppose the data follow the distribution specified by Feigin and Cohen's model. In the sequel, r stands for $[\binom{k}{2}/2]$, where $[x]$ denotes the largest integer not exceeding x. Without loss of generality, it is assumed that $ABCD...$ is the median ordering ω_0, and that $\binom{k}{2}+1$ is odd. Hence, $r = \binom{k}{2}/2$. Analogous arguments hold true if $\binom{k}{2}+1$ is even. The proof consists in showing that the total probability of admissible orderings from an optimal median-compatible J scale is larger than the total probability from an optimal non-compatible scale.

Probability for the median-compatible J scale

If the k-scale is median-compatible, ω_0 is an admissible ordering of the J scale and, in principle, can have all possible positions on the J scale. Let $I_i (i = 1, 2, \dots 2r + 1)$ denote the position of ω_0 on the scale, then $X(\omega_0, p_i|I_i)$ represents the number of inversions between each admissible ordering and the median ordering, given the position of the latter on the J scale. These distances are given in Table 3.

Since the Feigin and Cohen model has the strong unimodality property, $P_\theta(p_i)$ increases towards ω_0 and decreases away from ω_0 on the J scale (ω_0 and p_1 are stochastically ordered, as are ω_0 and p_{2r+1}). Hence, the special J scale that includes ω_0 in the midranking has the largest probability, and must be the dominant J scale if the data satisfy Feigin and Cohen's model. No other median-compatible J scale can have a larger probability. Let the *total probability* of the J scale denote

$$P_{p_i} = \sum_i P_\theta(p_i) \quad (i = 1, 2, \dots \binom{k}{2} + 1) \tag{7}$$

then for this scale, the total probability of the admissible orderings is:

$$P_{p_i} = \pi_0 + 2\pi_1 + 2\pi_2 + \dots + 2\pi_r. \tag{8}$$

Probability for the w-non-compatible J scale

If the k-scale is w-non-compatible, it does not include ω_0, and $X(\omega_0, p_i) \geq w (i = 1, \dots, 2r + 1)$. By definition, there is an admissible ordering p_i, say Q, for which $X(\omega_0, Q) = w$, and each step on the J scale from p_i to p_{i+1} (to the right of Q) or to p_{i-1} (to the left of Q) moves farther away from Q. This step moves away from ω_0 too, if p_i and ω_0 initially agree on the order of the options concerned, but *towards* ω_0 otherwise. Since $X(\omega_0, p_i) \geq w$,

Table 3.

The number of inversions of each admissible ordering p_j, indexed by $j = 1, \ldots, 2r+1$ (columns) with respect to the median ordering, for each possible position I_i (rows) of the median ordering on the J scale.

	Distances of admissible orderings from median ordering										
I_i	1	2	$\cdots$	$r-1$	r	$r+1$	$r+2$	$r+3$	$\cdots$	$2r$	$2r+1$
$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$r-1$	$r-2$	$r-3$	$\cdots$	0	1	2	3	4	$\cdots$	$r+1$	$r+2$
r	$r-1$	$r-2$	$\cdots$	1	0	1	2	3	$\cdots$	r	$r+1$
$r+1$	r	$r-1$	$\cdots$	2	1	0	1	2	$\cdots$	$r-1$	r
$r+2$	$r+1$	r	$\cdots$	3	2	1	0	1	$\cdots$	$r-2$	$r-1$
$r+3$	$r+2$	$r+1$	$\cdots$	4	3	2	1	0	$\cdots$	$r-3$	$r-2$
$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$2r+1$	$2r$	$2r-1$	$\cdots$	$r+2$	$r+1$	r	$r-1$	$r-2$	$\cdots$	1	0

Table 4.

The number of inversions of each admissible ordering p_j indexed by $j = 1, \ldots, 2r+1$ (columns) with respect to the median ordering, for optimal positions I_i (rows) of Q in a non-compatible J scale.

	Distances of admissible orderings from median ordering												
I_i	1	2	$\cdots$	$r-2$	$r-1$	r	$r+1$	$r+2$	$r+3$	$r+4$	$\cdots$	$2r$	$2r+1$
r	$r+1$	r	$\cdots$	3	2	1	2	3	2	3	$\cdots$	$r-2$	$r-1$
r	$r+1$	r	$\cdots$	3	2	1	2	3	4	3	$\cdots$	$r-2$	$r-1$
r	$r+1$	r	$\cdots$	3	2	1	2	3	4	5	$\cdots$	$r-2$	$r-1$
.	.	.	$\cdots$	.	.	.	.	.	.	.	$\cdots$	.	.
r	$r+1$	r	$\cdots$	3	2	1	2	3	4	5	$\cdots$	r	$r-1$
$r-1$	r	$r-1$	$\cdots$	2	1	2	3	2	3	4	$\cdots$	$r-1$	r
$r-1$	r	$r-1$	$\cdots$	2	1	2	3	4	3	4	$\cdots$	$r-1$	r
$r-1$	r	$r-1$	$\cdots$	2	1	2	3	4	5	4	$\cdots$	$r-1$	r
.	.	.	$\cdots$	.	.	.	.	.	.	.	$\cdots$	.	.
$r-1$	r	$r-1$	$\cdots$	2	1	2	3	4	5	6	$\cdots$	$r+1$	r

and only w out of $\binom{k}{2}$ steps are towards ω_0, it is evident that the total distance

$$X_{p_i} = \sum_i X(\omega_0, p_i), \quad (i = 1, \dots \binom{k}{2} + 1)$$

must be larger for a w-non-compatible J scale than for a non-compatible J scale; hence, P_{p_i} (from (8)) must be smaller than for a non-compatible J scale. Thus, a w-non-compatible J scale cannot be the dominant J scale, and we restrict our attention to non-compatible J scales. If the J scale is 1-non-compatible, there is an admissible ordering $p_i = Q$, for which $X(\omega_0, Q) = 1$, and $Q = \sigma_{s_t} \bullet \omega_0$, where options S and T are adjacent both in Q and ω_0; suppose the order is ST in ω_0 and TS in Q. Thus, S and T must be embedded in a triple RTS, and of the three options, T is the least preferred. Somewhere on the J scale, options S and T are interchanged (there is a midpoint between S and T), so there is also a $p_j (j \neq i)$ in which S and T are adjacent and their order agrees with that in ω_0. From (5) it follows that $X(Q, p_j) \geq 3$. With a proper placement of the triple RTS among remaining options on the J scale, Q may have a middle position on the J scale (see above). Hence, $X(\omega_0, Q) = 1$ in the middle of the J scale, and $X(\omega_0, p_i)$ monotonically increases towards one end of the J scale (say, the left end), and monotonically increases towards the other end apart from one drop by one inversion somewhere on the J scale (this follows from (6) and the discussion above). For the non-compatible J scales with Q in the midranking or next to it, the distances $X(\omega_0, p_i), i = 1, \dots, \binom{k}{2} + 1$ (apart from taking reverses), are given in Table 4. Using the same reasoning as in the case of median- compatible scales, there are now two candidates for the dominant J scale, namely those for which $X(\omega_0, p_i)$ shows the pattern:

$$r, r-1, r-2, \dots, 3, 2, 1, 2, 3, 2, 3, 4, 5, \dots, r-1, r, \text{ or}$$

$$r+1, r, r-1, \dots, 4, 3, 2, 1, 2, 3, 2, 3, 4, \dots, r-2, r-1,$$

and for these J scales,

$$P_{p_i} = \pi_1 + 3\pi_2 + 3\pi_3 + 2\pi_4 + \dots + 2\pi_{r-1} + 2\pi_r, \text{ or} \tag{9}$$

$$P_{p_i} = \pi_1 + 3\pi_2 + 3\pi_3 + 2\pi_4 + \dots + 2\pi_{r-1} + \pi_r + \pi_{r+1}. \tag{10}$$

Since $\pi_r > \pi_{r+1}$ (from the strong unimodality property of Feigin and Cohen's model), it follows that the J scale defined by (9) has the largest probability and must be the best non-compatible candidate for the dominant J scale. This scale will be compared with its best median-compatible counterpart (defined by its total probability from (8), to decide whether the dominant J scale is median-compatible or non-compatible. Suppose that the dominant J scale is non-compatible. Comparing (8) and (9), this would imply that

$$\pi_2 + \pi_3 > \pi_0 + \pi_1. \tag{11}$$

However, from the strong unimodality property of Feigin and Cohen's model, we know that rankings are stochastically ordered away from ω_0; thus,

$$\pi_0 > \pi_1 > \pi_2 > \pi_3 > \ldots > \pi_{r-1} > \pi_r, \tag{12}$$

hence, (11) is in contradiction with (12), therefore, the dominant J scale must be median-compatible. □

Corollary 1:

If the data follow the distribution specified by a strongly unimodal distance model for rankings, the dominant J scale is median compatible.

Proof:

Since we only used the strong unimodality property of Feigin and Cohen's model, the results of the theorem are valid for any strongly unimodal distance model for rankings. □

Corollary 2:

If the data are follow the distribution specified by a strongly unimodal distance model for rankings, the minimum-number-of-inversions J scale is median compatible.

Proof:

The dominant or ML J scale maximizes the probability of getting an admissible ordering, because probabilities of rankings decrease monotonically with increasing numbers of inversions away from the median ordering. Therefore, the ML J scale is at the same time the MNI J scale, and, hence, the results of the Theorem and Corollary 1 generalize to this case. □

This generalizes the found results for the 3-scale (see section 13.5). The plot for

$$G = (\theta^2 + \theta^3)/(1 + \theta)$$

i.e., the probability-ratio for the best non-compatible J scale versus the best median-compatible J scale, is given in Figure 5, for $k \geq 3$ and $0 \leq \theta \leq 1$. From this, it is evident that this ratio is monotonically increasing but smaller than 0.5 over the whole range of θ. Therefore, the best median-compatible J scale optimizes the probabilities of admissible orderings of a J scale as well as the total number of inversions of the individual rankings with respect to the J scale.

Folded J Scales and Consensus Orderings

The main purpose of this paper concerned the generalizability of Coombs' (1954) and Goodman's (1954) results. These authors showed that Simple Majority Rule ordering is the ranking of the median individual on the quantitative J scale if *all* individuals' preference functions are single-peaked on the J scale. We used theorems from Kemeny (1959), Kemeny and Snell (1972), and Bogart (1973, 1975) to show that Simple Majority Rule yields the median ordering when not all rankings are *SPF*'s. We proved that

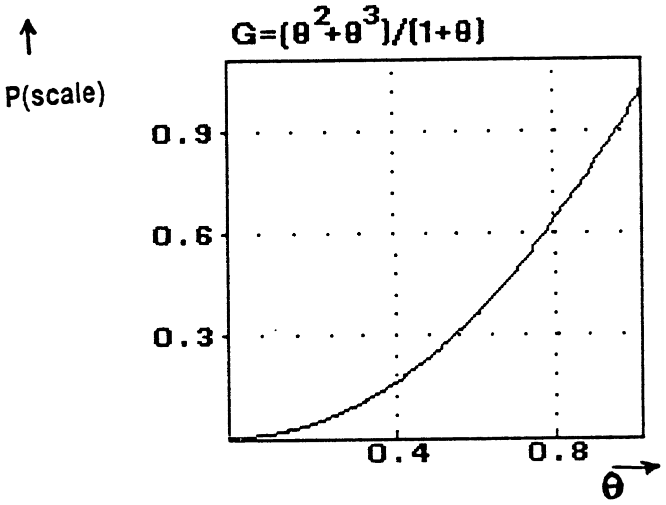


Figure 5. *Probability ratio P(best non-median scale)/ P (best median-compatible scale), for* $k \geq 4; 0 \leq \theta \leq 1$.

the median ordering is a folded J scale if probabilities of rankings strictly decrease with increasing Kemeny or Kendall distance from the median or modal ordering. This constitutes the generalizability of Coombs' and Goodman's assertions under rather general conditions. For all sets of data which have been analyzed to date, the median ordering proved to be a folded J scale. This result is the more remarkable since these sets often contained a high level of error, and the number of individuals was often small. It appears to hold quite generally for dichotomous data as well (see Van Blokland, 1991). In one case, the Andrich (1988) data on Capital Punishment, two pairs of options are tied in the median ordering and the median ordering is not a folded J scale in this case. This is precisely what is to be expected according to Bogart's theorems (see section 13.2).

Folded J Scales and Single-Peaked Social Preference Functions

Just as the normal distribution is found to mirror the frequency distribution of many variables, a unimodal or *single-peaked* distribution of rankings often arises on the unfolding scale. The single-peakedness of the social preference curve has important consequences for interpretation: the social preference decreases on either side of the median ordering towards the ends of the J scale. Socially most preferred options are found in the center of the J scale, less popular options towards both ends. Mostly, the ends of the J scale have opposite connotations: the J scale may be described in terms of a *bipolar* continuum. In folding the J scale in the *social ideal point* (point of highest social preference) in the center of the scale, the median ordering arises as a folded J scale, with options ranked in order of decreasing social preference. This is the Simple Majority Rule ordering, the *group consensus ordering.* This is what we found in all of the data sets analyzed up till now, except for the Andrich data (see above and next section).

13.7 Unfolding and Social Choice Theory: Illustrations

Two sets of data, one consisting of complete rankings and one consisting of dichotomous data, are presented for illustration of the connections between unfolding theory and social choice theory: the Coombs' (1950) Grade Expectations and Andrich' (1988) data on Capital Punishment. They represent two different kinds of data: complete and incomplete rankings of preference, respectively.

13.7.1 *Analysis of Complete Rankings: Coombs' (1950) Grade Expectations*

The first example concerns Coombs' (1950) Grade Expectations data ($n = 121$). Subjects were students in a graduate course in statistics ($n_1 = 40$) and in an undergraduate course in sociology ($n_2 = 81$), who completed a questionnaire about grades expected in the course (from 'most expected' to 'least expected'). The grades were: A, B, C, D, and E. The rankings, their frequencies, and the number of students in each of the courses are given in Table 5. For this data, both the dominant (ML) and the minimum-number-of-inversions (MNI) criterion yield the same results, namely, the qualitative J order $ABCDE$ and the quantitative J scale $ABCDE$. In addition, these results are obtained both in the total group and in the two subgroups separately. All other results (the median ordering is a folded J scale, and single-peakedness of the social preference function) are the same as well. For space considerations, only the results for the total group are presented here.

Best ML and MNI quantitative J scales

The order of the midpoints be and cd on the J scale is indeterminate because of zero frequencies for two of the admissible orderings. Hence, two different quantitative J scales are possible, both of which include the same number of individuals without inversions. Also, the total number of inversions from individuals' rankings with respect to the quantitative J scale is the same for both scales. In Figure 6, one of these J scales is displayed on a line. If an individual ranking is observed that is not equal to one of the admissible orderings of the quantitative J scale, it is assigned the admissible ordering from which it has a minimum number of inversions. The first number above the admissible orderings is the frequency of rankings that fit the indicated admissible ordering without inversions. The second number, if any, is the number of non-permissible individual rankings that

Table 5.

Coombs' data: number of individuals in two courses with the following rankings

	Statistics	Sociology	Total
Ranking	Class	Class	group
1. ABCDE	14	6	20
2. BACDE	10	22	32
3. BCADE	6	21	27
4. CBADE	1	4	5
5. CBDAE	1	11	12
6. CBDEA	2	3	5
7. CDBEA	1	0	1
8. DECBA	1	0	1
9. BCDAE	3	7	10
10. BCDEA	0	6	6
11. BACED	0	1	1
12. CABDE	1	0	1
	40	81	121

have been assigned the indicated admissible ordering with one inversion, and so on (see Van Blokland, 1991).

Our results can be compared with Coombs' results. Coombs sought the *dominant* J scale, that is, the scale that fits a maximum number of individuals' rankings perfectly (Coombs, 1964; Niemi, 1969), thus, the ML scale. Our criterion for the best J scale is the minimization of the total number of inversions from the individual rankings (see section 13.5), the MNI criterion. Despite this difference in optimization criteria, for this data, Coombs' and our results are essentially the same: (a) the best qualitative J order is the same (i.e., $ABCDE$) under the ML and MNI criterion; (b) the best quantitative J scale is the same (i.e., $ABCDE$) under the ML and MNI criterion; (c) the median ordering is a folded J scale; and (d) the social preference function is single-peaked on the quantitative J scale.

Goodness-of-Fit

The goodness-of-fit of the Feigin and Cohen model to the data is evaluated using the chi-squared testing procedure described in Feigin and Cohen (1978). As a test statistic, Pearson's X^2 is used. The number of degrees of freedom for this test is equal to the number of X-categories in which the data are combined, minus 2. One extra degree of freedom has to be subtracted because of the estimation of θ; for the estimation of ω_0 no degree of freedom need be subtracted (see Critchlow, 1985). The test for goodness-of-fit is presented in Table 6. In this Table, the observed frequencies of inversions, X are given. From this, $\bar{x} = \sum x/n = 153/121 = 1.26$, and $\hat{\theta}$

can then be determined: $\hat{\theta} = 2.75$ (from the Appendix, and by interpolating). Expected frequencies are determined from the Feigin and Cohen distribution from $\exp(x) = n \cdot P_{\hat{\theta}}(x)$. Expected values for $X = 6, 7, \ldots 10$ have been combined because of small values, leaving 6-2 = 4 degrees of freedom. On behalf of the resulting p-value, we conclude that the Feigin and Cohen model does not fit this data. This may be caused by a lack of symmetry: rankings with the same number of inversions from the median ordering do not have the same frequencies; this is reflected in the J scale: rankings are not symmetrically distributed about the median ordering.

Figure 6 shows that the median ordering is a folded J scale, and, is thus an admissible ordering of the J scale; moreover, it shows the ranking of the median individual on the J scale. Also, the social preference function is unimodal on the J scale. Despite the fact that the rankings do not fit Feigin and Cohen's unimodal distance model, Coombs' and Goodman's theorems apply for this data that do not only consist of SPF's (see section 13.2). Therefore, we conclude that a strictly unimodal distance model for rankings is a *sufficient* condition and not a necessary one to ensure that the best MNI scale includes the median ordering as an admissible ordering, and that the social preference function is unimodal on the quantitative J scale.

We have to distinguish between the unimodal distance model and the unfolding model. With Feigin and Cohen's model, we are fitting a strictly unimodal distance model, with the unfolding model, the total number of inversions of individuals' rankings with respect to all admissible orderings of the quantitative J scale (denoted $\sum Y$) is assessed (see section 13.1). The goodness of fit of the unfolding model to the data is tested in an analogous way as in Feigin and Cohen's model. The Coombs' data fit the unfolding model nearly perfectly (see Van Blokland, 1991).

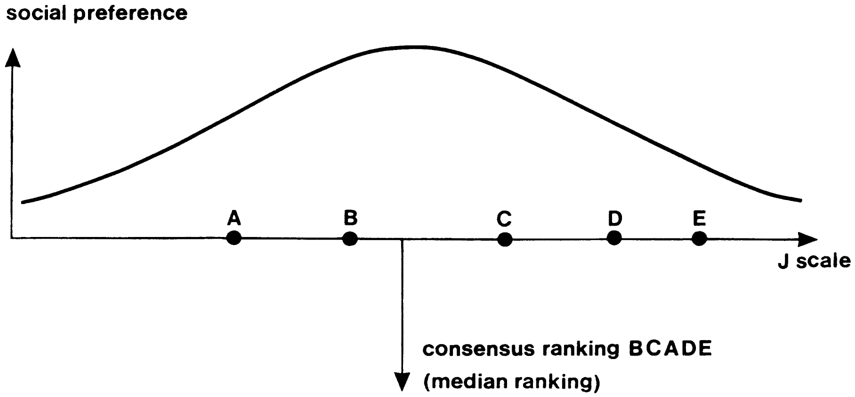


Figure 6 *One of two possible quantitative J scales ABCDE for Coombs' (1950) Grade Expectations.*

Table 6.

Number of inversions X, their observed frequencies (Obs) and expected frequencies (Exp) for the Coombs' (1950) Grade Expectations

# Inversions X	Obs	Exp	χ^2
0	27	37.21	2.80
1	47	40.92	0.90
2	40	25.33	8.50
3	5	11.60	3.76
4	1	4.26	2.49
5	0	1.28	0.23
6	0	0.33	
7	1	0.06	
8	0	0.00	
9	0	0.00	
10	0	0.00	
Total	121	121	χ^2=17.11 $df = 4, p \approx .001$

13.7.2 *Analysis of Incomplete Rankings: Andrich' Data*

The procedures for finding the best qualitative or quantitative J scale on the basis of the MNI criterion can be used for incomplete rankings as well. Incomplete rankings may consist of partial rankings, 'order r out of k' data ($r \leq k-1$), rating scores, and dichotomous data. In the unfolding procedure,[2] equal scores are treated as ties, which are untied using a primary approach to ties (Van Blokland, 1991). Andrich (1988) collected data concerning attitudes toward Capital Punishment. Subjects were 54 graduate students in an introductory course in Educational Measurement. The data are dichotomous scores: Agree (1) or Disagree (0). The students were asked to judge eight statements:

A Capital punishment is one of the most hideous practices of our time 'HIDEOU'

B The state cannot teach the sacredness of human life by destroying it 'SACRED'

C Capital punishment is not an effective deterrent to crime 'NONEFF'

D I do not believe in capital punishment but I am not sure it is not necessary 'NOSURE'

[2] For the unfolding analysis of the data, the computer program UNFOLD (Van Blokland and Van Blokland, 1990) has been used.

E I think capital punishment is necessary but I wish it were not 'INEVIT'

F Until we find a more civilised way to prevent crime we must have capital punishment 'NECESS'

G Capital punishment is justified because it does act as a deterrent to crime 'JUSTIF'

H Capital punishment gives the criminal what he deserves 'REVENG'

Median Ordering and Social Preference Function

The results of the unfolding analysis for Andrich' data are given in Table 7. The best quantitative J scales for $k = 4, \ldots 8$ form an order-preserving or 'nested set' of quantitative J scales: each smaller J scale is contained in a larger J scale, as can be seen from Figure 7. The resulting scale is $ABCDEFHG$, and is the same as the one Andrich obtained. The underlying continuum goes from Capital Punishment as a *Hideous Practice* to *Justified Deterrent to Crime*, and clearly is bipolar.

The median ordering for Andrich' data is $CBED(AF)(GH)$, where () means that options have equal mean ranks, and inverting these options in the median ordering yields the same total number of inversions from individuals' rankings. Thus, the median ordering is not unique. The same conclusion follows from the observation that the number of individuals preferring A to F equals the number preferring F to A. Also, the number of individuals preferring G to H equals the number preferring H to G. This illustrates Bogart's (1975) theorem that the median is unique unless for some options X and Y, the number of individuals preferring X to Y is equal to the number preferring Y to X (see section 13.2).

Apart from the ties in the median ordering, the median ordering also is not a folded J scale for this data: the order of D and E on the J scale is not in agreement with their position in the median ordering. The number of individuals preferring E to D is 14, the number preferring D to E is 13, which, again, is a demonstration of Bogart's theorem: E just has a simple (14/27) majority over D, which is reflected in the median ordering where E dominates D. The fact that option D precedes E on the J scale is probably due to the large number of individuals (26) who prefer C and B (the most preferred options) to E, while the number preferring C and B to D is only 17 and 18, respectively. In fact, the mean scores of D and E are nearly equal: 28/54 and 29/54 respectively. On inverting these options in the median ordering, the total number of inversions with respect to the J scale, $\sum Y$, increases from 299 to 300. Hence, the social preference function is very close to single-peakedness.

Options:	(A)	(B)	(C)	(D)	(E)	(F)	(H)	(G)
Codes:	Hideou	Sacred	Noneff	Nosure	Inevit	Necess	Reveng	Justif
$k = 4$	Hideou				Inevit	Necess	Reveng	
$k = 5$	Hideou			Nosure	Inevit	Necess	Reveng	
$k = 6$	Hideou		Noneff	Nosure	Inevit	Necess	Reveng	
$k = 7$	Hideou		Noneff	Nosure	Inevit	Necess	Reveng	Justif
$k = 8$	Hideou	Sacred	Noneff	Nosure	Inevit	Necess	Reveng	Justif
		← Hideous Practice			Justified Deterrent →			

Figure 7 *Order preserving subsets of quantitative J scales for Andrich' data*

Table 7.

Best quantitative J scales for Andrich' data, $k = 4, \ldots, 8$. The total number of inversions for this J scale is given by '$\sum y$', the frequency of perfect fit is given under 'Fit'.

Andrich' Data			
k	Scale	$\sum y$	Fit
4	AEFH	1	53
5	ADEFH	5	50
6	ACDEFH	12	47
7	ACDEFHG	22	43
8	ABCDEFHG	37	40

13.8 Discussion

The unfolding technique was placed in the wider context of social choice theory, median procedures and strictly unimodal distance models for rankings. Social choice theory was used to construct a framework for the unidimensional unfolding model. We generalized Coombs' (1954) and Goodman's (1954) theorems: if the data follow a strictly unimodal distance model, the modal or median ranking is an admissible ordering of the quantitative J scale whose admissible orderings have the highest total probability according to the model. This is because the maximum likelihood and the minimum-number-of-inversions criterion yield the same ordering: the mean/modal/median ordering. From this, it follows that the group consensus ordering is transitive and is the median ordering. In addition, we proved that the social preference function is unimodal on the quantitative J scale if the data follow a strictly unimodal distance model.

The unfolding procedure was illustrated in Coombs' (1950) Grade Expectations. Results have been compared with Coombs'. Coombs sought the dominant or ML J scale, our criterion for the best J scale is the minimization of the number of inversions, the MNI criterion. Despite this difference, Coombs' and our results are essentially the same: the best quantitative J scale is the same. The median ordering is a folded J scale, and the social preference function is single-peaked on the J scale.

We presented the Andrich' data to show the importance of Bogart's (1973, 1975) theorems on the unicity of the median ranking. In the Andrich' data, the median ordering it is not unique: three pairs of options are tied in the median ordering. Combined with an option pair that shows only a bare majority, this results in a median ordering that is not a folded J scale. Consequently, the social preference function is not single-peaked, however, it is nearly so.

APPENDIX: $E_\theta(X)$ and $E_\theta(\tau)$

$E_\theta(X)$ (first row) and $E_\theta(\tau)$ for selected values of θ and k.

θ^k	3	4	5	6	7	8	9	10
.05	.100	.152	.205	.258	.310	.363	.415	.468
	.933	.949	.959	.966	.970	.974	.977	.979
.10	.199	.310	.421	.532	.643	.754	.865	.976
	.867	.897	.916	.929	.939	.946	.952	.957
.15	.297	.471	.647	.824	1.000	1.177	1.353	1.530
	.802	.843	.871	.890	.905	.916	.925	.932
.20	.392	.636	.884	1.134	1.384	1.634	1.884	2.134
	.738	.788	.823	.849	.868	.883	.895	.905
.25	.486	.803	1.132	1.464	1.797	2.130	2.463	2.796
	.676	.732	.774	.805	.829	.848	.863	.876
.30	.576	.972	1.388	1.813	2.240	2.668	3.096	3.525
	.616	.676	.722	.758	.787	.809	.828	.843
.35	.663	1.141	1.653	2.180	2.714	3.251	3.789	4.327
	.558	.620	.669	.709	.741	.768	.790	.808
.40	.747	1.309	1.924	2.566	3.221	3.882	4.547	5.212
	.502	.564	.615	.658	.693	.723	.747	.768
.45	.828	1.475	2.199	2.967	3.759	4.564	5.375	6.190
	.448	.508	.560	.604	.642	.674	.701	.725
.50	.905	1.638	2.477	3.382	4.326	5.295	6.277	7.268
	.397	.454	.505	.549	.588	.622	.651	.677
.55	.978	1.798	2.755	3.806	4.920	6.075	7.256	8.452
	.348	.401	.449	.492	.531	.566	.597	.624
.60	1.048	1.953	3.031	4.238	5.536	6.899	8.308	9.747
	.301	.349	.394	.435	.473	.507	.538	.567
.65	1.115	2.103	3.304	4.672	6.168	7.762	9.429	11.149
	.256	.299	.339	.377	.413	.446	.476	.504
.70	1.179	2.248	3.572	5.105	6.810	8.654	10.609	12.651
	.214	.251	.286	.319	.351	.382	.411	.438
.75	1.239	2.388	3.832	5.533	7.455	9.565	11.834	14.237
	.174	.204	2.34	.262	.290	.317	.343	.367
.80	1.297	2.522	4.085	5.953	8.096	10.483	13.088	15.885
	.135	.159	.183	.206	.229	.251	.273	.294
.85	1.352	2.650	4.329	6.362	8.726	11.396	14.350	17.565
	.099	.117	.134	.152	.169	.186	.203	.219
.90	1.404	2.772	4.563	6.757	9.340	12.293	15.601	19.248
	.064	.076	.087	.099	.110	.122	.133	.145
.95	1.453	2.889	4.786	7.137	9.932	13.164	16.823	20.902
	.031	.037	.043	.048	.054	.060	.065	.071
1.00	1.500	3.000	5.000	7.500	10.500	14.000	18.000	22.500
	0	0	0	0	0	0	0	0

13.9 References

[1] Andrich, D. The application of an unfolding model of the PIRT type to the measurement of attitude. *Appl. Psych. Meas.,* **12**:33-51, 1988.

[2] Andrich, D. A probabilistic IRT model for unfolding preference data. *Appl. Psych. Meas.,* **13**:193-216, 1989.

[3] Arrow, K.J. *Social Choice and Individual Values.* New York: Wiley. 1951.

[4] Black, D. On the rationale of group decision-making. *J. Pol. Econ.,* **56**:23-34, 1948a.

[5] Black, D. The decisions of a committee using a special majority. *Econometrica,* **16**:245-261, 1948b.

[6] Bogart, K.P. Preference structures I: Distances between transitive preference relations. *J. Math. Sociol.,* **3**:49-67, 1973.

[7] Bogart, K.P. Preference Structures II: Distances between asymmetric relations. *SIAM J. Appl. Math.,* **29**:254-262, 1975.

[8] Borda, J.-C. Mémoire sur les élections au scrutin. Read to the French Académy of Sciences in 1770, printed in: Histoire de l'Académy Royal des Sciences, 1781, published in 1784.

[9] Bradley, R.A. and Terry, M.E. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika,* **39**:324-345, 1952.

[10] Carroll, J.D. Individual differences and multidimensional scaling. In R.N. Shepard, A.K. Romney, and S.B. Nerlove (Eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Vol. I:* Theory. New York: Seminar Press. 1972.

[11] Coombs, C. H. Psychological scaling without a unit of measurement, *Psych. Rev.,* **57**:145-158, 1950.

[12] Coombs, C.H. Social choice and strength of preference. In R.M. Thrall, C. H. Coombs, & R.L. Davis, *Decision Processes.* Wiley: New York, 69-86, 1954.

[13] Coombs, C.H. *A Theory of Data.* New York: Wiley; republished in 1976 by Ann Arbor, MI: Mathesis Press. 1964.

[14] Critchlow, D.E. *Metric Methods for Analyzing Partially Ranked Data.* Berlin: Springer Verlag. 1985.

[15] D. E. Critchlow, M. A. Fligner, and J. S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, **35**:294-318, 1991.

[16] Davison, M.L. Testing a unidimensional, qualitative unfolding model for attitudinal or developmental data. *Psychometrika*, **44**:179-194, 1979.

[17] De Condorcet, M. Essai sur l'application de l'analyse à la probabilit des dácisions rendues à la pluralité des voix. Paris: L'Imprimerie Royale. 1785.

[18] DeSarbo, W.S. and Hoffman, D.L. Simple and weighted unfolding threshold models for the spatial representation of binary choice data. *Appl. Psych. Meas.,* **10**:247-264, 1986.

[19] Dijkstra, L. Ontvouwing: Over het Afbeelden van Rangordes van Voorkeur in Ruimtelijke Modellen. Assen (The Netherlands): Van Gorcum & Co. 1978.

[20] Feigin, P.D. and Cohen, A. On a model for concordance between judges. *J.R. Statist. Soc., B,* **40**: 203-213, 1978.

[21] Fishburn, P.C. *The Theory of Social Choice.* Princeton: Princeton University Press. 1972.

[22] Fligner, M.A. and Verducci, J.S. Distance based ranking models. *J.R. Stat. Soc. B,* **48**:359-369, 1986.

[23] Fligner, M.A. and Verducci, J.S. Multistage ranking models. *J. Am. Stat. Assoc.,* **83**:892-901, 1988.

[24] Formann, A.K. Latent class models for nonmonotone dichotomous items. *Psychometrika,* **53**:45-62, 1988.

[25] Goodman, L.A. On methods of amalgation. In R. M. Thrall, C.H. Coombs, & R.L. Davis (Eds.), *Decision Processes* . New York: Wiley, 39-48, 1954.

[26] Heiser, W.J. Unfolding analysis of proximity data. Ph.D. Thesis. University of Leiden, The Netherlands. 1981.

[27] Heiser, W.J. Joint ordination of species and sites: the unfolding technique. In P. Legendre and L. Legendre (Eds.), Developments in Numerical Ecology, Berlin: Springer-Verlag. 1987.

[28] Henery, R.J. Permutation probabilities as models for horse races. *J. R. Stat. Soc. B,* **43**:86-91, 1981.

[29] Jansen, P.G.W. Rasch analysis of attitudinal data. Ph.D. Thesis. University of Nijmegen, The Netherlands. 1983.

[30] Kemeny, J.G. Mathematics without numbers. *Daedalus,* **88**: 577-591, 1959.

[31] Kemeny, J.G. and Snell, *J.L. Preference rankings, an axiomatic approach. Mathematical Models in the Social Sciences* New York: Blaisdell, 1962; reprinted by MIT-Press, Cambridge, Mass., 9-23, 1972.

[32] Kendall, M.G. *Rank correlation methods.* London: Griffin. 1970, 4th ed.

[33] Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, **29**:115-129, 1964.

[34] Luce, R.D. and Raiffa, H. *Games and Decisions.* New York: Wiley. 1957.

[35] Malkevitch, J. and Meyer, W. *Graphs, Models, and Finite Mathematics.* Englewood Cliffs NJ: Prentice-Hall. 1974.

[36] Mallows, C.L. Non null ranking models I. *Biometrika*, **44**:114-130, 1957.

[37] May, K.O. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, **20**:680-684, 1952.

[38] Niemi, R.G. (1969). Majority decision-making with partial unidimensionality. *Am. Pol. Sc. Rev.,* **63**: 488-497, 1969.

[39] Riker, W.H. and Ordeshook, P.C. *Positive Political Theory.* Englewood Cliffs NJ: Prentice-Hall. 1973.

[40] Roberts, F.S. *Discrete Mathematical Models.* Englewood Cliffs NJ: Prentice-Hall. 1976.

[41] Roskam, E.E. Ch. I. *Metric Analysis of Ordinal Data.* Voorschoten: VAM. 1968.

[42] Sen, A. *Choice, Welfare and Measurement.* Oxford: Basil Blackwell. 1982.

[43] Schulman, R.S. Ordinal data: an alternative distribution. *Psychometrika*, **44**:3-20, 1979.

[44] Sixtl, F. Probabilistic unfolding. *Psychometrika*, **38**:235-248, 1973.

[45] Thurstone, L.L. A law of comparative judgment. *Psych. Rev.,* **34**:278-286, 1927.

[46] Torgerson, W.S. *Theory and Methods of Scaling.* New York: Wiley. 1958.

[47] Tversky, A. Intransitivity of Preferences. *Psych. Rev.,* **76**:31-48, 1969.

[48] Van Blokland-Vogelesang, A.W. (1989). Unfolding and consensus ranking: A prestige ladder for technical occupations. In G. De Soete, H. Feger, & K.C. Klauer (Eds.), *New Developments in Psychological Choice Modeling*, 237-258. Amsterdam: North-Holland.

[49] Van Blokland-Vogelesang, A. W., and Van Blokland, P.J. (1990). UNFOLD: Unidimensional Unfolding of Preference Data. User's manual. Groningen, N.L. : I.E.C. ProGAMMA.

[50] Van Blokland-Vogelesang, A.W. (1991). Unfolding and Group Consensus Ranking for Individual Preferences. Thesis. Leiden, N.L.: DSWO-Press.

[51] Van Schuur, H. *Structure in Political Beliefs.* Thesis, Amsterdam: CT-Press. 1984.

[52] Vickrey, W. Utility, strategy, and social decision rules. *Quart. J. Econ.,* **74**:507-535, 1960.

Miscellanea

Models on Spheres and Models for Permutations

Peter McCullagh [1]

ABSTRACT It is shown that the space of permutations is naturally ordered in a circular or spherical manner. By exploiting the geometry of the sample space it is shown that Mallows's ϕ-model with the Spearman metric is essentially equivalent to the Mallows-Bradley-Terry ranking model, which is essentially equivalent to the von Mises-Fisher model on the sphere. Extensions to bi-polar models are discussed briefly.

The Geometry of Permutations

In the discussion that follows, S_k^* is the set of permutations of k items, here labelled $a, b, \ldots, k$. Thus S_k^* has $k!$ elements, and if $k = 4$ a typical vector y in S_k^* is $y = (b, d, a, c)$. The *elements* of S_k^* are k-component vectors, but the *components* of $y \in S_k^*$ are non-numeric labels. Associated with each $y \in S_k^*$ is a vector of ranks $r(y)$. Thus if $y = (b, d, a, c)$, $r(y) = (3, 1, 4, 2)$. The components of the rank vector are ordinal numbers, namely the rank achieved by that item or competitor. By convention, the integers $1, \ldots, k$ are used to denote the ranks, and these are often treated as cardinal numbers rather than ordinal numbers. Use of the integers as cardinal numbers in this context (rather than squared integers or reciprocal integers) is clearly arbitrary, though in practice this choice is often surprisingly effective.

In visualizing the set S_k^* I find it helpful to imagine the $k!$ points lying on the surface of a sphere in Euclidean space of dimension $k - 1$. Neighboring points in this set differ by one transposition of adjacent letters or labels. The set S_3^* and the rank vectors for $k = 3$ are shown in Fig. 1. Figure 2 illustrates the nature of the set S_4^*.

[1] Department of Statistics, University of Chicago

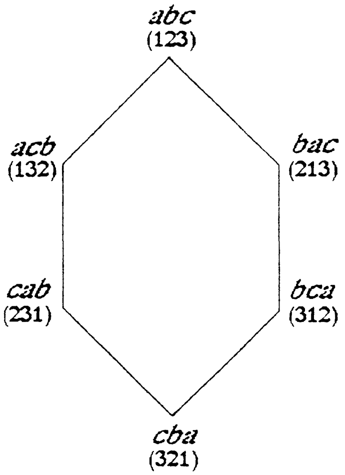


Fig 1. Schematic diagram of the permutation vectors and rank vectors for $k = 3$.

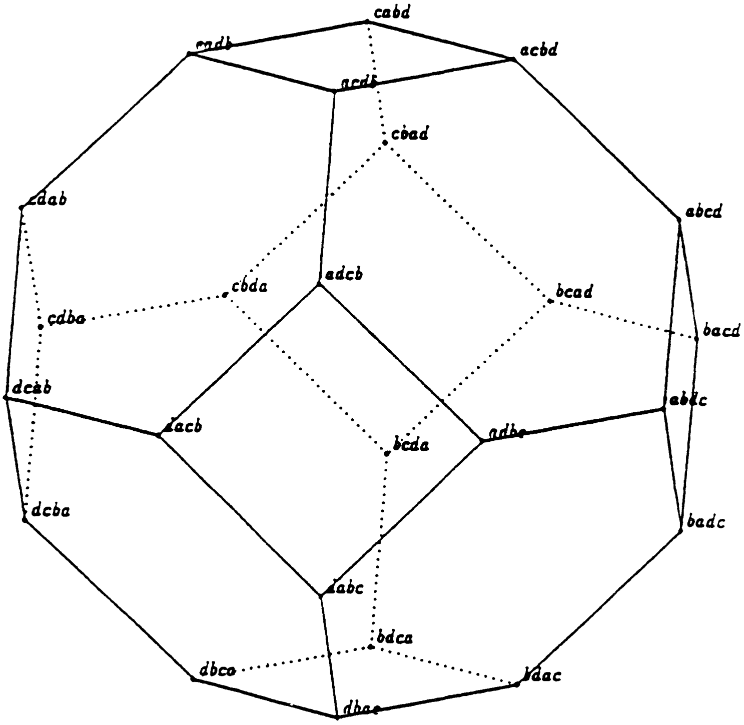


Fig 2. Sample space of permutations of *abcd*. The graph has 24 vertices, 36 edges, 6 square faces and 8 hexagonal faces.

The above representation of S_k^* as a network of points on the surface of a sphere is based on Kendall's metric. The distance between any two points y_1, y_2 in Kendall's metric is the shortest Euclidean distance within the network between the points: in other words the length of the shortest chain beginning at y_1 and ending at y_2. By contrast, Spearman's metric given by

$$d_S(y_1, y_2) = \| r(y_1) - r(y_2) \|,$$

the Euclidean distance between the two rank vectors, measures distance 'as the crow flies', i.e. not restricted to the network of edges of the graph. Kendall's metric treats the ranks as ordinal numbers and is unaffected by monotone transformation of the ranks. Spearman's metric on the other hand treats the rank vectors as points in a $(k-1)$-dimensional affine subspace of Euclidean k-space. Both metrics are entirely consistent with the diagrams shown above, but with slightly different interpretations of distance.

MALLOWS'S ϕ MODELS.

In his 1957 paper, Mallows discussed a class of unimodal models on the sample space of permutations. The simplest of these, discussed in section 7 of Mallows's paper, have the form

$$p(y) = C_\phi \exp\{-\phi d(y, \mu)\} \tag{1}$$

in which $d(y, \mu)$ is a measure of distance (or squared distance or 'generalized matching coefficient') between y and the modal permutation μ, ϕ is a concentration parameter and C_ϕ is a constant of integration chosen to make the probabilities sum to unity. For purposes of this discussion we take $d(\cdot, \cdot)$ to be

$$d(y, \mu) = d_S^2(y, \mu) = \| r(y) - \theta \|^2, \tag{2}$$

the squared Euclidean distance between $r(y)$ and θ, the rank vector for the modal permutation μ.

It is conventional to denote the ranks by the integers $1, \ldots, k$ so that the sum of the ranks is $\frac{1}{2} k(k+1)$ and the sum of squared ranks is $\frac{1}{6} k(k+1)(2k+1)$. This choice is to some extent arbitrary, and in what follows it is more convenient to arrange matters so that the average rank is 0 and the sum of squared ranks is unity. In other words we use numerical ranks of the form

$$i' = \big(i - (k+1)/2\big) / \sqrt{k(k^2-1)/12}$$

in which i is an integer from 1 to k. Thus these modified rank vectors lie on the unit sphere. With that choice in (2) we have

$$d(y, \mu) = 2 - 2\theta \cdot r(y)$$

where $\theta \cdot r(y)$ is the cosine of the angle between $r(y)$ and the modal rank vector. Thus Mallows's probability distribution becomes

$$p(y) = C_\phi \exp\{2\phi\theta \cdot r(y)\}, \tag{3}$$

in which θ and $r(y)$ are both rank vectors of unit length.

In the literature on directional data, probability distributions on the surface of the unit sphere S_p in p-space are often studied. Beginning with the uniform distribution on S_p, the associated exponential family

$$\frac{\exp(\kappa\theta^T y)\kappa^\nu}{2^{\nu+1} I_\nu(\kappa)\, \pi^{p/2}}, \tag{4}$$

with $|\theta| = |y| = 1$, $\kappa \geq 0$, $\nu = p/2 - 1$, is known as the von-Mises Fisher distribution. This is an exponential family distribution with canonical parameter $\kappa\theta$: θ is the mean or modal direction and κ is a concentration parameter.

There is a very strong similarity between (4) and (3) with $\kappa = 2\phi$, $p = k - 1$, and in fact it is advantageous to extend (3) by taking θ to be an arbitrary vector with zero sum and unit length, not necessarily one of the $k!$ rank vectors. With this extension we have

$$p(y) = C(\phi, \theta) \exp\{2\phi\theta \cdot r(y)\} \tag{5}$$

where the constant of integration now depends on both θ and ϕ. In interpreting θ as a modal ranking vector, it must now be borne in mind that θ may correspond to a vector on the unit sphere that is intermediate between two or more permutations. In other words, there may be two or more permutations with rank vectors close to θ, whose probabilities are approximately equal under (5). For example, $\theta = (-3, 1, 0, 2)/\sqrt{14}$ corresponds to a 'rank vector' $(1, 3, 2.5, 3.5)$. This is half-way between $(1, 2, 3, 4)$ and $(1, 4, 2, 3)$, though the nearest rank vector is clearly $(1, 3, 2, 4)$ corresponding to a modal permutation *acbd*. By extension $\theta = (-3, 1, 1, 1)/\sqrt{12}$ gives rise to a distribution that depends only on the rank assigned to a. The six modal permutations are those for which a is placed in first position.

Evidently, by reparameterization, (5) is a full exponential family model in which the sum of the rank vectors is the complete sufficient statistic. It follows that (5) must be a re-parameterization of the Mallows-Bradley-Terry ranking model, for which the average rank vector is also the sufficient statistic. Thus we are led to the approximate equivalence

$$M\text{-}S^2 = M\text{-}B\text{-}T = vM\text{-}F$$

meaning that Mallows's model with the squared Spearman metric is equivalent to the Mallows-Bradley-Terry ranking model, which is in turn equivalent to the von Mises-Fisher model. Strictly speaking, the three models are distinct: $M\text{-}S^2$ is equivalent to $M\text{-}B\text{-}T$ with the restriction that the modal vector in the *M-B-T* model be proportional to one of the $k!$ rank vectors. The Mallows-Bradley-Terry ranking model is equivalent to the von Mises-Fisher distribution except for the discrete nature of the sample space: the likelihood functions are both of the linear exponential-family form.

Using the fact that the set of $k!$ permutations lie on a sphere in $(k-1)$-space, we can use the normalization constant from the von Mises Fisher distribution to obtain an approximate normalization constant for the Mallows-Bradley-Terry ranking model (5). On approximating the sum in (5) by an integral over the sphere we find that the approximate normalization constant depends only on ϕ and not on the modal direction θ. The approximation thus obtained is

$$C(\phi,\theta) \simeq \frac{\phi^{\nu}}{I_{\nu}(2\phi)\, k!\, \Gamma(\frac{k-1}{2})},$$

where $\nu = (k-3)/2$, and $I_{\nu}(\phi)$ is the modified Bessel function of order ν. This approximation seems to be very accurate particularly for small values of ϕ, say $\phi < 1$. For example, if $k = 4$ and θ corresponds to any rank vector, the exact sum reduces to

$$2\cosh(2\phi) + 6\cosh(8\phi/5) + 2\cosh(6\phi/5) + 8\cosh(4\phi/5) + 4\cosh(2\phi/5) + 2.$$

The approximation gives

$$C^{-1}(\phi,\theta) = \frac{4!\ \sinh(2\phi)}{2\phi}.$$

In the range $\phi < 1$, these expressions differ by no more than 0.3%. For other values of θ the error of approximation can be larger, but it never exceeds 1% for $\phi < 1$. The maximum error seems to decrease rapidly as k increases.

Bi-polar models

The logarithm of the von Mises-Fisher distribution (4) is linear in y. Such linear functions on the sphere are called first-order harmonics. They form a space of dimension p that is invariant under orthogonal transformation of coordinates. The second-order harmonics on S_p are quadratic functions of the form $\sum a_{ij} y_i y_j$ with coefficient matrix A satisfying $a_{ij} = a_{ji}$ and $\sum a_{ii} = 0$. These functions form an invariant subspace of dimension $p(p+1)/2 - 1$. The second-order exponential-family model on the sphere is therefore

$$p(y;\theta) = \exp\Bigl\{\sum \theta_i y_i + \tfrac{1}{2}\sum \theta_{ij} y_i y_j - K(\theta)\Bigr\}$$

with $\sum \theta_{ii} = 0$. Unfortunately there is no simple closed form expression for the cumulant function $K(\theta)$. Depending on the choice of parameters, the density may be unimodal or bimodal: in addition, there may be stationary points that are neither local maxima nor minima. If $\theta_i = 0$ the density is antipodally symmetric, and is known as the Bingham distribution. The shape of the density is then governed by the eigenvalues of the matrix θ_{ij}.

Analogous probability distributions on the set of permutations are obtained by replacing y by the rank vector, and summing over the permutations. Such models are automatically invariant under re-labelling of items or candidates.

References

[1] Mallows, C.L. Non-null Ranking Models. I. *Biometrika* **44**:114-130, 1957.

Complete Consensus and Order Independence: Relating Ranking and Choice

Hans Colonius[1]

ABSTRACT Complete Consensus has been introduced as a plausible independence from irrelevant alternatives' property of probability models on rankings. It is shown here that complete consensus implies order-independence for the choice probabilities of the corresponding random utility models for choice.

Introduction

In a *simple choice experiment*, a person is asked to select one element from a set of available alternatives according to some specified criterion (preference, loudness, brightness, etc.). Similarly, in a *ranking experiment* the task consists of rank-ordering the set (or a subset) of available alternatives according to some criterion. For example, the person may be asked to assign rank 1 to the most preferred alternative, rank 2 to the second-best, and so on. For both tasks, various theoretical accounts have been proposed in the psychological and the economics literature. The purpose of this note is to point out how, under certain conditions, a property of a probability model for rankings (*viz.* 'complete consensus') constrains corresponding probabilistic models for choice. We first introduce some definitions and fix notation.

[1] Institut für Kognitionsforschung, Universität Oldenburg, Germany

Let $T = \{1, \ldots, k\}$ be a set of alternatives labelled arbitrarily as alternative 1 to alternative k and suppose that for all $i \in A \subseteq T$, there is a probability $P(i, A)$ that i is chosen from an available set A. The set $\{P(i, A) : i \in A \subseteq T\}$ is a *complete system of choice probabilities.* If only sets of cardinality two are available, the set $\{P(i, A) : i \in A \subseteq T, \ |A| = 2\}$ is called a *pair comparison system.* A *ranking* of the k alternatives corresponds to a permutation function π from T onto T, where $\pi(i)$ is the rank assigned to alternative i, $i = 1, \ldots, k$. With composition of rankings defined by $(\pi \circ \sigma)(i) = \pi[\sigma(i)]$, the set S_T of all rankings on T constitutes the permutation group. A *transposition* is a permutation τ_{ij} defined by $\tau_{ij}(i) = j$, $\tau_{ij}(j) = i$, and $\tau_{ij}(m) = m$ for all $m \neq i, j$. Note that $\pi \circ \tau_{ij}$ is the permutation that agrees with π except that the ranks assigned to alternative i and alternative j are exchanged. A probability mass function $P(\pi)$, $\pi \in S_T$, represents a *probability model on rankings* (see Critchlow, Fligner, & Verducci [2] for an excellent review of various classes of these models).

Complete Consensus and Order Independence

Note that there are two different, although formally equivalent, ways to think about the probability mass function P on rankings. First, if the data are collected from a single person, $P(\pi)$ is the (multinomial) probability with which ranking π is produced by the person in a series of independent replications of the ranking task. Second, if the data are collected from a population, $P(\pi)$ represents the (multinomial) probability of randomly sampling a person with a fixed ranking π. In both situations, an important question is whether there exists – in a sense to be made precise – an underlying "true" or "consensus" ordering of the alternatives. Define alternative i to be *strongly preferred* to alternative j, written $i \succ_s j$, if for any ranking π such that $\pi(i) < \pi(j)$, $P(\pi) \geq P(\pi \circ \tau_{ij})$ holds with strict inequality for at least one ranking π. A probability model on rankings is said to have *complete consensus* with consensus ranking ν if for any i, j with $\nu(i) < \nu(j)$, $i \succ_s j$. The notion of complete consensus has many implications. For example, the consensus ranking agrees with the ordering of alternatives according to the probability of being ranked first (see Henery [6], Fligner & Verducci [4]). In the following, its implications for corresponding models of choice are explored.

There are numerous ways to derive a probability model for rankings from a system of choice probabilities. The Babington-Smith model, for example, is defined in terms of a pair comparison system (see Joe & Verducci [7] and, for further examples, Critchlow *et al.* [2]). Conversely, the most natural way to define choice probabilities in terms of rankings seems to be the following. The probability to select i from a subset A of available alternatives is given by summing the probability mass function P over all permutations on T

where i is ranked before all other elements of A. Formally,

$$P(i,A) = \sum_{\pi \in R(i,A)} P(\pi) \tag{1}$$

for $i \in A \subseteq T$ where $R(i,A) = \{\pi \in S_T | \pi(i) < \pi(j),\ j \in A \setminus \{i\}\}$. Assumption (1) puts some constraints on a system of choice probabilities. Specifically, it is equivalent to the existence of a *random utility presentation* for $\{P(i,A) : i \in A \subseteq T\}$, i.e., there is a random vector $\mathbf{U}$ with components U_i, $i \in T$, such that $P(i,A) = \Pr(U_i = \max_{j \in A} U_j)$ (see Block & Marschak [1], Falmagne [3] and, for a recent review, Suppes, Krantz, Luce, & Tversky [9], Chpt. 17). Another property of choice probabilities has gained some notoriety in the literature. The choice probabilities are said to satisfy *order-independence* if for all $i, j \in B \setminus C$ and $k \in C$

$$P(i,B) \geq P(j,B) \quad \text{if and only if} \quad P(k, C \cup \{i\}) \leq P(k, C \cup \{j\}) \tag{2}$$

provided the choice probabilities on the two sides of either inequality are not both 0 or 1. In a pair comparison system (2) takes the simpler form

$$p_{ij} \geq \frac{1}{2} \qquad \text{if and only if} \qquad p_{ik} \geq p_{jk} \tag{3}$$

for all i, j, k where p_{ij} denotes $P(i, \{i,j\})$. Order-independence assumes that the ordering of the alternatives is independent of context. Although it may hold in many circumstances, it is clear from both empirical evidence and theoretical considerations that it does not hold in general. For example, let $A = \{i, j, k\}$ and suppose i and j are very similar to each other whereas k is very different from both of them. Assume that all pair comparison probabilities are equal one- half. It follows from order-independence that all trinary choice probabilities equal one-third. It seems unlikely, however, that the choice between j and k is greatly affected by the addition of the very similar alternative i. More generally, it appears that the addition of an alternative to an available set "hurts" alternatives that are similar to it more than those that are dissimilar to it (see Suppes *et al.* [8] for further examples including Savage's case of 'dominated alternative'). These considerations led Tversky's to developing more general models of choice where order- independence is no longer implied (Tversky [9]).

Given the empirical status of the hypothesis of order-independence for choice probabilities the following result seems of interest.

Theorem If a probability model on rankings has complete consensus, then the complete system of choice probabilities (or, the corresponding pair comparison system) defined by (1) satisfies order-independence.

The proof consists of showing that, under complete consensus, either side of (2) implies the other side of (2). For convenience, we consider the pair comparison case, i.e., condition (3). Note that complete consensus implies that all binary probabilities are different from $\frac{1}{2}$ and write (3) as strict inequalities. Obviously, under complete consensus, the left hand side of (3), $p_{ij} > \frac{1}{2}$, implies that i is strongly preferred to j. For the right hand side, we introduce the following notation: let $[ijk]$ be the set of all permutations on T where $\pi(i) < \pi(j) < \pi(k)$. From assumption (1)

$$p_{ij} = \sum_{[ikj]} P(\pi) + \sum_{[ijk]} P(\pi) + \sum_{[jik]} P(\pi)$$

and

$$p_{jk} = \sum_{[jki]} P(\pi) + \sum_{[ijk]} P(\pi) + \sum_{[jik]} P(\pi).$$

Obviously, the right hand side of (3), i.e., $p_{ik} > p_{jk}$, holds if and only if the respective sums in the first position above are ordered accordingly. This, however, follows from $i \succ_s j$. Conversely, $p_{ik} > p_{jk}$ is compatible only with $i \succ_s j$ implying $p_{ij} > \frac{1}{2}$. The proof in the general case, for complete systems of choice probabilities, is analoguous and is omitted here.

In Fligner and Verducci [5] it is noted that the relation of strong preference that underlies the complete consensus concept may be motivated by the idea of "independence from irrelevant alternatives", and it is correctly pointed out that complete consensus is a less restrictive realization of that idea than the wellknown Luce-Plackett model. Nonetheless, the above result relating complete consensus of rankings and order-independence of choice probabilities suggests that complete consensus is a property that may still be too strong in many situations.

References

[1] H. D. Block and J. Marschak. Random orderings and stochastic theories of responses. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann (Eds.), *Contributions to Probability and Statistics*: 97-132. Stanford University Press, Stanford, CA, 1960.

[2] D. E. Critchlow, M. A. Fligner, and J. S. Verducci. Probability models for rankings. *Journal of Mathematical Psychology*, **35**: 294-318, 1991.

[3] J.-C. Falmagne. A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, **18**: 52-72, 1978.

[4] M. A. Fligner and J. S. Verducci. Multistage ranking models. *J. Amer. Statist. Ass.*, **83**: 892-901, 1988.

[5] M. A. Fligner and J. S. Verducci. Posterior probabilities for a consensus ordering. *Psychometrika*, **55**: 53-63, 1990.

[6] R. J. Henery. Permutation probabilities as models for horse races. *J. Roy. Statist. Soc. B.*, **43**: 86-91, 1981.

[7] H. Joe and J. S. Verducci. On the Babington Smith class of models for rankings. *This volume.*

[8] P. Suppes, D. H. Krantz, R. D. Luce, and A. Tversky. *Foundations of Measurements: Volume II. Geometrical, Threshold, and Probabilistic Representations.* Academic Press, San Diego, 1989.

[9] A. Tversky. Elimination by aspects: A theory of choice. *Psychological Review*, **79**: 281-299, 1972.

Ranking From Paired Comparisons by Minimizing Inconsistency

Edwin L. Crow[1]

ABSTRACT A criterion is presented for the best ranking of items or individuals who have been compared in pairs in an unbalanced fashion. The criterion is to choose the ranking or rankings that minimize the sum over all contestants of the absolute differences between the number of net wins over players ranked above and the number of net losses to players ranked below. A method for reaching the minimum is presented. There are two variations of the criterion. They are illustrated on a small set of 1989 tennis player data

Introduction

I consider the problem of ranking k items or individuals when each of them has been compared in separate pairs with one or more of the others and a preference or winner is declared as the result of each comparison. I let n_{ij} be the number of comparisons between individual (player) i and individual j with $n_{ij} \geq 0$ and s_{ij} the number of times i wins, so that $s_{ij} + s_{ji} = n_{ij}$ $(i, j, = 1, 2, \ldots, k, i \neq j)$. The extensive literature on the problem has been summarized and unified by David (1988). Here I propose a nonparametric criterion for a best ranking and a method for attaining it.

The data may be presented in a win-loss chart (preference table), a matrix with the i-th row and the i-th column identified with player i and s_{ij}

[1] Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U. S. Department of Commerce, Boulder, Colorado 80303

the element of the i-th row and j-th column. Only $s_{ij} > 0$ need be entered. If the players are ordered according to their ability to defeat other players and the abler player always wins, then all the s_{ij} are above the main diagonal. If the abler player does not always win, then an upset or inconsistency with the given order or ranking occurs, and an s_{ij} appears below the main diagonal. Slater (1961) proposed to determine a ranking by minimizing the *total number of inconsistencies* (TNI). The TNI provides a nonparametric criterion for the efficacy of a given ranking. However, if $n_{ij} > 1$, an upset may be nullified by one or more opposite outcomes of the n_{ij} comparisons. Thus a more appropriate criterion when the n_{ij} may exceed 1 is the *net number of inconsistencies* (NNI), in which only non-nullified upsets are summed (Crow, 1990). More symmetrically, two players having equal numbers of wins against each other would result in a contribution of $\frac{1}{2}$ to NNI for each mark below the main diagonal.

Still another nonparametric criterion is suggested by a pair of results printed out by the STAR computer program (System for Tournament Administration and Ranking) made available to rankers by the United States Tennis Association, *WnAb* (number of wins above) and *LsBl* (number of losses below) for each player. These are the inconsistencies. Crow (1990) included the criterion

$$WALB = \sum |WnAb - LsBl|,$$

the sum of the absolute differences between wins above and losses below. However, a better criterion, by analogy with the advantage of NNI over TNI, is the sum of the absolute differences between *net* wins above and *net* losses below,

$$NWALB = \sum |NWnAb - NLsBl|.$$

since any wins above may be partly or wholly nullified by losses above to the same players and likewise for losses below. In the case of two players with equal numbers of wins against each other, one is credited with half a win, the other half a loss for each mark below the main diagonal. Only net wins above and net losses below provide true inconsistency with any given ranking. The criterion may result in several equally good rankings and indicate that the players should be tied (co-ranked). A further criterion, David's score (1987; 1988, p. 108) in particular, may be used to break a tie.

A further modification in the criterion is proposed. Two or more net inconsistencies of the same type (win or loss) by a player against the same player are to be counted only once; thus the inconsistency is measured by the number of players out of line rather than the number of matches. This is analogous to the reckoning in David's score, in which a series of matches between any two players has a total value of 1.

The *NWALB* criterion has an advantage over *NNI*: The algebraic term for each player tells how to improve the ranking. If $NWnAb - NLsBl > 0$, move him up; in the contrary case, down. Furthermore, move the player with the largest absolute difference first. This type of operation is then repeated, on the player with the remaining largest absolute difference (who may or may not be the same as the first player moved). The process is continued until all the rankings yielding the minimum *NWALB* have been found. Since a win above by any player is a loss below for his opponent, the algebraic sum of the $NWnAb - NLsBl$ is zero, providing a check.

Relatively few of the $k!$ possible rankings should have to be considered. The initial ranking can be made by using David's score. The process can be easily programmed for, and performed by, a computer.

Although the best ranking for a group of players is obtained by minimizing *NWALB* over the entire group this criterion can be applied to any convenient subgroup. Inconsistencies may often occur over a limited range of ranks. If a subgroup has no inconsistencies outside it, then the minimum for it is part of the solution for the entire group. In other words, the global minimum can be obtained by combining regional minima.

An illustration of the use of the criterion is provided by a subgroup of four closely matched players in the 70-&-over men's singles of the 1989 Colorado Tennis Association annual ranking: Cougnenc (Co), Herr (He), Tanner (Ta), and Trostorff (Tr). They had no wins over the top three players and no losses to the only other player eligible for ranking, so minimizing *NWALB* over them provides their correct order in the overall ranking. Their initial order in the table below is determined by their David scores S, which include results of matches against the other five players as well as against several other, ineligible players (players with too few tournaments and non-residents).

	4	5	6	7	Total	NWab	NLB1	Diff.	S
4. Ta	-	1	1		2	0	1	-1	0
5. He		-	1	2	3	0	$\frac{1}{2}$	$-\frac{1}{2}$	$-2\frac{1}{2}$
6. Co		1	-	1	2	$\frac{1}{2}$	0	$\frac{1}{2}$	$-5\frac{1}{2}$
7. Tr	1			-	1	1	0	1	-7
Total	1	2	2	3	8	$1\frac{1}{2}$	$1\frac{1}{2}$	3	(Absolute)

Can this order be improved by moving players as indicated by the differences, i.e., Ta and He down and Co and Tr up? I try the following order:

	4	5	6	7	Total	NWab	NLB	Diff.
4. Co	-		1	1	2	0	$-1\frac{1}{2}$	$-1\frac{1}{2}$
5. Ta	1	-		1	2	1	1	0
6. Tr		1	-		1	1	1	0
7. He	1		2	-	3	$1\frac{1}{2}$	0	$1\frac{1}{2}$
Total	2	1	3	2	8	$3\frac{1}{2}$	$3\frac{1}{2}$	3

This is as good a ranking as the first one according to $NWALB$, even though $TNI = 5$ versus 2 and $NNI = 4\frac{1}{2}$ versus $1\frac{1}{2}$. Similarly one can confirm that the order TaCoHeTr yields $NWALB = 3$, but five other orders also turn out to have $NWALB = 3$, while nine have $NWALB = 4$ and seven have $NWALB = 5$ (completing the 24 permutations). One can resolve the tie by choosing the original order above on the basis of the NNI or the David scores, but the large number of ties is disturbing. Perhaps the $NWALB$ criterion is not a very refined measuring tool, although it must be recognized that the four players are closely bunched.

At the Amherst Conference for ranking data David commented on a presentation of the above criterion that it gave no more credit for a big upset than for a mild upset. That can be allowed for by adding the difference in ranks for each upset instead of simply tallying it, thus obtaining the criterion

$$NDWALB = \sum |NDWnAb - NDLsBl|$$

where $NDWnAb$ denotes the difference in ranks of the two players in a net upset win and $NDLsBl$ the same in a net upset loss. This criterion makes sense; a big upset raises more question than a mild one. For the first table above, the differences for this criterion are -3, $-\frac{1}{2}$, $\frac{1}{2}$, 3 respectively for a numerical total of 7, whereas the second table yields a total of 5, the minimum, as does also the order HeTaTrCo. $NDWALB$ is reminiscent of the criterion Net Difference in Ranks (NDR) (Crow, 1990), but it differs in considering only upsets and in classifying them.

The complete distribution of NDWALB over the 24 permutations of the four players, as well as that of TNI, is:

Value	2	3	4	5	6	7	8	9	10	11
NDWALB Frequency				2	8	2	1	2	7	2
TNI Frequency	2	6	8	6	8					

In the example at least, and apparently often, $NDWALB$ thus does substantially better than $NWALB$ in reducing ties for the best ranking. The NNI is not tabulated above because in this example (not in general) it is uniformly less than TNI by a constant. The criteria $NWALB$ and $NDWALB$ are attractive in corresponding closely to the views of players being ranked about the fairness of their individual rankings; they believe they should never fall below a player they have defeated unless there is a counteracting loss to a player they outrank.

References

[1] E. L. Crow. Ranking paired contestants. *Commun. Statist. Simula.*, **19**:749-769, 1990.

[2] H. A. David. Ranking from unbalanced paired comparison data. *Biometrika,* **74**:432-436, 1987.

[3] H. A. David. *The Method of Paired Comparisons,* 2nd ed., Oxford University Press, New York, 1988.

[4] P. Slater. Inconsistencies in a schedule of paired comparisons. *Biometrika,* **48**:303-312, 1961.

ADDENDUM

A refinement of the *NWALB* criterion that makes fuller use of the data is preferable. This is achieved by defining the net number of wins above of a given player, i, against any particular player, j, ranked above him, *NWnAb*, as the proportion of the matches between them won by the given player (s_{ij}/n_{ij} in the notation of the first paragraph). The same change in definition applies to *NLsBl.* In practice there is no change from the previous definition unless the two players play three or more matches against each other. The refined *NWALB* criterion is slightly less subject to ties than the previous criterion.

Graphical Techniques for Ranked Data

G. L. Thompson[1]

Exploratory graphical methods are critically needed for displaying ranked data. Fully and partially ranked data are functions on the symmetric group of n elements, S_n, and on the related coset spaces. Because neither S_n nor its coset spaces have a natural linear ordering, graphical methods such as histograms and bar graphs are inappropriate for displaying ranked data. However, a very natural partial ordering on S_n and its coset spaces is induced by two reasonable measures of distance: Spearman's ρ and Kendall's τ. Graphical techniques that preserve this partial ordering can be developed to display ranked data and to illustrate related probability density functions by using permutation polytopes. A polytope is the convex hull of a finite set of points in $\Re^{n-1}$, and a permutation polytope is the convex hull of the $n!$ permutations of n elements when regarded as vectors in $\Re^n$ (see, for example, Yemelichev, et. al.[9]). This concept is closely related to the observation by McCullagh [7] that the $n!$ elements of S_n lie on the surface of a sphere in $\Re^{n-1}$.

To illustrate a rudimentary version of the proposed graphical technique, we will consider the paired ranking data of Critchlow and Verducci [4] in which $n = 4$. This data consists of pairs of orderings in which 38 students have ranked 4 styles of literary criticism in their order of preference, both before and after taking a course. The question of interest is whether the students' preferences have moved toward the teacher's preferred ordering $< p, c, a, t >$, i.e, toward the ordering in which style p is ranked first, style c is ranked second, style a is ranked third, and style t is ranked fourth. As illustrated by figure 1 of McCullagh [7], the 24 possible orderings of 4 items form the vertices of an Archimedean solid, the truncated octahedron, which is the permutation polytope for $n = 4$. The edges connect orderings

[1] Department of Statistical Science, Southern Methodist University, Dallas, Texas

that differ by exactly one pairwise transposition so that Kendall's τ is the minimum number of edges that must be traversed to get from one ordering to another. The straight line distance between two points is proportional to Spearman's ρ. In the proposed graphical technique, the frequency with which each ordering appears in a data set is visually indicated, for example, by the size of a circle, at the corresponding vertex. The frequencies of the 38 pre-rankings are shown in Figure 1 and the 38 post-rankings are shown in Figure 2. To make the plots perceptually accurate, the areas of the circles are based on Steven's Law (Cleveland [1]) which says that a person's perceived scale, p, of the size of an area is proportional to $(\text{area})^{.7}$. Hence the radii are scaled so as to be proportional to $(\text{frequency})^{5/7}$.

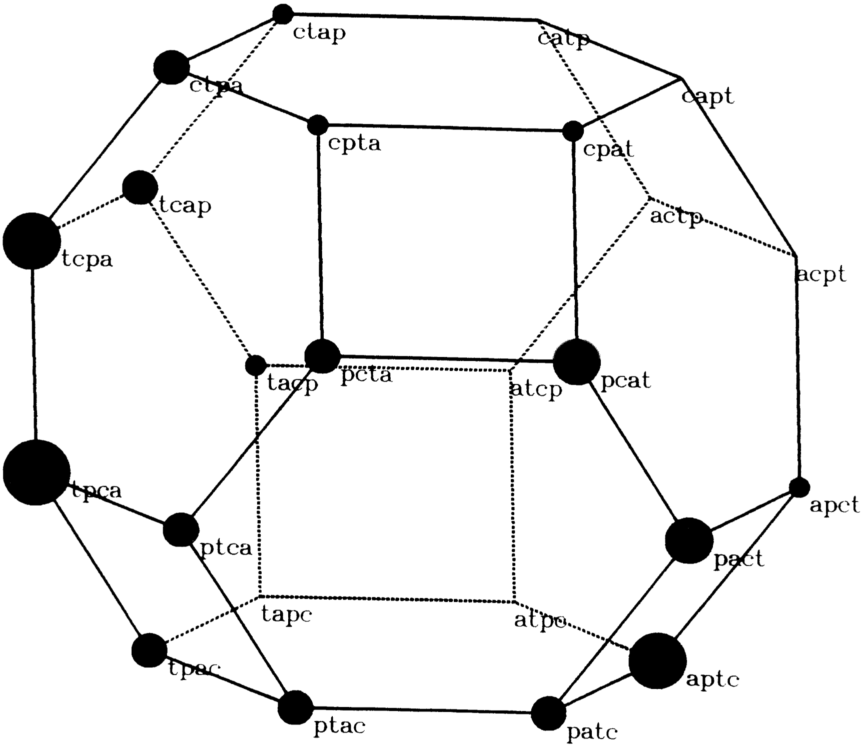


Fig. 1: Frequencies of the pre-rankings.

Although the bivariate nature of the data is lost, valuable insight into this data can be obtained from the above figures. Just as with paired univariate data, it is often fruitful to do exploratory data analysis by comparing the histograms of the "before" and "after" components. When Figures 1 and 2 are compared, it is obvious that the frequencies do change a great deal between the two sets of rankings. There is a notable increase in the frequencies of the vertices of the hexagon corresponding to the 6 orderings beginning with c, and a decrease in the frequencies of the 6 vertices corresponding to orderings that end with c. Hence, it appears that c has a higher level of preference in the post-rankings than in the pre-rankings. In fact, one might hypothesize that the orderings have moved toward $< c, p, t, a >$ because almost half of the post-rankings lie either on $< c, p, t, a >$ or on the three vertices within one edge (pairwise transposition) of $< c, p, t, a >$. This is not inconsistent, however, with the conclusion of Critchlow and Verducci's [4] that the rankings have moved closer to $< p, c, a, t >$. Because the figures

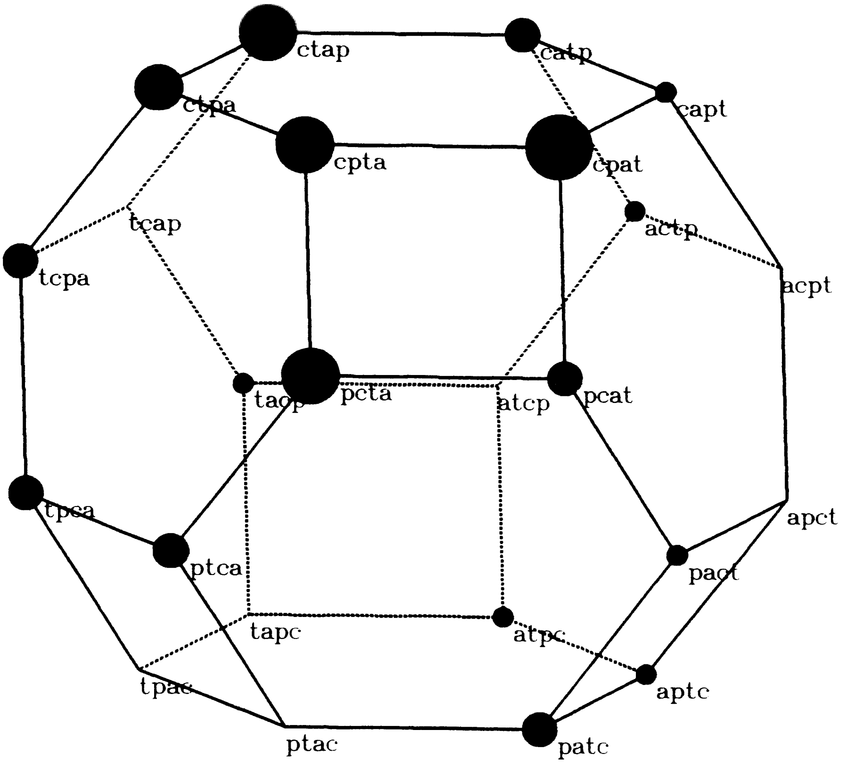


Fig. 2: Frequencies of the post-rankings.

are based on partial ordering of S_n, and not on a full linear ordering, movement closer to a permutation is not necessarily the same as movement toward it. Figures 1 and 2 also show that 1) style a is rarely chosen as either first or second choice after the course is completed; 2) the incidence of style t as a first choice decreases; and 3) there does not seem to be any movement toward $< a, c, p, t >$, the ordering of the styles on the second questionnaire. Different rotations of Figures 1 and 2 would make some of these observations more apparent.

Figures 1 and 2 are fairly elementary examples of the potential of permutation polytopes in developing graphical techniques for ranked data. Significant improvements are immediately possible in the following three areas. First, the availability of information would be greatly enhanced if the truncated octahedrons could be arbitrarily rotated about any axis. This can be accomplished via interactive software, especially if the coordinates of the permutation polytope are known. For S_n the coordinates of the vertices of the permutation polytope in $\Re^{n-1}$ with center at $\mathbf{0}$ are found by using the Helmert transformation. Let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)' \epsilon S^n$ be any ranking (not ordering) of n items. This means that the item i has rank π_i. Let $\mathbf{r}$ be the n dimensional column vector in which every element equal $(n+1)/2$, and let H be the Helmert transformation which maps the hyperplane $\sum_{i=1}^{n} x_i = 0$ onto the hyperplane $x_n = 0$. Note that H is orthogonal and preserves Euclidian distances. The coordinates of the vertices of the permutation polytope in $\Re^{n-1}$ are $H(\pi - \pi)$, $\pi \epsilon S^n$. At the same time capabilities are introduced to rotate the polytopes, the frequencies associated with each vertex could be color coded on a scale chosen to highlight desired features. With color and rotational capabilities, it also might be worthwhile to experiment with illustrating the observed frequencies on the duals of the polytopes. For $n = 4$, the dual of the truncated octahedron is the tetrakis hexahedron. It has 24 faces which, instead of the 24 vertices of the truncated octahedron, would correspond to the 24 possible orderings. See Cundy and Rollett [5] for a discussion of the duals of Archimedean solids.

Second, these graphical techniques can be extended to the case where n is greater than 4 even though the dimension of the corresponding permutation polytope is greater than three. One approach to this is interactive software to successively view adjacent three dimensional faces. Possibly, the faces of greatest interest are those generated by the set of orderings in which four items are permuted while the remaining $n - 4$ items remain fixed. Characterizations of all the faces of permutation polytopes are given by Yemelichev, et. al. [9]. The application of other new graphical methods of viewing higher dimensions may be very useful.

Third, the above arguments can be extended to partially ranked data by considering the multiset $M = (1^{a_1}, 2^{a_2}, \ldots, k^{a_k})$. This multiset has k levels and a_i items are ranked together in the i-th level. It is assumed that $a_i \epsilon \mathcal{Z}^+$ and $\sum_{i=1}^{k} a_i = n$. Stanley [8] discusses multisets. The set of permutations of M, when regarded as vectors in $\Re^n$, become the vertices of an integral polytope. These vertices can be shown to lie on an $n-1$ dimensional sphere $\Re^n$. Just as with fully ranked data, software to rotate the vertices and techniques to view the three dimensional faces for $n > 4$ are needed. The characterizations of the faces of integral polytopes formed by permutations of multisets is a straightforward generalization of the results in Yemelichev, et. al. (1984) for permutation polytopes. These integral polytopes can also be thought of as Cayley diagrams (see Coxeter and Moser [2]) and have interesting connections with the generators of the coset space $S_n/\Pi S_{a_i}$. Furthermore, they induce an interesting extension of Kendall's τ to partially ranked data that has properties quite different from both the Haussdorf metric (Critchlow [3]) and the metric $i(\pi)$ discussed by Diaconis [6]. Similarly, the straight line distance between vertices in $\Re^n$ is an extension of Spearman's ρ.

References

[1] W. S. Cleveland. *The Elements of Graphing Data*, Monterey: Wadsworth Advanced Books and Software, 1985.

[2] H. S. M Coxeter and W. O. Moser. *Generators and Relations for Discrete Groups* (4th ed.), Berlin: Springer-Verlag, 1980.

[3] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*, Berlin: Springer-Verlag, 1985.

[4] D. E. Critchlow and J. S. Verducci. "Detecting a Trend in Paired Rankings," Technical Report #418, Department of Statistics, The Ohio State University, 1989.

[5] H. M. Cundy and A. P. Rollett. *Mathematical Models*, London: Oxford University Press. 1951.

[6] P. Diaconis. *Group Representations in Probability and Statistics*, Hayward, California: Institute of Mathematical Statistics. 1988.

[7] P. McCullagh. "Models on Spheres and Models on permutations," Technical Report, Department of Statistics, University of Chicago. 1990.

[8] R. Stanley. *Enumerative Combinatorics*, Monterey, California: Wadsworth and Brooks/Cole. 1986.

[9] V. A. Yemelichev, M. M. Kovalev and M. K. Kravtsov. *Polytopes, Graphs and Optimisation*, Cambridge: Cambridge University Press, 1984.

Matched Pairs and Ranked Data

Peter McCullagh[1]
Jianming Ye[2]

ABSTRACT We consider the Babington Smith and Bradley-Terry Models for ranked data. Both models are based on inversions. In a matched pairs design the pair-specific nuisance parameters are eliminated by a conditioning argument. The conditional likelihood has a form similar to that of a logistic model, so that conditional likelihood computations are straightforward. An example previously considered by Critchlow and Verducci is analysed using the new method.

Introduction

We consider a class of models for a matched pairs design in which each response is a permutation of k objects. All models considered include a pair-specific vector-valued nuisance parameter, together with a treatment effect that is assumed to be constant across pairs. As is usual in matched pairs models, no assumptions are made concerning the pair-specific parameters, though in some circumstances it might be reasonable to assume that they are independent and identically distributed according to an arbitrary unspecified distribution. The treatment effect, on the other hand, is modelled in such a way as to take account of the nature of the reponse variable. We assume here that the effect of treatment is to make certain inversions relatively more likely. The treatment effect is therefore a vector-valued parameter having $k(k-1)/2$ components, although an important

[1] Department of Statistics, University of Chicago
[2] Department of Statistics, University of Chicago

sub-model having rank $k-1$ is also considered.

The pair-specific nuisance parameters are eliminated by using a conditioning argument previously employed for matched binary pairs by Cox [1], and for multinomial matched pairs by McCullagh [6]. This argument follows the general theory for constructing similar regions (See Lehmann [4], chapter 4). The resulting conditional models are related to quasi-symmetry and generalize the first example in McCullagh [6] to the case where the response is a permutation rather than a purely nominal set of response categories. One the other hand, they generalize the model for permutation data (See McCullagh [5]) to allow for control and treatment effect. Since the conditional likelihood function has the formal appearance of a likelihood derived from a series of Bernoulli trials with varying probability, standard computer programs such as GLIM can be used to compute the conditional maximum likelihood estimate and its asymptotic standard error.

The Model

In a matched pairs experiment, $(\mathbf{y}_i, \mathbf{y}_i^{'})$ is observed for each pair. Here the first response refers to the control and the second to the treated individual, with each response a permutation of k objects, say, a, b, $\ldots$, k. So, each $\mathbf{y}$ has $k!$ possible values. Our interest is to investigate the effect of treatment on the response probabilities.

For example, if the response for each individual is a permutation of 3 objects, a, b and c, then $\mathbf{y}$ takes one of the 6 values, *abc*, *acb*, *bac*, *bac*, *cab* or *cba*. We take alphabetical order as the standard and measure all inversions relative to this order. Any two letters, not necessarily adjacent in the sequence, are either in standard order or inverted. So, in *cab*, *ab* is in standard order, *ca*, *cb* are in inverted order. A permutation is uniquely determined by the order of all pairs. This kind of inversion is called a 'first-order inversion'. The incidence matrix $\mathbf{X}$, of order $k! \times k(k-1)/2$, is a matrix of 0s and 1s whose rows are indexed by permutations and whose columns are indexed by inversions. For the above example with $k=3$, the first-order incidence matrix is

$$\mathbf{X} = \begin{array}{c} \\ abc \\ acb \\ bac \\ bca \\ cab \\ cba \end{array} \begin{array}{c} \begin{array}{ccc} ba & ca & cb \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}$$

Let $\boldsymbol{\pi}_i$, $\boldsymbol{\pi}_i^{'}$ be $k! \times 1$ probability vectors giving the probability distribution for the ith pair and let $\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i^{'}$ be the vectors of log probabilities. Consider the following model:

$$\boldsymbol{\eta}_i = \boldsymbol{\lambda}_i \tag{1}$$

$$\boldsymbol{\eta}_i^{'} = \boldsymbol{\lambda}_i + \mathbf{X}\boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{\lambda}_i$ is a vector-valued nuisance parameter specific to the ith pair, $\boldsymbol{\beta}$ measuring the effect of treatment is a $k(k-1)/2$ dimensional parameter vector assumed constant across pairs, and $\mathbf{X}$ is the incidence matrix of first-order inversions. Let X_ω be the ωth row of $\mathbf{X}$, and let $\pi_{i\omega}$, $\eta_{i\omega}$ be the elements of the probability vector and log probability vector for the ith pair, corresponding to $\omega \in \Omega$, where Ω is the set of all possible permutations of the k objects. The likelihood contribution from one pair is then (omitting the subscript i)

$$\begin{aligned}\pi_y \pi_{y'}^{'} &= \frac{\exp\{\lambda_{\mathbf{y}}\}\exp\{\lambda_{\mathbf{y}'} + X_{\mathbf{y}'}\boldsymbol{\beta}\}}{\Sigma_{\omega\in\Omega}\exp\{\lambda_\omega\}\Sigma_{\omega'\in\Omega}\exp\{\lambda_{\omega'} + X_{\omega'}\boldsymbol{\beta}\}} \\ &= \frac{\exp\{\Sigma_{\omega\in\Omega}\lambda_\omega[I(\mathbf{y}=\omega) + I(\mathbf{y}^{'}=\omega)] + X_{\mathbf{y}'}\boldsymbol{\beta}\}}{\Sigma_{\omega\in\Omega}\exp\{\lambda_\omega\}\Sigma_{\omega'\in\Omega}\exp\{\lambda_{\omega'} + X_{\omega'}\boldsymbol{\beta}\}}.\end{aligned}$$

If $\boldsymbol{\beta}$ is known, the minimal sufficient statistics for $\boldsymbol{\lambda}$ is the unordered set $S_{\boldsymbol{\lambda}} = \{\mathbf{y}, \mathbf{y}^{'}\}$, which has marginal distribution

$$\begin{aligned}&\mathrm{pr}\{(\mathbf{Y}, \mathbf{Y}^{'}) = (\mathbf{y}, \mathbf{y}^{'}) \text{ or } (\mathbf{y}^{'}, \mathbf{y})\} \\ &= \frac{\exp\{\Sigma_{\omega\in\Omega}\lambda_\omega[I(\omega=\mathbf{y}) + I(\omega=\mathbf{y}^{'})]\}(e^{X_{\mathbf{y}}\boldsymbol{\beta}} + e^{X_{\mathbf{y}'}\boldsymbol{\beta}})}{\Sigma_{\omega\in\Omega}\exp\{\lambda_\omega\}\Sigma_{\omega'\in\Omega}\exp\{\lambda_{\omega'} + X_{\omega'}\boldsymbol{\beta}\}}.\end{aligned}$$

Standard procedure for generating similar regions (Lehmann 1986, chapter 4) leads to consideration of the conditional distribution of the ordered pair $(\mathbf{y}, \mathbf{y}^{'})$ given $S_{\boldsymbol{\lambda}}$. This conditional distribution depends only on the treatment parameter $\boldsymbol{\beta}$ and not on the pair-specific parameter $\boldsymbol{\lambda}_i$. The contribution to the conditional likelihood given by the ith pair is then the probability that $\mathbf{Y}$ takes the value $\mathbf{y}$ conditional on $\mathbf{Y} = \mathbf{y}$ or $\mathbf{y}^{'}$, i.e. $\mathrm{pr}\{Y_i = \mathbf{y}_i \mid \{\mathbf{Y}_i, \mathbf{Y}_i^{'}\} = \{\mathbf{y}_i, \mathbf{y}_i^{'}\}\}$. Omitting the subscript i, this is equal to

$$\frac{e^{X_{\mathbf{y}'}\boldsymbol{\beta}}}{e^{X_{\mathbf{y}}\boldsymbol{\beta}} + e^{X_{\mathbf{y}'}\boldsymbol{\beta}}} = \frac{e^{(X_{\mathbf{y}'} - X_{\mathbf{y}})\boldsymbol{\beta}}}{1 + e^{(X_{\mathbf{y}'} - X_{\mathbf{y}})\boldsymbol{\beta}}}. \tag{3}$$

For example, if $\mathbf{y} = (cabd)$ and $\mathbf{y}^{'} = (acdb)$ then the inversions vectors $X_{(cabd)}$ and $X_{(acdb)}$ in the order (ba, ca, da, cb, db, dc) are $(0,1,0,1,0,0)$ and $(0,0,0,1,1,0)$ respectively and the common inversions cancel. The resulting procedure is closely analogous to working with differences in the standard normal-theory matched pairs problem. The conditional likelihood for this pair is

$$\frac{1}{1 + \exp\{\beta_{ca} - \beta_{db}\}}.$$

The interpretation of the parameters is now clear. For example, β_{db} measures the effect of treatment on the log probability that d precedes b. If $\beta_{db} > 0$, the order db is more probable after the treatment than before if all other relative rankings remain unchanged. Finally, we note that if $\mathbf{y} = \mathbf{y}'$, the contribution to conditional likelihood is constant and can be ignored in the conditional analysis.

Let $\mathbf{d}_i = X_{\mathbf{y}'_i} - X_{\mathbf{y}_i}$ be the vector of net changes in inversions for the ith pair. The conditional likelihood contribution (3) is the same as the likelihood contribution of a success in a Bernoulli experiment $B(1, \pi)$, with $\pi = \exp(\mathbf{d}_i\boldsymbol{\beta})/(1 + \exp(\mathbf{d}_i\boldsymbol{\beta}))$. As a consequence numerical maximization of the conditional likelihood is straightforward using programs for fitting linear logistic models. The full conditional likelihood is

$$\prod \frac{e^{X_{\mathbf{y}'}\boldsymbol{\beta}}}{e^{X_{\mathbf{y}}\boldsymbol{\beta}} + e^{X_{\mathbf{y}'}\boldsymbol{\beta}}} = \frac{\exp(\sum \mathbf{d}_i\boldsymbol{\beta})}{\prod(1 + \exp(\mathbf{d}_i\boldsymbol{\beta}))} \tag{4}$$

Note that the denominator on the left is symmetric in the pair $(\mathbf{y}_i, \mathbf{y}'_i)$, so that $\sum \mathbf{d}_i$ is sufficient for $\boldsymbol{\beta}$ conditional on $\{S_{\boldsymbol{\lambda}_i}\}$. If we use 1,2,3,4 to indicate the rank of an object in a ranking, as in $dcba$ where b takes rank 3, then $\sum \mathbf{d}_i$ is the aggregate net total change in inversion for every two-object combination.

To test transitivity (see Critchlow [2]), we may make the further assumption that

$$\beta_{ij} = \theta_i - \theta_j \quad \text{for } i \neq j \tag{5}$$

This means, if order ba is relatively more probable than ab in the treatment group ($\beta_{ba} > 0$) and ac is more probable than ca ($\beta_{ac} = -\beta_{ca} > 0$), then bc should be relatively more probable than cb after treatment ($\beta_{bc} = -\beta_{cb} > 0$) and $\beta_{cb} = \beta_{ca} + \beta_{ab}$. This is a submodel of the first-order inversion model and usually referred to as the Bradley-Terry model. The order of the θs measures the change in relative ranking of objects. The reduced model (2) now becomes

$$\boldsymbol{\eta}_i = \boldsymbol{\lambda}_i + \mathbf{X}^o\theta. \tag{6}$$

where $\mathbf{X}^o$ is the incidence matrix corresponding to (5). For the case $k = 3$

$$\mathbf{X}^o = \mathbf{X}\begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \begin{array}{c|ccc} & a & b & c \\ \hline abc & 0 & 0 & 0 \\ acb & 0 & -1 & 1 \\ bac & -1 & 1 & 0 \\ bca & -2 & 1 & 1 \\ cab & -1 & -1 & 2 \\ cba & -2 & 0 & 2 \end{array}.$$

Since $\mathbf{X}^o$ has rank $(k - 1)$, one of the θs can be set to zero. Every row indicates the rank changes of the k objects in a permutation relative to their rank in standard order. Let $\mathbf{d}_i^o = X^o_{\mathbf{y}'_i} - X^o_{\mathbf{y}_i}$. The full conditional likelihood corresponding to (4) is

$$\frac{\exp(\sum \mathbf{d}_i^o \theta)}{\prod(1+\exp(\mathbf{d}_i^o \theta))} \tag{7}$$

so $\sum \mathbf{d}_i^o$, the net change in rank for each object, is sufficient for θ conditional on $\{S_{\boldsymbol{\lambda}_i}\}$.

Example

In the example discussed by Critchlow and Verducci [3], 38 students rank four styles of textual criticism before and after a course in writing and literary criticism. The four styles are "Authorial"(A), "Comparative"(C), "Personal"(P) and "Textual"(T). They show that the post-treatment rankings have moved toward the direction of the teacher's own ranking, which is considered as "idealized".

To use the models presented in the last section, we need maximize the conditional likelihood, which is a product of terms like (3), one for each pair. The estimates for the first-order inversion model, obtained using GLIM, are as follows:

Parameter	$\sum \mathbf{d}_i$	*Estimate*	*s.e.*
CA	7	0.529	0.872
PA	-2	0.009	0.928
PC	-13	-1.500	0.843
TC	-14	-1.281	0.766
TP	-1	-0.240	0.748

The scaled deviance reduction due to inversions *PC* and *TC* is 16.88 on 2 degrees of freedom, and all the other factors contribute only 0.55 on 4 degrees of freedom. When *CA*, *PA*, *TA*, *TP* are excluded from the model, the estimates of parameters are

Parameter	$\sum \mathbf{d}_i$	*Estimate*	*s.e.*
PC	-13	-1.448	0.683
TC	-14	-1.529	0.670

From these analyses we conclude that the orders *CT* and *CP* are much more probable after the course than before, the odds being increased by 4.61 and 4.25, respectively. It does not follow, though it happens in this case to be true, that the post-treatment orders *CT* and *CP* are more probable than *TC* and *PC*, respectively.

The Bradley-Terry model fits almost as well as the first-order inversion model, the deviance increasing by only 0.28 on three degrees of freedom. There is no evidence of lack of transitivity in the treatment effect. The parameter estimates are as follows, showing that the major effect of the course is to decrease the rank (or to enhance the perceived importance) of C mainly at the expense of P and T.

Parameter	$\sum d_i^o$	*Estimate*	*s.e.*
A	3	0.000	0.000
C	-33	0.853	0.374
P	13	-0.309	0.372
T	17	-0.439	0.345

Figure 1 is an attempt to depict the rank vectors geometrically. The 4! sample points lie on the surface of a sphere as shown in the diagram. Neighbouring points differ by one transposition of adjacent letters. The point $\bar{y}$ is the average of the pre-course ranks, namely $(2.95, 2.74, 1.97, 2.34)$ in the order A, C, P, T. The modal pre-course 'direction' is thus closest to PTCA. Note that, although the individual sample points lie on the surface of the sphere, averaged ranks lie in the convex hull of the vertices. If the pre-course ranks were uniformly distributed $\bar{y}$ would be depicted as having zero length. Thus, the origin O in the diagram corresponds to the 'rank vector' $(2.5, 2.5, 2.5, 2.5)$.

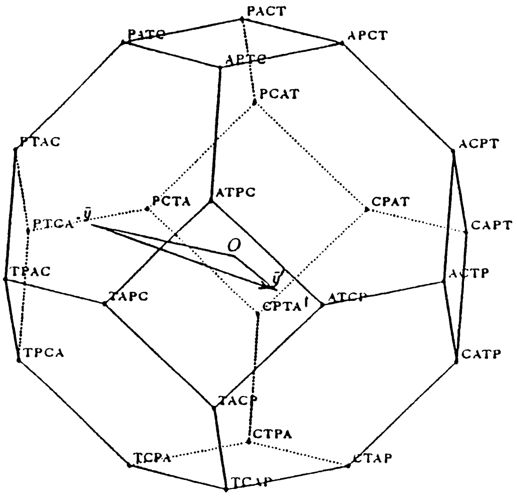


Figure 1: *Sample space of permutations of ACPT. The graph has 24 vertices, 36 edges, 6 square faces, and 8 hexagonal faces. Vectors y and y' are shown enlarged by a factor of 2.*

The average post-course rank vector is $(3.03, 1.87, 2.32, 2.79)$, denoted by $\bar{y}'$ in the diagram, so that the modal post-course direction is closest to CPTA. The difference vector, on which our conditional analysis is based, is $(0.079, -0.87, 0.34, 0.45)$. The effect of treatment is to move the rank toward $(C, A, -, -)$, i.e. in the direction between (C, A, P, T) and (C, A, T, P), even though A has the lowest rank before and after treatment, and the favoured post-treatment ranking is (C, P, T, A). Critchlow and Verducci (1989) find that the post-rankings are getting closer to the assumed idealized ranking, (P, C, A, T), but here we show that the idealized ranking is not the exact direction of movement and the effect of the writing course is not exactly the same as what the teacher wanted it to be.

Acknowledgements:

This research was supported in part by the National Science Foundation Grant DMS 88-01853. This manuscript was prepared using computer facilities supported in part by the National Science Foundation Grants DMS 89-05292, DMS 87-03942, and DMS 86-01732 awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

REFERENCES

[1] D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, **34**:562–565, 1958.

[2] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data*, **34** of *Lecture Notes in Statistics*. New York: Springer-Verlag, 1985.

[3] D. E. Critchlow and J. S. Verducci. Detecting a trend in paired rankings. *Applied Statistics*, 1992. To appear.

[4] E. L. Lehmann. *Testing Statistical Hypotheses.* New York: Wiley, 1985.

[5] P. McCullagh. Permutations and regression models. *In this volume.*

[6] P. McCullagh. Some applications of quasisymmetry. *Biometrika*, **69**:303-308, 1982.

[7] P. McCullagh. Models on spheres and models for permutations. *In this volume*, 1990.

Pre- and Post-Course Rankings of Four Critical Styles

Subject No.	*Pre-Ranking*	*Post-Ranking*	*Subject No.*	*Pre-Ranking*	*Post-Ranking*
	A C P T	A C P T		A C P T	A C P T
10	2 4 1 3	3 1 4 2	29	3 4 1 2	4 2 1 3
11	3 4 1 2	3 1 4 2	30	3 2 1 4	2 3 1 4
12	3 2 1 4	1 2 4 3	31	4 3 2 1	2 4 1 3
13	1 3 2 4	3 1 4 2	32	2 3 4 1	4 1 3 2
14	3 2 1 4	2 1 4 3	33	2 3 1 4	3 1 2 4
15	4 3 2 1	4 2 1 3	34	3 2 4 1	2 1 4 3
16	4 1 3 2	1 4 3 2	35	2 3 1 4	1 4 2 3
17	4 2 1 3	4 2 3 1	36	4 3 2 1	4 1 2 3
18	1 4 2 3	4 3 2 1	37	3 4 2 1	4 3 1 2
19	3 1 2 4	3 2 1 4	38	4 2 3 1	4 3 1 2

†Source: Critchlow and Verducci [3].

Vol. 1: R.A. Fisher: An Appreciation. Edited by S.E. Fienberg and D.V. Hinkley. XI, 208 pages, 1980.

Vol. 2: Mathematical Statistics and Probability Theory. Proceedings 1978. Edited by W. Klonecki, A. Kozek, and J. Rosinski. XXIV, 373 pages, 1980.

Vol. 3: B.D. Spencer, Benefit-Cost Analysis of Data Used to Allocate Funds. VIII, 296 pages, 1980.

Vol. 4: E.A. van Doorn, Stochastic Monotonicity and Queueing Applications of Birth-Death Processes. VI, 118 pages, 1981.

Vol. 5: T. Rolski, Stationary Random Processes Associated with Point Processes. VI, 139 pages, 1981.

Vol. 6: S.S. Gupta and D.-Y. Huang, Multiple Statistical Decision Theory: Recent Developments. VIII, 104 pages, 1981.

Vol. 7: M. Akahira and K. Takeuchi, Asymptotic Efficiency of Statistical Estimators. VIII, 242 pages, 1981.

Vol. 8: The First Pannonian Symposium on Mathematical Statistics. Edited by P. Révész, L. Schmetterer, and V.M. Zolotarev. VI, 308 pages, 1981.

Vol. 9: B. Jørgensen, Statistical Properties of the Generalized Inverse Gaussian Distribution. VI, 188 pages, 1981.

Vol. 10: A.A. McIntosh, Fitting Linear Models: An Application of Conjugate Gradient Algorithms. VI, 200 pages, 1982.

Vol. 11: D.F. Nicholls and B.G. Quinn, Random Coefficient Autoregressive Models: An Introduction. V, 154 pages, 1982.

Vol. 12: M. Jacobsen, Statistical Analysis of Counting Processes. VII, 226 pages, 1982.

Vol. 13: J. Pfanzagl (with the assistance of W. Wefelmeyer), Contributions to a General Asymptotic Statistical Theory. VII, 315 pages, 1982.

Vol. 14: GLIM 82: Proceedings of the International Conference on Generalised Linear Models. Edited by R. Gilchrist. V, 188 pages, 1982.

Vol. 15: K.R.W. Brewer and M. Hanif, Sampling with Unequal Probabilities. IX, 164 pages, 1983.

Vol. 16: Specifying Statistical Models: From Parametric to Non-Parametric, Using Bayesian or Non-Bayesian Approaches. Edited by J.P. Florens, M. Mouchart, J.P. Raoult, L. Simar, and A.F.M. Smith, XI, 204 pages, 1983.

Vol. 17: I.V. Basawa and D.J. Scott, Asymptotic Optimal Inference for Non-Ergodic Models. IX, 170 pages, 1983.

Vol. 18: W. Britton, Conjugate Duality and the Exponential Fourier Spectrum. V, 226 pages, 1983.

Vol. 19: L. Fernholz, von Mises Calculus For Statistical Functionals. VIII, 124 pages, 1983.

Vol. 20: Mathematical Learning Models — Theory and Algorithms: Proceedings of a Conference. Edited by U. Herkenrath, D. Kalin, W. Vogel. XIV, 226 pages, 1983.

Vol. 21: H. Tong, Threshold Models in Non-linear Time Series Analysis. X, 323 pages, 1983.

Vol. 22: S. Johansen, Functional Relations, Random Coefficients and Nonlinear Regression with Application to Kinetic Data, VIII, 126 pages, 1984.

Vol. 23: D.G. Saphire, Estimation of Victimization Prevalence Using Data from the National Crime Survey. V, 165 pages, 1984.

Vol. 24: T.S. Rao, M.M. Gabr, An Introduction to Bispectral Analysis and Bilinear Time Series Models. VIII, 280 pages, 1984.

Vol. 25: Time Series Analysis of Irregularly Observed Data. Proceedings, 1983. Edited by E. Parzen. VII, 363 pages, 1984.

Vol. 26: Robust and Nonlinear Time Series Analysis. Proceedings, 1983. Edited by J. Franke, W. Härdle and D. Martin. IX, 286 pages, 1984.

Vol. 27: A. Janssen, H. Milbrodt, H. Strasser, Infinitely Divisible Statistical Experiments. VI, 163 pages, 1985.

Vol. 28: S. Amari, Differential-Geometrical Methods in Statistics. V, 290 pages, 1985.

Vol. 29: Statistics in Ornithology. Edited by B.J.T. Morgan and P.M. North. XXV, 418 pages, 1985.

Vol 30: J. Grandell, Stochastic Models of Air Pollutant Concentration. V, 110 pages, 1985.

Vol. 31: J. Pfanzagl, Asymptotic Expansions for General Statistical Models. VII, 505 pages, 1985.

Vol. 32: Generalized Linear Models. Proceedings, 1985. Edited by R. Gilchrist, B. Francis and J. Whittaker. VI, 178 pages, 1985.

Vol. 33: M. Csörgo, S. Csörgo, L. Horváth, An Asymptotic Theory for Empirical Reliability and Concentration Processes. V, 171 pages, 1986.

Vol. 34: D.E. Critchlow, Metric Methods for Analyzing Partially Ranked Data. X, 216 pages, 1985.

Vol. 35: Linear Statistical Inference. Proceedings, 1984. Edited by T. Calinski and W. Klonecki. VI, 318 pages, 1985.

Vol. 36: B. Matérn, Spatial Variation. Second Edition. 151 pages, 1986.

Vol. 37: Advances in Order Restricted Statistical Inference. Proceedings, 1985. Edited by R. Dykstra, T. Robertson and F.T. Wright. VIII, 295 pages, 1986.

Vol. 38: Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits. Edited by R.W. Pearson and R.F. Boruch. V, 129 pages, 1986.

Vol. 39: J.D. Malley, Optimal Unbiased Estimation of Variance Components. IX, 146 pages, 1986.

Vol. 40: H.R. Lerche, Boundary Crossing of Brownian Motion. V, 142 pages, 1986.

Vol. 41: F. Baccelli, P. Brémaud, Palm Probabilities and Stationary Queues. VII, 106 pages, 1987.

Vol. 42: S. Kullback, J.C. Keegel, J.H. Kullback, Topics in Statistical Information Theory. IX, 158 pages, 1987.

Vol. 43: B.C. Arnold, Majorization and the Lorenz Order: A Brief Introduction. VI, 122 pages, 1987.

Vol. 44: D.L. McLeish, Christopher G. Small, The Theory and Applications of Statistical Inference Functions. VI, 124 pages, 1987.

Vol. 45: J.K. Ghosh (Ed.), Statistical Information and Likelihood. 384 pages, 1988.

Vol. 46: H.-G. Müller, Nonparametric Regression Analysis of Longitudinal Data. VI, 199 pages, 1988.

Vol. 47: A.J. Getson, F.C. Hsuan, {2}-Inverses and Their Statistical Application. VIII, 110 pages, 1988.

Vol. 48: G.L. Bretthorst, Bayesian Spectrum Analysis and Parameter Estimation. XII, 209 pages, 1988.

Vol. 49: S.L. Lauritzen, Extremal Families and Systems of Sufficient Statistics. XV, 268 pages, 1988.

Vol. 50: O.E. Barndorff-Nielsen, Parametric Statistical Models and Likelihood. VII, 276 pages, 1988.

Vol. 51: J. Hüsler, R.-D. Reiss (Eds.), Extreme Value Theory. Proceedings, 1987. X, 279 pages, 1989.

Vol. 52: P.K. Goel, T. Ramalingam, The Matching Methodology: Some Statistical Properties. VIII, 152 pages, 1989.

Vol. 53: B.C. Arnold, N. Balakrishnan, Relations, Bounds and Approximations for Order Statistics. IX, 173 pages, 1989.

Vol. 54: K.R. Shah, B.K. Sinha, Theory of Optimal Designs. VIII, 171 pages, 1989.

Vol. 55: L. McDonald, B. Manly, J. Lockwood, J. Logan (Eds.), Estimation and Analysis of Insect Populations. Proceedings, 1988. XIV, 492 pages, 1989.

Vol. 56: J.K. Lindsey, The Analysis of Categorical Data Using GLIM. V, 168 pages, 1989.

Vol. 57: A. Decarli, B.J. Francis, R. Gilchrist, G.U.H. Seeber (Eds.), Statistical Modelling. Proceedings, 1989. IX, 343 pages, 1989.

Vol. 58: O.E. Barndorff-Nielsen, P. Blæsild, P.S. Eriksen, Decomposition and Invariance of Measures, and Statistical Transformation Models. V, 147 pages, 1989.

Vol. 59: S. Gupta, R. Mukerjee, A Calculus for Factorial Arrangements. VI, 126 pages, 1989.

Vol. 60: L. Györfi, W. Härdle, P. Sarda, Ph. Vieu, Nonparametric Curve Estimation from Time Series. VIII, 153 pages, 1989.

Vol. 61: J. Breckling, The Analysis of Directional Time Series: Applications to Wind Speed and Direction. VIII, 238 pages, 1989.

Vol. 62: J.C. Akkerboom, Testing Problems with Linear or Angular Inequality Constraints. XII, 291 pages, 1990.

Vol. 63: J. Pfanzagl, Estimation in Semiparametric Models: Some Recent Developments. III, 112 pages, 1990.

Vol. 64: S. Gabler, Minimax Solutions in Sampling from Finite Populations. V, 132 pages, 1990.

Vol. 65: A. Janssen, D.M. Mason, Non-Standard Rank Tests. VI, 252 pages, 1990.

Vol. 66: T. Wright, Exact Confidence Bounds when Sampling from Small Finite Universes. XVI, 431 pages, 1991.

Vol. 67: M.A. Tanner, Tools for Statistical Inference: Observed Data and Data Augmentation Methods. VI, 110 pages, 1991.

Vol. 68: M. Taniguchi, Higher Order Asymptotic Theory for Time Series Analysis. VIII, 160 pages, 1991.

Vol. 69: N.J.D. Nagelkerke, Maximum Likelihood Estimation of Functional Relationships. V, 110 pages, 1992.

Vol. 70: K. Iida, Studies on the Optimal Search Plan. VIII, 130 pages, 1992.

Vol. 71: E.M.R.A. Engel, A Road to Randomness in Physical Systems. IX, 155 pages, 1992.

Vol. 72: J.K. Lindsey, The Analysis of Stochastic Processes using GLIM. VI, 294 pages, 1992.

Vol. 73: B.C. Arnold, E. Castillo, J.-M. Sarabia, Conditionally Specified Distributions. XIII, 151 pages, 1992.

Vol. 74: P. Barone, A. Frigessi, M. Piccioni, Stochastic Models, Statistical Methods, and Algorithms in Image Analysis. VI, 258 pages, 1992.

Vol. 75: P.K. Goel, N.S. Iyengar (Eds.), Bayesian Analysis in Statistics and Econometrics. XI, 410 pages, 1992.

Vol. 76: L. Bondesson, Generalized Gamma Convolutions and Related Classes of Distributions and Densities. VIII, 173 pages, 1992.

Vol. 77: E. Mammen, When Does Bootstrap Work? Asymptotic Results and Simulations. VI, 196 pages, 1992.

Vol. 78: L. Fahrmeir, B. Francis, R. Gilchrist, G. Tutz (Eds.), Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13-17 July 1992. IX, 225 pages, 1992.

Vol. 79: N. Schmitz, Optimal Sequentially Planned Decision Procedures. XII, 209 pages, 1992.

Vol. 80: M. Fligner, J. Verducci (Eds.), Probability Models and Statistical Analyses for Ranking Data. XXII, 306 pages, 1992.

General Remarks

Lecture Notes are printed by photo-offset from the master-copy delivered in camera-ready form by the authors of monographs, resp. editors of proceedings volumes. For this purpose Springer-Verlag provides technical instructions for the preparation of manuscripts. Volume editors are requested to distribute these to all contributing authors of proceedings volumes. Some homogeneity in the presentation of the contributions in a multi-author volume is desirable.

Careful preparation of manuscripts will help keep production time short and ensure a satisfactory appearance of the finished book. The actual production of a Lecture Notes volume normally takes approximately 8 weeks.

For monograph manuscripts typed or typeset according to our instructions, Springer-Verlag can, if necessary, contribute towards the preparation costs at a fixed rate.

Authors of monographs receive 50 free copies of their book. Editors of proceedings volumes similarly receive 50 copies of the book and are responsible for redistributing these to authors etc. at their discretion. No reprints of individual contributions can be supplied. No royalty is paid on Lecture Notes volumes.

Volume authors and editors are entitled to purchase further copies of their book for their personal use at a discount of 33.3% and other Springer mathematics books at a discount of 20% directly from Springer-Verlag. Authors contributing to proceedings volumes may purchase the volume in which their article appears at a discount of 20%.

Springer-Verlag secures the copyright for each volume.