Notes on Statistical Models for Ranking Data

John I. Marden
Department of Statistics
University of Illinois at Urbana-Champaign

Typeset using the *memoir* package (Madsen and Wilson, 2015) with LaTeX (Lamport, 1994).

Preface

Although I had done some work in ranking earlier, the conference *Probability Models and Statistical Analyses for Ranking Data* (Fligner and Verducci, 1993) held in 1990 at the University of Massachusetts, Amherst, opened my eyes to a wide array of modeling and analytic techniques. Subsequently, I prepared these notes around 1993 (in troff) for a course on ranking data taught at the University of Illinois, Urbana-Champaign. They provide an overview of ranking models presented at the conference, and in the literature, at that time. This current version (2017) is cleaned up a bit, and written in LaTeX, but the content is still the same.

The book *Analyzing and Modeling Rank Data* (Marden, 1996) greatly expands on these notes, adding various data analytic and graphical techniques, and providing details and examples for the models. Twenty years have passed, and fortunately Alvo and Yu (2014) have written *Statistical Methods for Ranking Data*, which brings the field up to date. It is especially strong in its treatment of Thurstonian (probit) models, using Markov chain Monte Carlo techniques for model fitting.

The work mentioned above is most appropriate for relatively small data sets, with a relatively small number of objects to rank. (How small depends on the model and/or technique.) There is currently important work being done on Big Data with a large number of objects (such as all the web sites, or movies, in the world) and millions of judges, entailing all manner of partial and incomplete rankings. Maybe someone will write a nice research monograph in this area soon.

Contents

Chapter 1

Models for Rank Data

1.1 Introduction

This chapter reviews a number of approaches to the statistical modeling of ranking data, dealing exclusively with complete rankings. Chapter 2 reviews some methods for categorizing various models. Chapter 3 reviews likelihood and other methods for inference in ranking models. Finally, Chapter 4 considers some extensions of the models to cases where there are ties or other types of incomplete rankings.

1.2 Rankings vs. orderings

The canonical experiment has m objects to be ranked by n judges. Typically, each judge gives either a **ranking** of the objects, assigning "1" to the favorite object, "2" to the second favorite, $\ldots$, and "m" to the least favorite; or an **ordering** of the objects, listing the objects in order from most favorite to least favorite. These two outcomes are equivalent, but some models are defined on rankings and some on orderings, so we have to be able to deal with either.

We will represent one judge's ranking by a pair of vectors of length m, $\boldsymbol{x}$ and $\boldsymbol{y}$. The $\boldsymbol{x}$ contains the **labels** of the objects. The labels may be numbers, letters, names, or anything else. Denote the set of labels by $\mathcal{O} = \{l_1, \ldots, l_m\}$. The $\boldsymbol{y}$ contains the rankings of the objects. The rankings must be integers from 1 to m, and each integer must appear in

$\boldsymbol{x}_i$'s equal The rankings				$\boldsymbol{y}_i$'s equal The orderings			
$\boldsymbol{x}$	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{y}_3$	$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$
Red	3	5	4	1	Green	Blue	Green
Yellow	6	4	2	2	Blue	Green	Yellow
Blue	2	1	3	3	Red	Orange	Blue
Orange	4	3	5	4	Orange	Yellow	Red
Green	1	2	1	5	Purple	Red	Orange
Purple	5	6	6	6	Yellow	Purple	Purple

Table 1.1: Comparison of orderings and rankings.

the vector exactly once. That is,

$$\boldsymbol{x} \in \mathcal{L}_m \equiv \{\text{All permutations of } \mathcal{O}\}$$

and

$$\boldsymbol{y} \in \mathcal{P}_m \equiv \{\text{All permutations of } \{1, 2, \ldots, m\}\}.$$

Thus there are $m!$ elements in both $\mathcal{L}_m$ and $\mathcal{P}_m$. The interpretation of a particular pair $(\boldsymbol{x}, \boldsymbol{y})$ is

$$\text{Object } x_i \text{ is ranked number } y_i \text{ among the } m \text{ objects.}$$

Thus if $\boldsymbol{x}$ = (Red, Blue, Yellow) and $\boldsymbol{y} = (2, 1, 3)$, $(\boldsymbol{x}, \boldsymbol{y})$ corresponds to having Blue ranked first, Red second and Yellow third. Note that $\boldsymbol{x}$ = (Blue, Yellow, Red) and $\boldsymbol{y} = (1, 3, 2)$ correspond to the exact same ranking, so that it is very important to know both $\boldsymbol{x}$ and $\boldsymbol{y}$.

Now consider the set of n rankings, one from each judge, so that the data are

$$(\boldsymbol{x}_1, \boldsymbol{y}_1),\ (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots,\ (\boldsymbol{x}_n, \boldsymbol{y}_n).$$

In order to present the data and to fit models to the data, it is convenient to have either all $\boldsymbol{x}_i$'s equal, or all $\boldsymbol{y}_i$'s equal. Table 1.1 illustrates the two methods.

In the first set, the objects are in a fixed order $\boldsymbol{x}_i$ = (Red, Yellow, Blue, Orange, Green, Purple), and the $\boldsymbol{y}_i$'s are the rankings of the colors in that order. In the second set, the rankings $\boldsymbol{y}_i$ are all in the regular order (1, 2, . . . , 6), and the $\boldsymbol{x}_i$'s are the orderings of the colors from favorite to least favorite.

1.3 Statistical models

Now let $(\boldsymbol{X}, \boldsymbol{Y})$ be a pair of random vectors. The joint sample space of the pair is $\mathcal{L}_m \times \mathcal{P}_m$. A probability distribution on $\mathcal{L}_m \times \mathcal{P}_m$ can be given by its density f, where f is a function of $(\boldsymbol{x}, \boldsymbol{y})$ such that

$$f(\boldsymbol{x}, \boldsymbol{y}) \geqslant 0 \text{ for all } (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{L}_m \times \mathcal{P}_m \text{ and } \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{L}_m \times \mathcal{P}_m} f(\boldsymbol{x}, \boldsymbol{y}) = 1.$$

(Here we are dealing with just discrete distributions, in fact, distributions on finite sample spaces.) Then $P[(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{x}, \boldsymbol{y})] = f(\boldsymbol{x}, \boldsymbol{y})$. A **statistical model** is a family of such probability distributions. Typically, the family can be parametrized by a small number of parameters, say $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$. Letting Θ be the range of $\boldsymbol{\theta}$, the **parameter space**, the model is written

$$\{f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{\theta} \in \Theta\}. \tag{1.3.1}$$

In almost all cases, either $\boldsymbol{x}$ or $\boldsymbol{y}$ is fixed as in the previous section. Thus typically the densities will be either

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \mathcal{L}_m \text{ or } f_{\boldsymbol{\theta}}(\boldsymbol{y}) \text{ for } \boldsymbol{y} \in \mathcal{P}_m.$$

For explicative reasons, densities will often be given in the form

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{c(\boldsymbol{\theta})}\, g_{\boldsymbol{\theta}}(\boldsymbol{y}),$$

where $g_{\boldsymbol{\theta}}$ is given explicitly, but $c(\boldsymbol{\theta})$ is not. In such cases, it is to be understood that $c(\boldsymbol{\theta})$ is whatever it must be so that the density sums to 1.

The following sections contain brief introductions to the main types of probability models used for rankings. Critchlow, Fligner, and Verducci (1991) contains a similar development. For each one, whether the $\boldsymbol{x}$ or the $\boldsymbol{y}$ is fixed must be specified. There are two extreme models: the **uniform** and the **saturated**. When $\boldsymbol{x}$ is fixed, the uniform model has no parameters, and each $\boldsymbol{y}$ is equally likely:

$$f(\boldsymbol{y}) = \frac{1}{m!} \text{ for every } \boldsymbol{y} \in \mathcal{P}_m\,; \text{ write } \boldsymbol{Y} \sim \text{Uniform}(\mathcal{P}_m). \tag{1.3.2}$$

Similarly, when $\boldsymbol{y}$ is fixed, we can define $\boldsymbol{X} \sim \text{Uniform}(\mathcal{L}_m)$. When $\boldsymbol{x}$ is fixed, the saturated model has $m!$ parameters ($m! - 1$ free ones) $\{p_{\boldsymbol{y}} \mid \boldsymbol{y} \in \mathcal{P}_m\}$, where $p_{\boldsymbol{y}} = P[\boldsymbol{Y} = \boldsymbol{y}]$. The parameter space is the simplex in $\mathbb{R}^{m!}$, that is, all sets of nonnegative values which sum to 1. The saturated model for $\boldsymbol{y}$ fixed is similar.

These extreme models are good to use as benchmarks. The uniform is the null model; one often first tests whether to reject it before trying to look for structure in the data. The uniform is a submodel of almost all models. On the other hand, every model is contained in the saturated model. Thus one can test for goodness-of-fit of a particular model by comparing it to the saturated model. The hope is to find a model which is as simple as possible but still fits well.

1.4 Thurstonian (≡ order statistic) models

Here, $\boldsymbol{y}$ is fixed to be $(1, 2, \ldots, m)$, so that $\boldsymbol{x}$ represents the ordering of the objects. Each object has a continuous but unobserved random variable associated with it. Call these variables $Z_{l_1}, Z_{l_2}, \ldots, Z_{l_m}$, where Z_{l_i} is that associated with the object l_i. A joint model is assumed for the vector $\boldsymbol{Z}$ of the Z_l's. The observed ordering of the objects is then given by the order of the Z_l's, that is,

$$\boldsymbol{x} = (l_{i_1}, l_{i_2}, \ldots, l_{i_m}) \text{ if and only if } Z_{l_{i_1}} < Z_{l_{i_2}} < \cdots < Z_{l_{i_m}}. \tag{1.4.1}$$

For example, if $\mathcal{O}$ = {Red, Yellow, Green}, then $\boldsymbol{x}$ = (Yellow, Green, Red) if and only if $Z_{\text{Yellow}} < Z_{\text{Green}} < Z_{\text{Red}}$. Thus the probability of any ranking of the objects is the probability of the corresponding order of the Z_l's.

The parameters for the model are those needed to describe the distribution of the vector $\boldsymbol{Z}$. Thurstone (1927), who invented the model, proposed using the normal distribution, hence the parameters included the m means, m variances, and $\binom{m}{2}(= \frac{m!}{k!(m-k)!})$ correlations. He also suggested simplifications such as equating the correlations, equating the variances, and/or setting the correlations to 0 so that the Z_{l_i}'s are independent. See Böckenholt (1993).

Daniels (1950) continued by looking at cases in which the Z's are independent and from a location-family model with possibly different location parameters. That is, for some continuous density g, the density of $\boldsymbol{Z}$ is

$$\prod_{i=1}^{m} g(z_{l_i} - \mu_{l_i}). \tag{1.4.2}$$

Yellott (1977) considered the Gumbel distribution $g(z) = e^{-e^{-z}}$ (see the Luce model in Section 1.6.1), and Henery (1983) and Stern (1987) used gamma distributions.

1.5 Paired comparison models

1.5.1 Babington Smith

An alternative to asking someone to rank the m objects is to have them choose which of each pair of objects is preferred. Thus instead of giving the ordering (Yellow, Red, Green), one would say "Yellow is preferred to Red," "Red is preferred to Green" and "Yellow is preferred to Green." For m objects there are $\binom{m}{2}$ such comparisons to make. It should be clear that given a ranking, one can easily determine what the pairwise preferences are. However, it is not always the case that a set of pairwise preferences corresponds to a ranking. For example, one might prefer Yellow to Red, Red to Green, and Green to Yellow. There are a number of papers which look at paired comparison data in which such nontransitivities are allowed. Kendall and Babington Smith (1940) is perhaps the seminal one. In contrast, we are taking the view that people construct a legitimate ranking by starting with paired comparisons, but report their preferences only after having a consistent set of comparisons, i.e., a set which yields an unambiguous ranking. The resulting ranking model is referred to as the **Babington Smith model**, because allegedly it appears in Babington Smith (1950).

First fix the vector $\boldsymbol{y} = (1, 2, \ldots, m)$, so that x_i is the label of the object ranked i^{th}. The general model is based on $\binom{m}{2}$ parameters $\boldsymbol{p} = \{p_{l_i l_j} \mid 1 \leqslant i < j \leqslant m\}$, where $p_{l_i l_j}$ is interpreted as the probability object l_i would be preferred to object l_j if only that comparison were to be made. Ties are not allowed, so that $p_{l_j l_i} = 1 - p_{l_i l_j}$. A ranking is obtained by making independently all the pairwise comparisons using those probabilities. If the comparisons yield a consistent ranking, that ranking is the $\boldsymbol{x}$. If not, start over with the pairwise comparisons. One repeats until the comparisons are consistent.

The probability that the paired comparisons are consistent is

$$c(\boldsymbol{p}) = \sum_{\boldsymbol{x} \in \mathcal{L}_m} \prod_{i<j} p_{x_i x_j}. \tag{1.5.1}$$

For $m = 3$, and $\mathcal{O} = \{A, B, C\}$, $\mathcal{L}_m$ = {ABC, ACB, BAC, BCA, CAB, CBA}, so that

$$c(\boldsymbol{p}) = p_{AB}p_{AC}p_{BC} + p_{AC}p_{AB}p_{CB} + p_{BA}p_{BC}p_{AC} \\ + p_{BC}p_{BA}p_{CA} + p_{CA}p_{CB}p_{AB} + p_{CB}p_{CA}p_{BA}.$$

Now the probability of an ordering $\boldsymbol{x}$ given that the pairwise comparisons are consistent is the probability that the comparisons yield $\boldsymbol{x}$ divided by the probability they are consistent. Thus the Babington Smith model is

$$f_{\boldsymbol{p}}(\boldsymbol{x}) = \frac{1}{c(\boldsymbol{p})} \prod_{i<j} p_{x_i x_j}. \tag{1.5.2}$$

Note that it is not necessary that the $p_{l_i l_j}$'s be consistent. That is, it is odd but legitimate to have $p_{l_1 l_2} = 0.9$ and $p_{l_2 l_3} = 0.95$ but $p_{l_1 l_3} = 0.1$. See (2.5.1) and (2.5.2) for transitivity constraints on the $p_{l_i l_j}$'s that avoid such behavior.

This model can be unwieldy if m is at all large because there are many parameters, and the constant $c(\boldsymbol{p})$ is not very compact: It consists of a sum of $m!$ products of $\binom{m}{2}$ terms. The next section considers some simplifications.

Warning. Recall that $p_{l_i l_j}$ is the probability of object l_i being preferred to object l_j if that comparison is the whole experiment. However, the probability that l_i is preferred to l_j after the entire ranking experiment has been performed is **not** $p_{l_i l_j}$.

1.5.2 Bradley-Terry-Mallows

Here, the parameters $p_{l_i l_j}$ are given special forms. Bradley and Terry (1952) proposed positive constants

$$\boldsymbol{\nu} = (\nu_{l_1}, \ldots, \nu_{l_m}),$$

where ν_{l_i} is associated with object l_i, setting

$$p_{l_i l_j} = \frac{\nu_{l_i}}{\nu_{l_i} + \nu_{l_j}}.$$

The idea is that the larger ν_{l_i}, the more preferred object l_i is. Bradley and Terry were thinking of just paired comparisons, but Mallows (1957) suggested substituting the Bradley-Terry form of the $p_{l_i l_j}$'s into the Babington Smith model for ranks. Note that for an ordering $\boldsymbol{x} \in \mathcal{L}_m$,

$$\prod_{i<j} p_{x_i x_j} = \frac{\prod_{i=1}^{m} \nu_{x_i}^{m-i}}{\prod_{i<j} (\nu_{l_i} + \nu_{l_j})},$$

since "ν_{x_i}" appears in the numerator the same number of times as object x_i is preferred to the other objects. Thus the model is

$$f_{\boldsymbol{\nu}}(\boldsymbol{x}) = \frac{1}{c(\boldsymbol{\nu})} \prod_{i=1}^{m} \nu_{x_i}^{m-i}. \tag{1.5.3}$$

This model now has only m parameters; in fact, since the probabilities are invariant to multiplying the ν_i's by a positive constant, there are only $m-1$ free parameters.

Mallows also simplified the model by assuming other special forms for the $p_{l_i l_j}$'s. His two-parameter model assumes that the objects are indexed in a meaningful way, e.g., l_1 is the most popular, l_2 is second most popular, ..., l_m is least popular. The model sets

$$p_{l_i l_j} = \frac{1 + \tanh((j - i)\log(\theta) + \log(\phi))}{2} \quad \text{for } i < j.$$

We will just look at the two one-parameter subfamilies obtained by setting either θ or ϕ to 1.

1.5.3 Mallows' ϕ model

This model is perhaps the most famous of all. One supposes that the paired-comparison probability of ranking object l_i before object l_j depends only on whether $i < j$ or $j > i$. Thus the $p_{l_i l_j}$'s for $i < j$ are all equal. It is usual to parametrize this probability as

$$p_{l_i l_j} = \frac{e^{\gamma I[i>j]}}{1 + e^{\gamma}}, \tag{1.5.4}$$

where $I[S]$ for an event S is the indicator of that event, that is, it equals 1 if the event occurs, 0 if it does not. If $\gamma < 0$, then objects with lower indices will tend to be ranked higher. Placing these (1.5.4) in the Babington Smith model (1.5.2) yields

$$f_{\gamma}(\boldsymbol{x}) = \frac{1}{c(\gamma)}\, e^{\gamma d_K(\boldsymbol{x})}, \tag{1.5.5}$$

where

$$d_K(\boldsymbol{x}) = \sum_{i<j} I[\text{index}(x_i) > \text{index}(x_j)],$$

and "index" gives the index of the object, i.e., $x_i = l_{\text{index}(x_i)}$. This $d_K(\boldsymbol{x})$ is the number of times a later object is preferred to an earlier object, and in fact is related to Kendall's τ distance in (1.8.3).

For example, if $\boldsymbol{x} = (l_3, l_2, l_5, l_1, l_4)$, then $d(\boldsymbol{x}) = 2 + 1 + 2 + 0 = 5$. That is, for $i = 1$, since $x_1 = l_3$, $\text{index}(x_1) = 3$. Then we check the indices of the l_j's after the first, and we see 2 of them less than 3 (obviously). Next we look at $i = 2$, seeing x_2's index is 2. There is just 1 index lower than 2 among those to the right of x_2. Then there are 2 to the right of x_3 lower than the index 5, and 0 to the right of x_1 lower than the index 1.

It turns out that the constant $c(\gamma)$ is reasonably tractable, an unusual situation so far. See Section 3.7.

1.5.4 Mallows' θ model

This model is a Bradley-Terry model with the special form for the ν_i's being

$$\nu_{l_i}(\gamma) = e^{\gamma i},$$

hence

$$p_{l_i l_j} = \frac{e^{\gamma i}}{e^{\gamma i} + e^{\gamma j}}. \tag{1.5.6}$$

Inserting these (1.5.6) into the Babington Smith model (1.5.2) gives

$$g_\gamma(\boldsymbol{x}) = \frac{1}{c(\gamma)} e^{\gamma s(\boldsymbol{x})}, \tag{1.5.7}$$

where

$$s(\boldsymbol{x}) = \sum_{i=1}^{m} (m - i) \cdot \text{index}(x_i).$$

Here, s is related to Spearman's ρ distance in (1.8.3). The difference between Mallows' ϕ and θ models is that in the latter $p_{l_i l_j}$ depends on the difference $i - j$, while in the former the $p_{l_i l_j}$ depends only on the sign of $i - j$.

1.6 Multistage models

1.6.1 Luce's choice axiom

Luce (1959) presents an axiom which he deems reasonable when declaring preferences among the objects. Let $\mathcal{T}$ be any subset of the objects' labels $\mathcal{O}$, and a be any of those in $\mathcal{T}$. Then Luce defines

$$P_{\mathcal{T}}(a) = \text{Prob[Object } a \text{ is the most preferred among those in } \mathcal{T}]$$

and for any subset $\mathcal{S} \subset \mathcal{T}$,

$$P_{\mathcal{T}}(\mathcal{S}) = \sum_{a \in \mathcal{S}} P_{\mathcal{T}}(a),$$

which is the probability the most preferred among $\mathcal{T}$ is one of those in S. The following axiom concerns the subset $\mathcal{T} \subset \mathcal{O}$.

Axiom 1.6.1. *Luce's choice axiom*

(i) If $P_{\{a,b\}}(a) \neq 0$ *for all* $a, b \in \mathcal{T}$, *then for* $r \in \mathcal{S} \subset \mathcal{T}$,

$$P_{\mathcal{T}}(r) = P_{\mathcal{S}}(r) P_{\mathcal{T}}(\mathcal{S});$$

(ii) If $P_{\{a,b\}}(a) = 0$ *for some* $a, b \in \mathcal{T}$*, then if* $r \in \mathcal{T}$*,* $r \neq a$*,*

$$P_{\mathcal{T}}(r) = P_{\mathcal{T}-\{a\}}(r).$$

The condition $P_{\{a,b\}}(a) \neq 0$ means it is possible to prefer a to b. Conversely, $P_{\{a,b\}}(a) = 0$ means b is always preferred to a. Thus condition (i) says that as long as any pairwise preference is possible among the objects in $\mathcal{T}$, the probability that r is the favorite is the same as the probability that the favorite is in the set $\mathcal{S}$ times the probability that r is the favorite in $\mathcal{S}$. Part (ii) says that if b is always preferred to a, then the probability r is the favorite does not change when a is thrown out.

For example, suppose $\mathcal{O} =$ {Coke, Pepsi, 7-up, Sprite}. The axiom applied to $\mathcal{T} = \mathcal{O}$ implies that if any pairwise preference is possible, then in particular

$$\begin{aligned} P[\text{Coke is the favorite among all four}] &= P[\text{Coke is preferred to Pepsi}] \\ &\quad \times P[\text{A cola is chosen as the favorite among all four}]. \end{aligned}$$

Thus the choosing of the favorite can be decomposed into a two-stage process. If 7-up is always preferred to Sprite, then part (ii) of the axiom implies that

$$\begin{aligned} &P[\text{Coke is the favorite among all four}] \\ &\qquad = P[\text{Coke is the favorite among \{Coke, Pepsi, 7-up\}}]. \end{aligned}$$

Now suppose we have a ranking model for $\mathcal{O}$ such that for all $a, b \in \mathcal{O}$, $P_{\{a,b\}}(a) \neq 0$. It is called **L-decomposible** if it satisfies (part (i) of) the choice axiom for all subsets $\mathcal{T} \subset \mathcal{O}$. One consequence is that, with $\boldsymbol{y} = (1, 2, \ldots, m)$,

$$P[\boldsymbol{X} = \boldsymbol{x}] = P_{\mathcal{O}}(x_1) \times P_{\{x_2,\ldots,x_m\}}(x_2) \times \cdots \times P_{\{x_{m-1},x_m\}}(x_{m-1}). \quad (1.6.1)$$

That is, the probability of ordering $\boldsymbol{x}$ is the probability x_1 is the favorite of all, times the probability x_2 is the favorite of all but x_1, times the probability that x_3 is the favorite of all but x_1 and x_2, etc.

Another consequence is the property **independence from irrelevant alternatives**: The relative ranking of the objects in some subset $\mathcal{T} \subset \mathcal{O}$ is independent of the relative ranking of the objects in $\mathcal{O} - \mathcal{T}$. In the example above, this means that the probability that Coke is preferred to Pepsi does not depend on whether 7-up is preferred to Sprite.

Finally, the axiom implies a form of the model which looks reminiscent of the Bradley-Terry model. There exist positive constants ν_i such that if $l_i \in \mathcal{S} \subset \mathcal{O}$,

$$P_{\mathcal{S}}(l_i) = \frac{\nu_{l_i}}{\sum_{l_j \in \mathcal{S}} \nu_{l_j}}. \quad (1.6.2)$$

Without loss of generality we assume that the ν_{l_i}'s sum to one. Then $P[X_1 = l_i] = \nu_{l_i}$, that is, ν_{l_i} is the probability that object l_i is ranked #1. Putting (1.6.1) and (1.6.2) together, we have the $(m-1)$-parameter model

$$f_{\boldsymbol{\nu}}(\boldsymbol{x}) = \frac{\nu_{x_1}}{1} \times \frac{\nu_{x_2}}{\nu_{x_2} + \cdots + \nu_{x_m}} \times \cdots \frac{\nu_{x_{m-1}}}{\nu_{x_{m-1}} + \nu_{x_m}}. \qquad (1.6.3)$$

Plackett (1975) presents this model as a model for an m-horse race. Thus $\nu_{\text{Sea Biscuit}}$ is the probability that Sea Biscuit wins when all m horses race, $\nu_{\text{Mr. Ed}}/(1-\nu_{\text{Sea Biscuit}})$ is the probability that Mr. Ed wins when all but Sea Biscuit race, etc. Silverberg (1980) calls this a Vase model. The idea is that one has a vase with an infinite number of balls, a proportion of ν_{Red} are Red, ν_{Blue} are Blue, etc. The ordering $\boldsymbol{x}$ is generated by first randomly drawing a ball, and assigning its color to x_1. Another ball is drawn. If it is a different color, that color is assigned to x_2. If it is the same as x_1, it is thrown away and another ball is drawn, continuing this process until a new color is drawn for x_2. More draws are made until a color distinct from x_1 and x_2 is found, which is assigned to x_3, etc.

1.6.2 Free and ϕ component models

Fligner and Verducci (1986; 1988) define multistage ordering probabilities that do not satisfy Luce's choice axiom but do result in a plausible model. That is, (1.6.1) holds but (1.6.2) does not. The ranking vector $\boldsymbol{y}$ is fixed at $(1, 2, \ldots, m)$, and the ordering $\boldsymbol{x}$ proceeds by operating iteratively on the vector $(l_1, l_2, \ldots, l_m)$ as follows. First, choose which object will be ranked #1, assign it to x_1, and move it to the first slot in the vector, keeping the other objects in the same relative order. Let U_1 be the number of places the object moved. The second ranked object is chosen from the remaining. Assign it to x_2, move it to the second slot, and keep the other $m-2$ objects in the same order. Now U_2 is the number of places x_2 moved. Continue until there are two objects left. Decide which is preferred, and place it in the $(m-1)^{st}$ slot. Then U_{m-1} is the number of places it moved, hence is either 0 or 1.

It is easier to see an example. Suppose $\mathcal{O} = \{A, B, C, D, E\}$, and the resulting ordering is $\boldsymbol{x} = (B, D, E, C, A)$. Table 1.2 exhibits the stages in transforming ABCDE to BDECA. First, B is moved up one slot; then D is moved up 2; E is moved up 2; and finally C is moved up 1.

Of course, these U_i's can be defined for any model. However, the crux of the Fligner and Verducci model is that the distribution for $\boldsymbol{X}$ is that induced from one on $\boldsymbol{U} \equiv (U_1, U_2, \ldots, U_{m-1})$ in which the U_i's are independent. The range of U_i is $\{0, 1, \ldots, m-i\}$, so that the space of $\boldsymbol{U}$

	Stage 1	Stage 2	Stage 3	Stage 4
A	**B**	B	B	B
B	A	**D**	D	D
C	C	A	**E**	E
D	D	C	A	**C**
E	E	E	C	A
	$U_1 = 1$	$U_2 = 2$	$U_3 = 2$	$U_4 = 1$

Table 1.2: Steps in a component model

is

$$\mathcal{U}_m \equiv \{0, 1, \ldots, m-1\} \times \{0, 1, \ldots, m-2\} \times \cdots \times \{0, 1\}.$$

It can be shown that there is a one-to-one correspondence between $\mathcal{U}_m$ and $\mathcal{L}_m$, so that knowing the U_i's is enough to know $\boldsymbol{x}$. The models for $\boldsymbol{U}$ are of the form

$$h(\boldsymbol{u}) = h_1(u_1)h_2(u_2)\cdots h_{m-1}(u_{m-1}), \quad \boldsymbol{u} \in \mathcal{U}_m, \tag{1.6.4}$$

where h_i is a density on $\{0, \ldots, m-i\}$. The **free** model allows the individual h_i's to be unrestricted. Other models restrict the h_i's to some parametric form. A special case is the ϕ component model in which U_i is assumed to have a truncated geometric distribution with parameter θ_i, i.e.,

$$h_{i\theta_i}(u_i) = \frac{1}{c_i(\theta_i)} e^{\theta_i u_i}, \quad u_i = 0, \ldots, m-i,$$

where

$$c_i(\theta_i) = \frac{1 - e^{\theta_i(m-i+1)}}{1 - e^{\theta_i}}. \tag{1.6.5}$$

Then the density of $\boldsymbol{U}$ is

$$h_{\boldsymbol{\theta}}(\boldsymbol{u}) = \frac{1}{c(\boldsymbol{\theta})} e^{\sum_{i=1}^{m-1} \theta_i u_i}. \tag{1.6.6}$$

This model (1.6.6) is called the ϕ component model because when the θ_i's are all equal to γ, the density becomes that of Mallows' ϕ model (1.5.5) since $d_K(\boldsymbol{x}) = \sum u_i$.

The presentation here has assumed that the ranking proceeded by first deciding which object would be ranked #1, then which #2, etc. An alternative approach would be to fix the $\boldsymbol{x} = (l_1, l_2, \ldots, l_m)$, and first decide what the rank of l_1 should be, then that of l_2, etc. Thus one starts with the ranking vector $(1, 2, \ldots, m)$, and moves the **ranks**, one at each stage. The U_i can again be defined as how many places the rank of object l_i was moved, and the models can be applied to these U_i's. These models are mathematically the same as above, but concern rankings rather than orderings.

1.6.3 Orthogonal contrast models

In the models in the previous section each stage involved choosing the one favorite object among those left. More general stages could choose sets of objects, so that, for example, if $\mathcal{O} = \{A, B, C, D, E\}$, at the first stage one might divide the objects into two groups $\{A, C, D\}$ and $\{B, E\}$, the interpretation being that A, C and D are all preferred to B and E. The second stage may choose A as the favorite among $\{A, C, D\}$, the third stage choose C above D, and the fourth choose E above B. The resulting ordering is then ACDEB. A model analogous to the Fligner and Verducci ones would then declare the choices at each stage independent.

Chung and Marden (1991) define the stages in terms of ranks instead of objects. Some definitions are needed.

Definition 1.6.2. Given a set $\mathcal{O}$ of objects, a **contrast** $\boldsymbol{C}$ is an ordered set $(\mathcal{I}, \mathcal{J})$ of two nonempty disjoint subsets $\mathcal{I}$ and $\mathcal{J} \subset \mathcal{O}$.

The idea is that contrast $\boldsymbol{C}$ represents a comparison between the two sets of objects $\mathcal{I}$ and $\mathcal{J}$. For example, suppose

$$\begin{aligned} \mathcal{O} = \{&\text{Coke, Pepsi, 7-up, Sprite,} \\ &\qquad\qquad\text{Diet Coke, Diet Pepsi, Diet 7-up, Diet Sprite}\}. \end{aligned}$$

Then some possible contrasts are

$$\begin{aligned} \boldsymbol{C}_1 &= (\{\text{Coke, Pepsi, 7-up, Sprite}\}, \\ &\qquad\qquad \{\text{Diet Coke, Diet Pepsi, Diet 7-up, Diet Sprite}\}); \\ \boldsymbol{C}_2 &= (\{\text{Coke, Pepsi}\}, \{\text{7-up, Sprite}\}); \\ \boldsymbol{C}_3 &= (\{\text{Coke}\}, \{\text{Pepsi}\}); \\ \boldsymbol{C}_4 &= (\{\text{Coke, Pepsi, Diet Coke, Diet Pepsi}\}, \\ &\qquad\qquad \{\text{7-up, Sprite, Diet 7-up, Diet Sprite}\}). \end{aligned}$$

Thus $\boldsymbol{C}_1$ compares the non-diet to the diet drinks, $\boldsymbol{C}_2$ compares the non-diet colas to the non-diet uncolas, $\boldsymbol{C}_3$ compares Coke and Pepsi, and $\boldsymbol{C}_4$ compares the colas to the uncolas. Marden (1992) extends the definition of contrast to include comparisons of more than two groups of objects.

The value a particular judge has for a particular contrast is defined to be the set of ranks of the objects in $\mathcal{I}$ relative to the objects in $\mathcal{I} \cup \mathcal{J}$. The relative ranks of the objects **within** group $\mathcal{I}$ are irrelevant. Consider contrast $\boldsymbol{C}_2$ above. A judge who likes both colas better than either uncola would have a value of $\{1,2\}$ for the contrast. This value is not meant to give any information about which cola is preferred, i.e., $\{1,2\}$ and $\{2,1\}$ are the same value. A judge who prefers both uncolas to the colas has the value $\{3,4\}$. The value $\{1,3\}$ means Coke and Pepsi are ranked 1 and 3, or 3 and 1, among the four non-diet drinks.

The value a judge has for a contrast can be obtained from the judge's ranking $\boldsymbol{y}$ of the objects in the order $(l_1, l_2, \ldots, l_m)$. The formal definition follows.

Definition 1.6.3. For $\boldsymbol{y} \in \mathcal{P}_m$ and $\boldsymbol{C} = (\mathcal{I}, \mathcal{J})$ a contrast of objects in $\mathcal{O}$, the **value** of the contrast at the ranking is

$$\boldsymbol{C}(\boldsymbol{y}) = \{\overline{y}_i \mid i \in \mathcal{I}\},$$

where $\overline{y}_i$ is the rank of y_i relative to the ranks $\{y_i \mid i \in \mathcal{I} \cup \mathcal{J}\}$.

Continuing the soft drink example, suppose $\boldsymbol{y} = (3,8,5,7,6,2,4,1)$. The values of the four contrast are

Values for $\boldsymbol{y}$ = 38576241

Contrast	$\boldsymbol{C}_1$	$\boldsymbol{C}_2$	$\boldsymbol{C}_3$	$\boldsymbol{C}_4$
Value	$\{3,5,7,8\}$	$\{1,4\}$	$\{1\}$	$\{2,3,6,8\}$

The value for $\boldsymbol{C}_2$ is found by first finding the ranks for the non-diet drinks: 3857. Then these are reranked relative to each other: 1423. These numbers are the $\overline{y}_i$'s. Then since $\mathcal{I} = \{l_1, l_2\}$, the value is $\boldsymbol{C}_2(\boldsymbol{y}) = \{1,4\}$. The value for $\boldsymbol{C}_3$ is $\{1\}$ since of the two colas, Coke is ranked first and Pepsi second.

Knowing the value of a contrast gives only partial information about the entire ranking $\boldsymbol{y}$. However, if you know the values for enough of the contrasts, you should be able to reconstruct the entire ranking. In particular, it is enough to know just the pairwise contrasts $(\{l_i\}, \{l_j\})$ for $i < j$. The minimal number needed is $m-1$. Special sets of contrasts, orthogonal contrasts, are especially efficient. Another definition is needed.

Definition 1.6.4. Two contrasts $\boldsymbol{C} = (\mathcal{I}, \mathcal{J})$ and $\boldsymbol{D} = (\mathcal{K}, \mathcal{L})$ are **orthogonal** if either

$$\begin{gathered}
\text{(i) } (\mathcal{I} \cup \mathcal{J}) \cap (\mathcal{K} \cup \mathcal{L}) = \emptyset; \\
\text{(ii) } (\mathcal{I} \cup \mathcal{J}) \subset \mathcal{K} \;\text{ or }\; (\mathcal{I} \cup \mathcal{J}) \subset \mathcal{L}; \text{ or} \\
\text{(iii) } (\mathcal{K} \cup \mathcal{L}) \subset \mathcal{I} \;\text{ or }\; (\mathcal{K} \cup \mathcal{L}) \subset \mathcal{J}.
\end{gathered}$$

The idea here is that two contrasts are orthogonal if the comparisons they represent are not confounded. Condition (i) means the two contrasts compare totally different sets of objects; (ii) and (iii) state that all the objects in one of the contrasts are contained in the same group of the other contrast. In the example, $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are orthogonal since the first compares non-diets to diets, while the second is a comparison within the non-diets. $\boldsymbol{C}_1$ and $\boldsymbol{C}_4$ are **not** orthogonal since both involve comparison of {Coke, Pepsi} to {Diet 7-up, Sprite}. Other contrast pairs which are orthogonal are $\boldsymbol{C}_1$ and $\boldsymbol{C}_3$, $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$, and $\boldsymbol{C}_3$ and $\boldsymbol{C}_4$; $\boldsymbol{C}_2$ and $\boldsymbol{C}_4$ are not orthogonal.

An orthogonal contrast model depends on a set of q orthogonal contrasts, $(\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_q)$. The following lemma is a justification for the designation "orthogonal."

Lemma 1.6.5. *If* $(\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_q)$ *is a set of orthogonal contrasts and* $\boldsymbol{Y} \sim$ *Uniform*$(\mathcal{P}_m)$ *(see* (1.3.2)*), then* $\boldsymbol{C}_1(\boldsymbol{Y}), \boldsymbol{C}_2(\boldsymbol{Y}), \cdots, \boldsymbol{C}_q(\boldsymbol{Y})$ *are independent, each* $\boldsymbol{C}_i(\boldsymbol{Y})$ *being uniformly distributed over its space.*

The next three subsections give special cases of orthogonal contrast models. These models are nested: $\phi \subset$ Free $\subset$ Contingency table. The Fligner and Verducci (1988) component model on ranks, defined in the last paragraph of Section 1.6.2, uses a special case of orthogonal contrasts: $\boldsymbol{C}_1 = (\{l_1\}, \{l_2, \ldots, l_m\})$, $\boldsymbol{C}_2 = (\{l_2\}, \{l_3, \ldots, l_m\})$, $\ldots$, $\boldsymbol{C}_{m-1} = (\{l_{m-1}\}, \{l_m\})$. The free and ϕ component models are then special cases of the free and ϕ orthogonal contrast models.

The Free model

Start with densities $h_1, \ldots, h_q$, where h_i is a density for contrast C_i. The **free model** states that the contrasts are independent but otherwise unrestricted. Thus

$$f(\boldsymbol{y}) = \prod_{i=1}^{q} h_i(C_i(\boldsymbol{y})). \tag{1.6.7}$$

Compare this to the model (1.6.4). Since contrast $C_i = (\mathcal{I}_i, \mathcal{J}_i)$ has $\binom{\#\mathcal{I}_i + \#\mathcal{J}_i}{\#\mathcal{I}_i}$ possible values, the number of parameters in the model (1.6.7), $\sum\left(\binom{\#\mathcal{I}_i + \#\mathcal{J}_i}{\#\mathcal{I}_i} - 1\right)$, can be quite large. The next model reduces the number of parameters significantly.

The ϕ model

Analogous to (1.6.6), the ***ϕ* model** declares a particular parametric form for each of the h_i's in (1.6.7). First we need a single number for each value of a contrast to measure the degree of preference of the first group over the second. The simplest is to take the sum of the elements of the value. Thus if a contrast has $\boldsymbol{C}(\boldsymbol{y}) = \{3, 5, 7, 8\}$, then the statistic is

$3+5+7+8 = 23$. The lower the number, the more the first group is preferred. We will change it slightly so that the lowest number is 0. Thus if $\boldsymbol{C} = (\mathcal{I}, \mathfrak{J})$, the associated statistic will be

$$d(\boldsymbol{C}(\boldsymbol{y})) \equiv \sum_{i=1}^{\#\mathcal{I}} \overline{y}_i - \binom{\#\mathcal{I}+1}{2}. \tag{1.6.8}$$

Now the ϕ model has a q-parameter vector $\boldsymbol{\theta}$ where

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{c(\boldsymbol{\theta})} e^{\sum_{i=1}^{q} \theta_i d(C_i(\boldsymbol{y}))}. \tag{1.6.9}$$

Those familiar with nonparametrics will recognize d as equivalent to the Mann-Whitney or Wilcoxon statistic for testing the equality of two populations.

Contingency table models

The ϕ model, as the free model, assumes that the contrasts are independent. Such a requirement is not always tenable. For example, the two contrasts ({Coke}, {7-up}) and ({Diet Coke}, {Diet 7-up}) are orthogonal, but it would not be surprising if they were not independent. A contingency table model generalizes the free model by considering the contrasts as the factors in an q-way contingency table. Any well-known contingency table model can be contemplated. The free model is just the independence model. One could also look at the model with all second-order interactions, or all third-order interactions, or one in which the first two contrasts are conditionally independent given the others, or the saturated model (see below (1.3.2)).

1.7 Unfolding models – Non-stochastic

Coombs (1964) has a general theory for preference data which seeks to plot judges and objects in the same space in such a way that the more a judge likes an object, the closer the judge is to the object. One can also look at the inter-judge and inter-object distances. The corresponding one-dimensional unfolding model places the objects and judges on the real line. A particular judge then ranks the objects in the order of their distance from the judge. For example, suppose there are five wines A, B, C, D, and E, all the same except for their oak content, A having the least, E the most. Also, let there be four judges, J1, ..., J4. A possible plotting of the judges and wines is

J1__A____B__J2____________J3__C__D___J4____E_.

Judge 1 does not like oak much. His ordering is ABCDE. Judge 2 and Judge 3 both prefer oak content to be somewhere between those of wines B and C, but Judge 2 is closer to B. Their orderings are, respectively, BACDE and CDEBA. Judge 4 likes oak. Her ordering is DECBA. The word "unfolding" arises since one can obtain the ordering of a particular judge by folding the scale at that judge's point. The objects are then in the correct order.

The non-stochastic version of the unfolding model posits that there is a scale on which the objects and all the judges can be placed which returns every judge's ordering exactly. In typical data sets, such a scale is impossible to find. For example, if Judge 5's order were DCBEA, there would be no way to place the J5 on the scale. Thus some way to incorporate slight deviations from the scale must be incorporated. Distance models seem appropriate. Section 1.9 will consider the stochastic version.

One can also have multidimensional unfolding models, although they are harder to fold. Basically, the objects and judges are placed in p-dimensional space. A judge's ordering is again in order of Euclidean distance from the objects. For example, with $p = 3$ one might have the wines differ in oak content, sweetness and viscosity.

1.8 Sufficient statistic models

The models up until now have a hard-modeling flavor. Each model assumes that ranking is performed by building upon smaller units, whether paired comparisons or stages or distances. Once the probability density is determined, it is possible to determine whether there are any sufficient statistics, that is, whether the information in the data about the parameters can be obtained from some functions of the data, which is especially useful when there is a large number of judges.

Formally, suppose $\boldsymbol{W}$ is a random vector or matrix with associated model $\{f_{\boldsymbol{\theta}}(\boldsymbol{w}) \mid \boldsymbol{\theta} \in \Theta\}$. We will just consider the discrete case. Suppose $S(\boldsymbol{w})$ is a function on the space of $\boldsymbol{W}$.

Definition 1.8.1. Sufficiency. S is sufficient for the model if the conditional distribution of $\boldsymbol{W}$ given S does not depend on $\boldsymbol{\theta}$. That is, there is a function $h(\boldsymbol{w} \mid \boldsymbol{s})$ such that

$$P[\boldsymbol{W} = \boldsymbol{w} \mid S(\boldsymbol{W}) = \boldsymbol{s}] = h(\boldsymbol{w} \mid \boldsymbol{s}) = \frac{g_{\boldsymbol{\theta}}(\boldsymbol{w})}{\sum_{\{\boldsymbol{v} \mid S(\boldsymbol{v}) = \boldsymbol{s}\}} g_{\boldsymbol{\theta}}(\boldsymbol{v})} \text{ if } S(\boldsymbol{w}) = \boldsymbol{s}.$$

The Fisher factorization theorem states that S is sufficient if and only if there exist functions $a(\boldsymbol{w})$ and $b_{\boldsymbol{\theta}}(s)$ such that

$$g_{\boldsymbol{\theta}}(\boldsymbol{w}) = a(\boldsymbol{w}) b_{\boldsymbol{\theta}}(S(\boldsymbol{w})). \tag{1.8.1}$$

Thus the parameter $\boldsymbol{\theta}$ "touches" the data $\boldsymbol{w}$ only through the function $S(\boldsymbol{w})$.

As an example, suppose there are n observed ranking vectors $\boldsymbol{X}_1$, $\boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ which are independent with the same Mallows' ϕ model density (1.5.5). The $\boldsymbol{W}$ consists of all n vectors. The joint density of the $\boldsymbol{X}_i$'s is the product of the individual f_γ's:

$$g_\gamma(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} f_\gamma(\boldsymbol{x}_i) = \frac{1}{c(\gamma)^n} e^{\gamma \sum_{i=1}^{n} d(\boldsymbol{x}_i)}.$$

Thus with $S(\boldsymbol{w}) = \sum d(\boldsymbol{x}_i)$, (1.8.1) holds where $b_\gamma(s) = c(\gamma)^{-n} e^{\gamma s}$, hence $\sum d(\boldsymbol{x}_i)$ is sufficient. This represents a substantial reduction in the data, from nm numbers to just one.

Other models show other reductions. The Thurstonian models typically do not allow much of any reduction. The Babington Smith paired comparison models have $\binom{m}{2}$-dimensional sufficient statistics being the number of people who prefer object l_i to object l_j, $i < j$. In any case, making such reductions when possible can be very helpful.

A softer, more data-analytic approach to modeling does not start with a mechanism and then try to find a density and hence sufficient statistic; rather, one starts with some numerical summaries of the data one thinks capture the main features of the data, and then creates a model that has those summaries for the sufficient statistics. For example, one might have data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, and decide to summarize the data by looking at just the average ranks $\overline{\boldsymbol{y}} = \sum \boldsymbol{y}_i/n$. The corresponding model would be as in (1.8.1) with S being the vector of means. One still must choose a, b and $\boldsymbol{\theta}$. One common method for choosing them is to try to find a distribution which is as close to being uniform as possible given fixed values for the expected values of the vector S. If closeness to uniform is measured using the entropy function, then the resulting distribution is the **exponential family** distribution with parameter $\boldsymbol{\theta}$, there being one parameter for each element of S. That is, if $S = (S_1, S_2, \ldots, S_p)$, then

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \frac{1}{c(\boldsymbol{\theta})} e^{\sum_{i=1}^{p} \theta_i S_i(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)}. \tag{1.8.2}$$

This is a very flexible and rich set of models. The next two subsections exhibit some particular ones. We note that the exponential family model with $S = \overline{\boldsymbol{y}}$, the vector of mean ranks, is equivalent to the Bradley-Terry-Mallows model (1.5.3).

1.8.1 Counts of rankings

Paul Holland (see Diaconis (1988), p.172) and Verducci (1982) suggest using S to be counts of the number of people giving object l_i the rank of j, $1 \leqslant i, j \leqslant m$. There are potentially $(m-1)^2$ statistics, hence parameters, but simpler distributions can be found by making some restrictions. For example, let S_{l_i} be the number of people who rank object l_i #1. Another model might have S_{kl_i} be the number of people who rank object l_i k^{th}, but only $k = 1$ and 2 are used. More complicated models can be built with statistics of the form "the number of people who give objects l_i and $l_{i'}$ the ranks j and j′, respectively. See Silverberg (1980) and Plackett (1975).

Diaconis (1988; 1989) presents a spectral analysis approach to analysis of rank data. He shows how to decompose the overall distribution into first-order, second-order, etc., effects. Although he does not create models out of this analysis, Verducci (1982) uses the spectral effects as sufficient statistics for some exponential family models.

1.8.2 Distance-based models

Often, one assumes there is a "modal" ranking $\boldsymbol{\mu} \in \mathcal{P}_m$ of the objects, and judges are expected to have rankings more or less close to $\boldsymbol{\mu}$. Appropriate models would give higher probability to rankings closer to the modal ranking; the question is how to measure the distance. There are a number of well-known metrics and distances used for rankings in $\mathcal{P}_m$. Diaconis (1988) and Critchlow (1985) have extensive discussions. Suppose $\boldsymbol{y}, \boldsymbol{z} \in \mathcal{P}_m$. Some measures are

$$\begin{aligned}
\text{Kendall's } \tau \text{ distance} &: \sum_{i<j} I[(y_i - y_j)(z_i - z_j) < 0]; \\
\text{Spearman's } \rho \text{ distance} &: \sum_{i=1}^{m} (y_i - z_i)^2; \\
\text{Spearman's footrule} &: \sum_{i=1}^{m} |y_i - z_i|; \\
\text{Hamming distance} &: \#\{i \,|\, y_i \neq z_i\}; \\
\text{Cayley distance} &: \text{ Minimum number of transpositions to} \\
&\qquad\qquad\qquad \text{take } \boldsymbol{y} \text{ to } \boldsymbol{z}; \\
\text{Ulam distance} &: m - \text{ length of longest set of monotone} \\
&\qquad\qquad\qquad \text{pairs}(y_i, z_i).
\end{aligned} \tag{1.8.3}$$

Exponential family distance-based models are those in (1.8.2) with the one statistic $S = \sum_{i=1}^{n} d(\boldsymbol{y}, \boldsymbol{\mu})$, where d is the particular distance used. Mallows' ϕ model is of this form with Kendall's τ metric, and Mallows' θ model uses Spearman's ρ metric. These and other such models can be found in Fligner and Verducci (1986). In some cases, the modal ranking $\boldsymbol{\mu}$ is known, and in other cases it is of interest to estimate it. In the latter case these models have parameters $(\theta, \boldsymbol{\mu}) \in \mathbb{R} \times \mathcal{P}_m$.

1.9 Unfolding models – Stochastic

Recall the unfolding models in Section 1.7, where it was assumed that there was a linear scale so that all observed rankings could be found by appropriate placing of objects and judges on the scale. In typical sets of data, there is no such placement possible. One approach to extending the non-stochastic model is to allow people to make errors with certain probabilities. Thus for a given placement of the objects on a scale, there is a set of orderings $\mathcal{O}_U \in \mathcal{O}$ which contains all the orderings which are exactly obtainable from the scale. Then any particular ranking $\boldsymbol{y}$ is either in $\mathcal{O}_U$, or a certain distance from some element of $\mathcal{O}_U$. van Blokland-Vogelesang (1989) uses Kendall's τ distance to create the following model which allows variance in the rankings. Assume that for each $\boldsymbol{\mu} \in \mathcal{O}_U$, there is a proportion $q_{\boldsymbol{\mu}}$ of people in the population with $\boldsymbol{\mu}$ as their "true" ranking. The distribution of the ranking $\boldsymbol{Y}$ for those people then follows Mallows' ϕ model. Thus the probability of any ranking $\boldsymbol{y}$ is a mixture

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \sum_{\boldsymbol{\mu} \in \mathcal{O}_U} q_{\boldsymbol{\mu}} \frac{1}{c(\theta_i)} e^{\theta_i d(\boldsymbol{y}, \boldsymbol{\mu})}. \tag{1.9.1}$$

Chapter 2

Distinctions and Commonalities

2.1 Introduction

Chapter 1 revealed a number of approaches to finding models for rank data. In fact, it would be quite easy to become confused about what differences and similarities exist among the models. This chapter discusses various ways in which to classify the models, thus to bring a sense of order to the proceedings.

2.2 Thurstone + Luce choice models

Sections 1.4 and 1.6.1 present two different choice models. Using the notation of Section 1.6.1, we have that for $\mathcal{S} \subset \mathcal{O}$ and $r \in \mathcal{S}$, $p_{\mathcal{S}}(r)$ is the probability of choosing object r as the favorite among the objects in $\mathcal{S}$. For each fixed $\mathcal{S}$, $\{p_{\mathcal{S}}(r) \,|\, r \in \mathcal{S}\}$ is a probability distribution on $\mathcal{S}$. The totality of such distributions is

$$\boldsymbol{p} \equiv \{\{p_{\mathcal{S}}(r) \,|\, r \in \mathcal{S}\} \,|\, \mathcal{S} \subset \mathcal{O}\}. \tag{2.2.1}$$

With no restrictions, the $\boldsymbol{p}$ has an $(1 + (m-2)2^{m-1})$-dimensional range. Thus except for fairly small m, it is necessary to demand some coherence from the set of probabilities. One idea is to disallow interactions of the objects, that is, one's response to a particular object depends only on the object itself, not on which other objects it happens to be compared to. Thus we wish to avoid situations with probabilities such as the following:

$$P_{\{\text{Coke, Pepsi, 7-up, Sprite}\}}(\text{Coke}) = 90\%$$
$$\text{while } P_{\{\text{Coke, Pepsi, 7-up}\}}(\text{Coke}) = 10\%.$$

Luce's approach is to posit a reasonable axiom that would govern the probabilities. This axiom is Luce's choice axiom of Section 1.6.1. Note that the axiom places a requirement on the (probabilities of) observed results of choice experiments, but does not try to explain how one arrives at the choices. By contrast, Thurstone's approach proposes an actual mechanism to model responses of the individual to stimuli, from which the choice probabilities arise. Thus if $\boldsymbol{Z}$ is the random vector associated with the objects as in Section 1.4, then the choice probabilities are found to be

$$P_{\mathcal{S}}(r) = P[Z_r < Z_s \text{ for all } s \in \mathcal{S} - \{r\}]. \tag{2.2.2}$$

When, if ever, does a Thurstonian model satisfy Luce's axiom? This question is addressed in Yellott (1977). Make the conventions that

1. There is always a chance any one of the objects will be chosen, i.e., $P_{\mathcal{S}}(r) > 0$ for all $r \in \mathcal{S} \neq \emptyset$;
2. For Thurstonian models, the $\boldsymbol{Z}$ follows a location-family model (1.4.2).

The following is Yellott's Theorem 5.

Theorem 2.2.1. ***Yellott's Theorem.*** *When $m \geqslant 3$, a Thurstonian model has probabilities* (2.2.2) *that satisfy the choice axiom if and only if the Z_{l_i}'s have the Gumbel distribution, that is, for some $a > 0$ and b,*

$$P[Z_{l_i} - \mu_{l_i} \leqslant z] = e^{-e^{-a(z-b)}}. \tag{2.2.3}$$

This is quite an interesting result from the foundational point of view. Whereas Thurstonian models in general do capture the notion of non-interaction of responses, only for a very special distribution do they satisfy the choice axiom. Whether there is something fundamentally wrong with the axiom or with the general Thurstonian model is up to the reader to decide. However, the following conjecture may, in practical terms, ease the conflict.

Robustness of the Thurstone models. All reasonable functions g in (1.4.2) yield approximately the same choice probabilities in (2.2.2). Thus they all approximately satisfy the choice axiom.

U_r	$U_r - U_s$
N(0,1)	N(0,2)
Gumbel	Logistic
Exponential	Double Exponential
Uniform	Tent

Table 2.1: Some distributions

The terms "reasonable" and "approximately" are suitably vague; the extent to which the conjecture is true is fruit for future research. Evidence is based on the calculation of paired-comparison probabilities arising from different g's. In general, under the Thurstonian model (2.2.2), for objects r and s,

$$p_{rs} = P[Z_r < Z_s] = P[U_r - U_s < -(\mu_r - \mu_s)] \equiv D(-(\mu_r - \mu_s)), \quad (2.2.4)$$

where $U_r = Z_r - \mu_r$ and $U_s = Z_s - \mu_s$ are independent with density g, and D is the distribution function of $U_r - U_s$. Some examples are given in Table 2.1

One implication of the conjecture is that when using Thurstonian models with location family density for $\boldsymbol{Z}$, the exact distribution assumed is not overly crucial; in fact, for most experimental situations it would be impossible to distinguish between these models based on choice data alone, hence impossible to reject the choice axiom. Thus one may as well use whichever model is easiest to work with analytically and computationally. My guess is that that one would be the Gumbel $\equiv$ Luce model.

Recall from Section 1.6.1 that when the Luce model holds, and convention 1 holds, there are positive constants ν_{l_i} corresponding to objects l_i, $i = 1, \ldots, m$, such that the choice probabilities have the form in (1.6.2). On the other hand, the Thurstone/Gumbel model has the location parameters μ_{l_i}. What is their relation? It can be shown that it is enough to take $a = 1$ and $b = 0$ in (2.2.3). Then from Table 2.1, since the logistic distribution function is $D(z) = (1 + e^{-z})^{-1}$,

$$\begin{aligned} \frac{\nu_r}{\nu_r + \nu_s} &= P[Z_r < Z_s] = D(-(\mu_r - \mu_s)) \\ &= \frac{1}{1 + e^{\mu_r - \mu_s}} = \frac{e^{-\mu_r}}{e^{-\mu_r} + e^{-\mu_s}}. \end{aligned} \quad (2.2.5)$$

Thus we can make the identification $\mu_r = -\log(\nu_r)$.

2.3 From choice to rankings

Choice probabilities as in (2.2.1) by themselves do not yield probabilities for ranks. There are three paradigms for connecting choice and rank probabilities presented in the first chapter:

- **Thurstonian.** Refer to the underlying mechanism that produces both choice and rank probabilities;
- **Conditional.** Perform a number of predetermined choice experiments independently, and report a ranking if the choice experiments result in a complete and unambiguous ranking; otherwise, repeat the set of choice experiments until a ranking is found (Babington Smith);
- **Multistage.** Perform a number of choice experiments independently, which ones to perform being sequentially decided given the outcomes of the previous (L-decomposable models (1.6.1)).

Before continuing, some notation is needed to provide an overall framework for the models to be discussed. A **choice experiment** is indexed by subsets $\mathcal{S} \subset \mathcal{O}$: $E_{\mathcal{S}}$ is the experiment in which one chooses the best among the objects in $\mathcal{S}$. The **outcome** of a particular performance of the experiment, $\text{Outcome}(E_{\mathcal{S}})$, is a random variable. Thus from (2.2.1),

$$P_{\mathcal{S}}(r) = P[\text{Outcome}(E_{\mathcal{S}}) = r].$$

A **choice-based ranking experiment** is a (possibly infinite) sequence of choice experiments, where which experiment to perform at the i^{th} stage is allowed to depend on the outcomes of the previous $i-1$ stages, together with the specification of when to stop and how to determine the ranking at the end. The only restrictions placed on the ranking experiment and choice probabilities are the following:

1. The distribution of the outcome of the i^{th} choice experiment $E_{\mathcal{S}_i}$ is independent of the previous outcomes given the choice experiment selected.
2. The ranking experiment stops with probability 1.

A number of models from Chapter 1 can be given as such a ranking experiment:

- **Babington Smith**. First perform the $\binom{m}{2}$ pairwise choice experiments $E_{\{l_i,l_j\}}$ for $i < j$. If the outcomes yield a definitive ranking, then stop. Otherwise, perform the $\binom{m}{2}$ pairwise experiments again. Keep going until the last set of pairwise comparisons yields a ranking. Bradley-Terry, Mallows′ ϕ and Mallows′ θ, component and orthogonal contrast models are special cases.

- **L-decomposable models** (1.6.1). Perform the $m-1$ experiments $E_{\mathcal{S}_1}, E_{\mathcal{S}_2}, \ldots, E_{\mathcal{S}_{m-1}}$, where $x_i = \text{Outcome}(E_{\mathcal{S}_i})$ and

$$\mathcal{S}_{i+1} = \mathcal{S}_i - \{x_j \mid j = 1, \ldots, i\}.$$

Then the resulting ranking is $\boldsymbol{x} = (x_1, \ldots, x_{m-1}, x_m)$, where x_m is the object left in $\mathcal{S}_m$. These models include:

  - **Babington Smith** model in Section 1.5.1. The actual choice probabilities are complicated. See Critchlow, Fligner, and Verducci (1991).
  - **Luce model**, which is also Plackett′s and Silverberg′s model (1.6.3), where there are ν_i′s, positive, such that the probabilities of the outcomes are given in (1.6.2).
  - **Component models** in Section 1.6.2. Here, the probabilities of outcomes are given by

$$P_{\mathcal{S}}(l_i) = h_i(u_i),$$

where $u_i + 1$ is the rank of l_i among the elements in $\mathcal{S}$. The ϕ component model is a special case, where the h_i′s are given in (1.6.5), and Mallows′ ϕ model is a further specialization, where the θ_i′s in (1.6.6) are set equal.

Most Thurstonian and distance-based models are not of the above form, except when they happen to coincide with the models just mentioned.

The choice-based ranking experiments based on choice experiments can be looked at as hypothesized models for the way an individual proceeds when asked to rank the objects, or as a method to design an experiment to ascertain someone′s ranking by sequentially presenting the subject with a number of choice experiments. Many other schemes can be devised. For example, one might use an adaptive sequence of paired comparisons chosen such that there is no chance for nontransitivities to arise. A possible sequence, with the objects A, B, C, D and E, is given in Table 2.2.

Such a scheme may be superior to Babington Smith, since it probability takes fewer comparison′s (Babington Smith takes at least 10 ($= \binom{5}{2}$), and possibly a lot more), and superior to the L-decomposable schemes since pairwise comparisons are generally easier to make.

Experiment	$\mathcal{S}$	Outcome	Ranking so far
1	A,B	A	(A,B)
2	A,C	A	??
3	B,C	C	(A,C,B)
4	A,D	D	(D,A,C,B)
5	D,E	D	??
6	A,E	A	??
7	C,E	E	(D,A,E,C,B)

Table 2.2: A possible ranking sequence

Note that the choice experiments above all involve choosing the one best among the subset $\mathcal{S}$. Another set of models can be built on experiments in which one picks the one worst from $\mathcal{S}$, or even the two best, or four worst, etc. It is reasonable to suppose that actual ranking proceeds with a complicated mixture of choice experiments. For example, one might first decide which is worst of all, then divide the rest into three groups, then within each group use a number of paired comparisons. More research in this area would be interesting.

The conditional and multistage models referred to above are choice-based. The Thurstonian models are not. In Thurstonian models, there is just one response, Z_{l_i}, for each object. The responses are noted, and the ranking obtained by ordering the Z_{l_i}'s. Thus the ranking occurs as a one-shot deal. For the choice-based models, one has to respond (possibly) several times independently to each object. Of course, it is possible that the individual choice experiments follow Thurstonian models, but there is still a difference in choice-based ranking models and Thurstonian ranking models. To illustrate, suppose a Thurstonian model holds for the responses to the stimuli (objects). Consider the Thurstonian ranking model, and the L-decomposable model (1.6.1) where the choice probabilities are given by the same Thurstonian model. Possible ways to obtain the ranking (B, C ,D, A, E) from the two experiments:

Thurstone: Observe $Z_B < Z_C < Z_D < Z_A < Z_E$;

and

Model (1.6.1) :

1. Observe $Z_B^{(1)} < Z_D^{(1)} < Z_A^{(1)} < Z_E^{(1)} < Z_C^{(1)}$, then
2. Observe $Z_C^{(2)} < Z_E^{(2)} < Z_D^{(2)} < Z_A^{(2)}$, then
3. Observe $Z_D^{(3)} < Z_E^{(3)} < Z_A^{(3)}$, and, finally
4. Observe $Z_A^{(4)} < Z_E^{(4)}$.

All the $Z_{l_i}^{(k)}$'s are independent. In general, it should be clear that the probabilities of the ranking under the two models need not be at all the same. What is amazing is that they can be:

Lemma 2.3.1. *Suppose the Thurstonian model* (1.4.2) *holds, where the* $-Z_{l_i}$*'s are Gumbel random variables. Then the Thurstonian ranking model and the L-decomposable model* (1.6.1) *yield the same distribution on* $\mathcal{L}_m$.

I am inclined to see this result as a bit of a fluke, rather than one that reveals a deep connection between these models. I may be wrong. A corollary to the lemma says that if the Z_{l_i}'s themselves are Gumbel, then the model is equivalent to the *backwards* analog of (1.6.1), that is, one first chooses the worst of all, then the worst of what remains, etc.

2.4 Luce's choice axiom: Backwards and forwards

In Section 1.6.1, Luce's choice axiom was presented as a reasonable axiom for a person's choice probabilities to satisfy. These probabilities are for choosing the best one of a subset of objects. By the same token, it may also be reasonable to expect choice experiments in which the person chooses the worst one of a subset to satisfy the axiom. Replace the one set of choice probabilities in (2.2.1) with the two sets

$$\boldsymbol{p} \equiv \{\{p_{\mathcal{S}}(r) \mid r \in \mathcal{S}\} \mid \mathcal{S} \subset \mathcal{O}\} \text{ and } \boldsymbol{q} \equiv \{\{q_{\mathcal{S}}(r) \mid r \in \mathcal{S}\} \mid \mathcal{S} \subset \mathcal{O}\},$$

where

$$Q_{\mathcal{S}}(r) = \text{Prob[Object } r \text{ is the least preferred among those in } \mathcal{S}].$$

Make the following assumptions:

(1) Both $\boldsymbol{p}$ and $\boldsymbol{q}$ satisfy Axiom 1.6.1.

(2) $P_{\{r,s\}}(r) > 0$ for all $r, s \in \mathcal{O}$.

(3) For any ordering of any subset of objects, the forward and backward choice-based probabilities are the same.

Luce (1959, page 57) shows that if $m \geqslant 3$, no Thurstonian model yields choice probabilities that satisfy all three assumptions. Yellott (1977) amplifies on this result by noting that because the axiom holds for a Thurstonian model for $\boldsymbol{p}$ if and only if the Z_{l_i}'s are Gumbel, it holds for $\boldsymbol{q}$ if and only if the Z_{l_i}'s are negatives of Gumbels. Since Gumbels are not symmetric, the axiom cannot hold for both $\boldsymbol{p}$ and $\boldsymbol{q}$. (An exception is if all the μ_{l_i}'s are equal, in which case the model is just the uniform model.) Luce and Yellott seem to take this result as a strike against Thurstonian models. However, without any reference to Thurstonian models, we have the following.

Lemma 2.4.1. *Suppose Assumptions 1, 2 and 3 hold. Then all objects are equally preferred, i.e.,* $P_{\mathcal{S}}(r) = Q_{\mathcal{S}}(r) = 1/\#\mathcal{S}$ *for all* $r \in \mathcal{S} \subset \mathcal{O}$.

Proof. Assumptions 1 and 2 imply that (1.6.2) holds, and analogously, that there exist positive u_{l_i}'s such that

$$Q_{\mathcal{S}}(l_i) = \frac{u_{l_i}}{\sum_{l_j \in \mathcal{S}} u_{l_j}}. \tag{2.4.1}$$

Assumption 3 implies that $P_{\{r,s\}}(r) = Q_{\{r,s\}}(s)$, which can be used to show that without loss of generality we can take $\nu_{l_i} = 1/u_{l_i}$. Now try ranking three of the objects, A, B, and C, using the forward choice-based experiment (1.6.3); also, using the backward choice-based experiment. The probability of a given ordering is the same no matter which method is used. Thus, e.g.,

$$\begin{aligned} \text{Prob}[\boldsymbol{x} = (A,B,C)] &= P_{\{A,B,C\}}(A)P_{\{B,C\}}(B) \\ &= Q_{\{A,B,C\}}(C)Q_{\{A,B\}}(B). \end{aligned}$$

Letting $\nu_A = u_A = 1$, (1.6.2), (2.4.1), and (2.4.2) can be used to show that $\nu_C = \nu_B^2$. Doing the same for $P[\boldsymbol{x} = (A,C,B)]$ shows that $\nu_B = \nu_C^2$. Thus it must be that $\nu_B = \nu_C = 1 = \nu_A$, hence $u_A = u_B = u_C = 1$. The same argument shows that in fact $\nu_{l_i} = u_{l_i} = 1$ for all objects l_i. □

What are the implications of the lemma? Since no non-uniform model can satisfy the three assumptions, it is hardly fair to blame Thurstonian models for being unable to satisfy them. Actually, the lemma shows that whatever leads one to accept the axiom for $\boldsymbol{p}$ must not lead one to accept it for $\boldsymbol{q}$ unless one believes the uniform model. Thus what may at first seem like a rather innocuous axiom is quite strong, and one may be tempted to not expect models to adhere to it strictly.

2.5 Unidimensionality, unimodality and consensus

The idea that there is an overall ranking for the objects is very important in ranking situations. For example, in competitions, several judges' rankings must be combined into one overall ranking; in sociology, the problem is that of arriving at a social consensus. The question can also be asked of an individual, that is, whether after repeated observations of the person's rankings, one can ascertain an overall ranking. Basically, one is asking for a placing of the objects on a one-dimensional scale, where objects on the left are more likely to be preferred.

There are a number of ways to define an overall ranking given a set of rankings or a distribution on $\mathcal{P}_m$ that yield a result no matter how inconsistent the individual rankings may be. Thus one actually desires more than just a composite ranking; one wishes to be assured that the ranking tells the whole story. Some intuitive notions of what is meant by "telling the whole story" include

1. Objects with lower overall rank are more likely to be preferred than those with a higher overall rank in paired comparisons. (If one object is preferred to all others in paired comparisons, it is called the **Condorcet** choice. See Wikipedia (2017).)

2. An object being #1 means that it is most likely to be ranked first, second most likely to be ranked second, ..., and least likely to be ranked last. Similarly, the #2 object is second most likely to be ranked #1, ..., and second least likely to be ranked last. Etc.

3. There is no interaction among objects in that the relative preference of two objects is not affected by the relative ordering of the others.

It is not hard to find examples in which these notions are violated. For example, number 1 need not hold since it could be that A has a lower average rank than B, but in paired comparisons most people prefer B to A. Number 2 is easily violated. For example, suppose $P[(A, B, C)] = 0.55$ and $P[(C, B, A)] = 0.45$. Though A is more likely to be #1 than B is (B is never #1), A is also more likely to be #3 (B is never #3). Number 3 avoids situations such as having 7-up preferred to Sprite when Coke is ranked first, but Sprite is preferred to 7-up when Coke is ranked second.

Next, some requirements are presented that try to formalize the intuitive ideas above. The properties below are formulated in terms of the probability density f on the ranks $\boldsymbol{y} \in \mathcal{P}_m$, where the ordering $\boldsymbol{x}$ is fixed at $(l_1, l_2, \ldots, l_m)$.

- **Unimodality**. The density f is unimodal if there exists one ranking $\boldsymbol{\mu}$ that has higher probability than any other, i.e., $f(\boldsymbol{\mu}) > f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathcal{P}_m - \{\boldsymbol{\mu}\}$. The ranking $\boldsymbol{\mu}$ is the **modal** ranking.
- **Strong unimodality** with respect to a given partial ordering $<_*$. The density is **strongly** unimodal with respect to $<_*$ if it is unimodal, and whenever $\boldsymbol{\mu} <_* \boldsymbol{y} <_* \boldsymbol{z}$, $f(\boldsymbol{y}) > f(\boldsymbol{z})$.
- **Order in expectation**. The objects are said to be ordered in expectation if $E[Y_1] < E[Y_2] < \cdots < E[Y_m]$.
- **Marginal stochastic ordering**. The objects are marginally stochastically ordered if the marginal distributions of the Y_i's are distinct, and for any number k, $P[Y_1 \leqslant k] \geqslant P[Y_2 \leqslant k] \geqslant \cdots \geqslant P[Y_m \leqslant k]$.
- **Consensus.** The ordering is a consensus ordering if $P[Y_i < Y_j] > 1/2$ for all $i < j$.
- **Complete consensus**. There is complete consensus about the ordering if for all $i < j$

  $$P[Y_i < Y_j \mid Y_k = r_k, \text{ for all } k \neq i, j) > 1/2$$

  for all sets of ranks r_k.

See Critchlow, Fligner, and Verducci (1991), Henery (1981), and Fligner and Verducci (1988). In this section, we will refer to these three papers as CFV, H, and FV, respectively. The definition of consensus here differs from that in FV.

Unimodality and order in expectation are very weak properties. They are satisfied as long as there are enough differences among the probabilities of the rankings. They do not require any real cohesion among the rankings. Marginal stochastic ordering and consensus are stronger conditions, but do not eliminate the possibility of interactions among the objects. For example, one might prefer A to B 95% of the time if C is ranked third, but prefer B to A 55% of the time when C is ranked first. Strong unimodality (with nontrivial $<_*$) and complete consensus do restrict the possible interactions, at least qualitatively.

Different partial orderings $<_*$ yield different strong unimodalities, of course. We will give three possibilities, all based on the modal ranking being $\boldsymbol{\mu} = (1, 2, \ldots, m)$. (If $\boldsymbol{\mu}$ is not $(1, 2, \ldots, m)$, relabel the objects so that it is.) CFV suggest $\boldsymbol{y} <_{CFV} \boldsymbol{z}$ if for some $i < j$,

$$y_j = z_i = y_i + 1 = z_j + 1 \text{ and } y_k = z_k \text{ for } k \neq i, j.$$

H's ordering says $\boldsymbol{y} <_H \boldsymbol{z}$ if

$$y_j = z_i < y_i = z_j \text{ and } y_k = z_k \text{ for } k \neq i, j.$$

Model	Number associated with object l_i	See
Thurstone	μ_{l_i}	(1.4.1), (1.4.2)
Bradley-Terry	$1/\nu_{l_i}$	(1.5.3)
Plackett-Luce	$1/\nu_{l_i}$	(1.6.3)
Distance-based	μ_i	Section 1.8.2
ϕ component	θ_i	(1.6.6)

Table 2.3: Some item parameters

Kendall's τ ordering says $\boldsymbol{y} <_K \boldsymbol{z}$ if $d(\boldsymbol{\mu}, \boldsymbol{y}) < d(\boldsymbol{\mu}, \boldsymbol{z})$, where d is Kendall's τ metric as in Section (1.8.2). Other orderings immediately arise from other distances. The partial orderings are themselves ordered:

$$[\boldsymbol{y} <_{\mathrm{CFV}} \boldsymbol{z}] \implies [\boldsymbol{y} <_{\mathrm{H}} \boldsymbol{z}] \implies [\boldsymbol{y} <_{\mathrm{K}} \boldsymbol{z}].$$

The idea is that for CFV, if two rankings are identical except for two objects with adjacent ranks, then the one with the two objects in the wrong order (with respect to $\boldsymbol{\mu}$) is farther from $\boldsymbol{\mu}$ than the other. The H ordering is similar, except that the requirement that the ranks have to be adjacent is dropped. The K ordering just looks at the number of discordances between $\boldsymbol{\mu}$ and each of the rankings.

Some implications:

$$\begin{aligned} \text{Complete consensus} &\iff \text{Strong Unimodality}[<_{\mathrm{H}}] \\ &\implies \left\{ \begin{array}{l} \text{Unimodality} \\ \text{Marginal stochastic ordering} \\ \qquad\qquad \implies \text{Order in Expectation} \\ \text{Consensus} \end{array} \right. \end{aligned}$$

Unidimensionality is not precisely defined, but the idea is that there is a real number attached to each object, and the lower the number the more likely the object will be preferred. CFV call these numbers **item parameters**. In Table 2.3, these parameters are exhibited for some of the models in Chapter 1.

Now to see which models satisfy which of the unimodality and consensus conditions.

Thurstonian models with $\boldsymbol{Z}$ of the form (1.4.2) have complete consensus if $\mu_{l_1} < \cdots < \mu_{l_m}$, and the density g has monotone likelihood ratio, i.e.,

$$\frac{g(z - \mu_{l_i})}{g(z - \mu_{l_j})} \text{ is decreasing in } z \text{ for } i < j.$$

See H. Even without the monotone likelihood ratio property, any Thurstonian model with the μ_{l_i}'s strictly increasing will have marginal stochastic ordering, hence order in expectation. See CFV.

Babington Smith models do not have item parameters in general. However, there are notions of qualitative ordering. A Babington Smith model has **weak stochastic transitivity** if for i, j,

$$p_{l_i l_j} \geqslant \frac{1}{2} \text{ and } p_{l_j l_k} \geqslant \frac{1}{2} \implies p_{l_i l_k} \geqslant \frac{1}{2}, \tag{2.5.1}$$

and has **strong stochastic transitivity** if

$$p_{l_i l_j} \geqslant \frac{1}{2} \text{ and } p_{l_j l_k} \geqslant \frac{1}{2} \implies p_{l_i l_k} \geqslant \max\{p_{l_i l_j}, p_{l_j l_k}\}. \tag{2.5.2}$$

See David (1988) for these definitions. CFV show that a Babington Smith model is strongly unimodal $[<_{CFV}]$ if and only if it has weak stochastic transitivity, and if it has strong stochastic transitivity, it has complete consensus.

Bradley-Terry and **Plackett-Luce** models show complete consensus whenever $\nu_{l_1} > \nu_{l_2} > \cdots > \nu_{l_m}$. See H and CFV.

Distance-based models show complete consensus whenever $f(\boldsymbol{y})$ decreases as $d(\boldsymbol{\mu}, \boldsymbol{y})$ increases, and $\boldsymbol{y} <_H \boldsymbol{z}$ implies that $d(\boldsymbol{\mu}, \boldsymbol{y}) < d(\boldsymbol{\mu}, \boldsymbol{z})$. It has already been noted above that the Kendall's τ distance has this property, hence Mallows' ϕ model has complete consensus if the parameter $\gamma < 0$. Also, if $\gamma > 0$, the model has complete consensus but with the modal ranking being the opposite of $\boldsymbol{\mu}$, which is $(m + 1 - \mu_1, \cdots, m - 1 - \mu_m)$. Spearman's ρ and footrule distances, Hamming distance, and Ulam distance have this property, as well.

The ϕ **component model** (1.6.6) has complete consensus if $\theta_i/\theta_{i+1} \geqslant ((m-1)-i)/(m-i)$ for $i = 1, \ldots, m-2$. See FV, Theorem 2.3.

Orthogonal contrast models in general do not lend themselves to such unidimensionality questions. In fact, part of their usefulness lies in being able to capture interesting deviations from consensus and unimodality. See Section 1.8 for some other such models. Unfolding models fit in to this framework by considering the mixture model (1.9.1). The idea is that for a given unfolding scale, there are $\binom{m}{2} + 1$ possible rankings, and for each of these rankings there is a population of judges and corresponding distribution of rankings. Consensus, unimodality and the other properties are said to hold for the model (1.9.1) if they hold for each of the subpopulations, where the modal ranking is that from the unfolding scale connected with that population.

Name	See	p	Parameters	Statistics
Babington Smith	(1.5.2)	$\binom{m}{2}$	$\log\left(\frac{p_{l_i l_j}}{1-p_{l_i l_j}}\right)$	$I[y_i < y_j]$
Bradley-Terry	(1.5.3)	$m-1$	$\log(\nu_{l_i})$	$m - \text{index}(x_i)$
Mallows' ϕ	(1.5.5)	1	γ	$d_K(\boldsymbol{x})$
Mallows' Θ	(1.5.7)	1	γ	$s(\boldsymbol{x})$
Free component	(1.6.4)	$\binom{m}{2}$	$\log\left(\frac{h_i(j)}{h_i(m-i)}\right)$	$I[u_i = j]$
ϕ component	(1.6.6)	$m-1$	$\boldsymbol{\theta}$	$\boldsymbol{u}$
ϕ orthogonal contrast	(1.6.9)	$m-1$	$\boldsymbol{\theta}$	$d(C_i(\boldsymbol{y}))$
Contingency table	Section 1.6.3	Lots	Loglinear	Cell counts

Table 2.4: Some sufficient statistics

2.6 Regular exponential families

Section 1.8 presented exponential family models. For ranking models, a **regular** exponential family has the form (1.8.2) with the additional requirements that

1. The parameter space Θ is $\mathbb{R}^p$;

2. The distribution is **identifiable**, i.e., if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, then the densities $f_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$ are different.

The reason for number 1 is that any of the ranking models is a submodel of the saturated model, which is an exponential family model with $p = m! - 1$, so that if any type of parameter space is allowed, all models are exponential family models. Number 2 just makes sure that there are no superfluous dimensions in $\boldsymbol{\theta}$. In a regular exponential family, p is referred to as the dimension of the model.

Many of the models in Chapter 1 are regular exponential family models. The ones in Sections 1.8.1 and 1.8.2 are because they constructed to be. About the only ones that are not are the Thurstonian models and some of the L-decomposable models such as Plackett-Luce. Table 2.4 exhibits the regular exponential models. The statistics are given for one observed ranking. The statistics for a sample consist of the sums of the statistics for the individual observations. Note that the free orthogonal contrast model is a contingency table model.

2.7 Hierarchical systems of models

Most of the models so far assume limited interactions among objects, as in Section 2.5. In many data sets, it is the interactions that are of interest. Thus there has been some effort in finding classes of models that mimic analysis-of-variance and categorical loglinear models, where one first tries no-interaction models, then models with two-way interactions, then three-way, and so on, until arriving at the saturated model. The object is to find the simplest model that fits. The contingency table model based on orthogonal contrasts in Section 1.6.3 is one such model. In this section, more systems are presented.

The general structure for each system below is that, for each order k, there exists a set of parameters that capture all the interactions of the objects up to the k^{th} order, where "order" has different meanings for different systems. The 0^{th}-order model is the uniform model, and the highest-order model is the saturated model. There is a hierarchy to each system in that the lower-order models are all contained in the higher-order ones.

In the Plackett system, the k^{th} order refers to the probability that each set of k objects is ranked #1 to #k. The model of Holland-Silverberg and Diaconis-Verducci is more general in that it considers every possible placement of the k objects in the ordering, not just at the beginning. The extensions of Babington Smith uses relative orderings of sets of objects instead of absolute placements. McCullagh (1993) systematizes this set by defining inversions. The orthogonal contrast contingency table approach is different than the others since the interactions are of contrasts of objects rather than of the objects themselves.

2.7.1 Plackett

The Plackett-Luce model (1.6.3) is the first order model for Plackett's (1975) system of hierarchical models. The model is defined on orderings, so the ranking $\boldsymbol{y}$ is fixed at $(1, 2, \ldots, m)$. Let the probability of the ordering $\boldsymbol{x} \in \mathcal{L}_m$ be $p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_m)$, and for any k, let

$$\begin{aligned} p(x_1, \ldots, x_k) &= \sum_{y_{k+1}} \cdots \sum_{y_m} p(x_1, \ldots, x_k, y_{k+1}, \ldots, y_m) \\ &\equiv p(x_1, \ldots, x_k, +, \ldots, +) \\ &= P[x_1 \text{ is ranked first}, \ldots, x_k \text{ is ranked } k^{th}]. \end{aligned}$$

The second line gives the analog of analysis-of-variance notation. The idea in the horse-race paradigm is that the probability of finishing first is proportional to the probability of finishing second if another horse finished first, etc. This idea is formalized by demanding that

$$\frac{p(x_1,\ldots,x_{k-1},a)}{p(x_1,\ldots,x_{k-1},b)} = \frac{p(a)}{p(b)}$$

for all $k, a, b, x_1, \ldots, x_{k-1}$.

Second-order models allow interactions between first and second places. That is, the probability that two objects are ranked first and second is proportional to the probability that they are ranked second and third given another object was ranked first, etc. Third-order interactions are defined by the probabilities that three objects are ranked 1, 2, and 3, etc. Formally, the l^{th}-order model demands that

$$\frac{p(x_1,\ldots,x_{k-1},a_1,\ldots,a_l)}{p(x_1,\ldots,x_{k-1},b_1,\ldots,b_l)} = \frac{p(a_1,\ldots,a_l)}{p(b_1,\ldots,b_l)}$$

for all $k, a_i's, b_i's, x_i's$.

Given a set of probabilities, it is straightforward to check whether any of the l^{th}-order models hold. However, one would also like to have a parametric representation of the models so that given the parameters, all the probabilities can be constructed. Plackett gives such parameters in terms of logs of ratios of the probabilities, hence calls the model a logistic model. It has nothing to do with the Thurstone-Gumbel-Luce model. The parameters he uses are

$$\lambda(a_1,\ldots,a_{l-1},b) \equiv \log\left(\frac{p(a_1,\ldots,a_{l-1},b)p(a_2,\ldots,a_{l-1},c)}{p(a_1,\ldots,a_{l-1},c)p(a_2,\ldots,a_{l-1},b)}\right),$$

where

$$c = \max[\{1,2,\ldots m\} - \{a_1,\ldots,a_{l-1}\}] \ \& \ b \in \{1,2,\ldots,m\} - \{a_1,\ldots,a_l,c\}.$$

The l^{th}-order model is obtained by setting all the λ's with more than l components to 0.

2.7.2 Holland-Silverberg/Diaconis-Verducci

Plackett's model allowed objects to interact, but only when they are ranked consecutively. More general two-way interactions may also be extant, such as the number of times objects a and b are ranked i and j for any i and j, as well as extensions to three- and higher-way interactions. A number of exponential family models have been suggested to capture these interactions.

Start with first-order (no interaction) models. Silverberg (1980; 1984), Paul Holland (Silverberg's Ph.D. advisor), and Verducci (1982) take as sufficient statistics

$$S(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\,;\, l_i\,;\, j) \equiv \#\{q \mid y_{q l_i} = j\}, \tag{2.7.2}$$

which is the number of observations $\boldsymbol{y}_q$ that give object l_i the rank j. The number of free parameters in the regular exponential family is $(m-1)^2$ since the sum of $S(\cdot\,;\, l_i\,;\, j)$ over any i or j is n. This model is more flexible than the Plackett first-order model, and the other models in the previous section, since it does not demand any unimodality.

Second- and higher-order models look at the rankings of groups of objects. The sufficient statistics for the k^{th}-order model are

$$S(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\,;\, l_{i_1}, \ldots, l_{i_k}\,;\, j_1, \ldots, j_k) \equiv \#\{q \mid y_{q l_{i_1}} = j_1, \ldots, y_{q l_{i_k}} = j_k\},$$

for all subsets $\{l_{i_1}, \ldots, l_{i_k}\} \subset \mathcal{O}$ and $\{j_1 < \ldots < j_k\} \subset \{1, 2, \ldots, m\}$. These statistics count the number of observations in which each set of objects receives each possible set of ranks. The dimension of the regular exponential family grows quickly with the order k, and it is not immediate how to find the dimension. Diaconis (1988; 1989) uses spectral analysis for the permutation group to find these dimensions as well as finer splittings of the models. This work is very interesting, but I do not understand it totally. The idea is that any function on $\mathcal{P}_m$ can be written as a linear combination of a number of functions, each corresponding to an invariant subspace of the space of all functions. This decomposition is analogous to the splitting of a time-series into components due to various frequencies, or an analysis-of-variance splitting into main, two-way, three-way, ..., effects. For each subspace, one can find a basis and use the coefficients of the basis vectors for the function $c(\boldsymbol{y}) = \#\{$observations with rank vector $\boldsymbol{y}\}$ for sufficient statistics. Verducci (1982) fits some exponential family models and submodels using these statistics

Diaconis gives a detailed analysis of an $m = 5$ example. The first subspace is just the constant function. The second yields the first-order statistics (2.7.2). The third and fourth subspaces yield the second-order statistics, where the third has functions that ignore the order of the objects, and the fourth does not. There are three other subspaces that take care of any further interaction. The dimensions of the seven subspaces are, respectively, 1, 16, 25, 36, 25, 16 and 1. Thus the second-order model has dimension $1 + 16 + 25 + 36 = 78$.

Silverberg calls the models loglinear models since one can choose parameters of which $\log(P[\boldsymbol{X} = \boldsymbol{x}])$'s is a linear function, analogous to loglinear models in contingency tables. With parameters

$$\alpha(l_{i_1}, \ldots, l_{i_k}\,;\, j_1, \ldots, j_k),$$

the first-order model is

$$\log(P[\boldsymbol{x} = \boldsymbol{x}]) = \alpha_0 + \alpha(l_1\,;\, 1) + \alpha(l_2\,;\, 2) + \cdots + \alpha(l_m\,;\, m).$$

Objects in x-order	Objects in original order	Inversion?
BDA	ABD	Yes
BDE	BDE	No
BDC	BCD	No
DAE	ADE	No
DAC	ACD	Yes
AEC	ACE	No

Table 2.5: Second-order inversions in $x = (\mathrm{B, D, A, E, C})$

The α_0 is the 0^{th}-order effect. It equals $\log(1/m!)$ in the 0-order model. The second-order model is

$$\log(p(l_1, \ldots, l_m)) = \alpha_0 + \sum_{i=1}^{m} \alpha(l_i\,;\, i) + \sum_{i<j} \alpha(l_i, l_j\,;\, i, j).$$

Constraints have to be placed on the parameters for them to be estimable, which constraints do not seem to be easily described without referring to Diaconis's theory.

2.7.3 Extended Babington Smith — McCullagh's inversion models

Babington Smith models have as sufficient statistics the number of times object l_i is preferred to object l_j, $i < j$. One can also look at sets of more than two objects, and the relative order within each set. Thus there are $\binom{m}{3}$ sets of three objects, and each set can be put in $3! = 6$ orders. The sufficient statistics for these count how many observations have each set of three objects in each possible order. Similarly for sets of any k objects. The dimensions of the corresponding parameter spaces and the summarization of the statistics appears complicated.

McCullagh (1993) has a systematic approach that produces interpretable parameters. Starting with a given ordering of the objects, he defines **inversions** for sets of objects to be orderings in which none of the objects are in their original spots. That is, suppose 5 objects start in the order A, B, C, D, E. Take any other ordering x. First order inversions are those pairs of objects that, in x, are in an order different than the original order. Thus if x = (B, D, A, E, C), then there are 4 first-order inversions: BA, DA, DC, and EC. Second-order inversions are triples of objects in which all three objects are in the wrong slot. The Table 2.5 gives the possible triples, their original order, and whether there is a second-order inversion.

Third-order inversions are sets of four objects in which none are in the original slot, etc. The k^{th}-order inversion model has as sufficient statistics all the number of occurrences of each inversion of order less than or equal to k. McCullagh shows that the number of inversions of order $k-1$ involving exactly k of the objects is

$$\binom{m}{k} k! \left(\frac{1}{2!} - \frac{1}{3!} + \cdots \pm \frac{1}{k!} \right).$$

Chapter 3

Testing and Estimation

3.1 Likelihood methods

Likelihood methods such as maximum likelihood estimation (MLE) and likelihood ratio testing (LRT) have proven to be extremely useful for inference in general models. There are many examples in which likelihood methods fail, or can be much improved, or are extremely difficult to implement computationally. However, for most of the ranking models we are considering, they work remarkably well. This section gives an introduction aimed at rank models.

The observations take values in the finite set $\mathcal{A}_m$ with M elements. For a typical ranking model, $\mathcal{A}_m$ is the set of rankings $\mathcal{P}_m$ or orderings $\mathcal{L}_m$, and $M = m!$. There is a corresponding vector $\boldsymbol{p}$ of probabilities, where p_a is the probability the observation is element a. The space of $\boldsymbol{p}$ is the M-dimensional simplex

$$\mathcal{S}_M = \{\boldsymbol{p} \in \mathbb{R}^M \,|\, p_a > 0 \text{ for all } a \in \mathcal{A}_m, \text{ and } \sum_{a \in \mathcal{A}_m} p_a = 1\}.$$

For an iid sample $W_1, \ldots, W_n$, we have

$$\boldsymbol{K} \sim \text{Multinomial}_M(n, \boldsymbol{p}), \text{ where } K_a = \#\{W_i = a\}.$$

3.1.1 Regular models

A model for the multinomial distribution is given by specifying a subset of $\mathcal{S}_M$ to which $\boldsymbol{p}$ is restricted. We will be mainly concerned with **regular** models, which will be defined in this section.

Start with a parametrization of the model, that is, assume there exists a p-dimensional vector $\boldsymbol{\theta}$ with space Θ and a function

$$\boldsymbol{p}(\boldsymbol{\theta}) : \Theta \longrightarrow \mathcal{S}_M,$$

the model allows $\boldsymbol{p}$ to have the range

$$\{\boldsymbol{p}(\boldsymbol{\theta}) \in \mathcal{S}_M \,|\, \boldsymbol{\theta} \in \Theta\}.$$

The following regularity conditions are required of a regular model.

Regularity conditions.

1. Identifiability: If $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, then $\boldsymbol{p}(\boldsymbol{\theta}_1) \neq \boldsymbol{p}(\boldsymbol{\theta}_2)$ for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$.

2. Θ is an open subset of $\mathbb{R}^p$.

3. As functions of $\boldsymbol{\theta}$, the $p_a(\boldsymbol{\theta})$'s have that all first, second and third (mixed) partial derivatives are continuous in $\boldsymbol{\theta}$.

4. For every a and $\boldsymbol{\theta}$, $p_a(\boldsymbol{\theta}) > 0$.

The fourth condition is actually redundant since $\mathcal{S}_M$ requires the p_a's to be positive. These conditions are automatically satisfied by regular exponential families as in Section 2.6, and in fact by almost all the models except the mixture models for unfolding scales. Thurstonian models may also violate the assumptions if the distributions of $\boldsymbol{Z}$ are irregular.

3.1.2 The likelihood function and Fisher information

The likelihood function L is a function of $\boldsymbol{\theta}$ for fixed value of the data $\boldsymbol{k}$. It is proportional to the density, i.e.,

$$L(\boldsymbol{\theta}\,;\,\boldsymbol{k}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(w_i) = \prod_{a \in \mathcal{A}_m} p_a(\boldsymbol{\theta})^{k_a}. \tag{3.1.1}$$

The loglikelihood is

$$l(\boldsymbol{\theta}\,;\,\boldsymbol{k}) = \log(L(\boldsymbol{\theta}\,;\,\boldsymbol{k})) = \sum_{a \in \mathcal{A}_m} k_a \log(p_a(\boldsymbol{\theta})). \tag{3.1.2}$$

The values of L are supposed to measure how "likely" various values of $\boldsymbol{\theta}$ are in light of the particular data $\boldsymbol{k}$. If $L(\boldsymbol{\theta}_1\,;\,\boldsymbol{k}) = 2L(\boldsymbol{\theta}_2\,;\,\boldsymbol{k})$, then $\boldsymbol{\theta}_1$ is twice as likely to be the true $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_2$. The **maximum likelihood estimate** (MLE) of $\boldsymbol{\theta}$ based on $\boldsymbol{k}$ is the value of $\boldsymbol{\theta} \in \Theta$ that maximizes the likelihood. Denote it by $\widehat{\boldsymbol{\theta}}$. It may be that there is no maximum, or the maximum is not unique, or the maximum lies outside of Θ. In such cases, the MLE does not exist.

If the MLE does exist, then for a regular model, the derivatives of l are zero at $\widehat{\boldsymbol{\theta}}$, that is,

$$\nabla l(\widehat{\boldsymbol{\theta}}\,;\boldsymbol{k}) = \mathbf{0}\,;\ (\nabla l(\widehat{\boldsymbol{\theta}}\,;\boldsymbol{k}))_j = \frac{\partial}{\partial\theta_j} l(\boldsymbol{\theta}\,;\boldsymbol{k}) \ \text{ for } \ 1 \leqslant j \leqslant p. \tag{3.1.3}$$

The first equation in (3.1.3) is termed the **likelihood equation(s)**.

Once the $\widehat{\boldsymbol{\theta}}$ is obtained, it is also of interest to ascertain how confident one should be in it as an estimate. The likelihood approach is to look at the likelihood for values of $\boldsymbol{\theta}$ near $\widehat{\boldsymbol{\theta}}$. If the likelihood tends to drop off precipitously, then one is quite sure of the estimate. If the likelihood is relatively flat around the MLE, then there are many other estimates which are nearly as likely as the MLE, hence not as much reliance can be put in the estimate. One way to measure how quickly the likelihood falls off is to look at the second derivative matrix. If it is large negative, then the likelihood has a relatively large amount of information. If it is near 0, then the likelihood has little information. Thus the bfseries observed Fisher information is defined to be

$$\widehat{\boldsymbol{I}}_n(\widehat{\boldsymbol{\theta}}) = -\left\{ \frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta};\boldsymbol{k})\big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right\}_{i,j=1}^{p}. \tag{3.1.4}$$

(It is the matrix of partial second derivatives.)

The $\widehat{\boldsymbol{I}}_n$ is a data-dependent measure of information. The corresponding population quantity is the **expected** Fisher information, which is the information about $\boldsymbol{\theta}$ one expects from a sample. It is usually called the Fisher information. The Fisher information for one observation is defined to be

$$\boldsymbol{I}_1(\boldsymbol{\theta}) = -\left\{ E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta}\,;W_1) \right] \right\}_{i,j=1}^{p}, \tag{3.1.5}$$

and the information for the sample of n observations is

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = \left\{ E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta}\,;\boldsymbol{K}) \right] \right\}_{i,j=1}^{p} = n\boldsymbol{I}_1(\boldsymbol{\theta}).$$

3.1.3 Maximum likelihood estimation

In the previous section, the MLE was defined to be the value of $\boldsymbol{\theta}$ which maximizes the likelihood (3.1.1). In a regular model, if it exists it satisfies the likelihood equations (3.1.3). In this section, we present asymptotic results for the MLE, asymptotic as $n \to \infty$. In order to state the result, we will assume the following additional regularity conditions.

5. For any $\boldsymbol{\theta}$, $P_{\boldsymbol{\theta}}$[The MLE exists] $\to 1$ as $n \to \infty$.

6. For any $\boldsymbol{\theta}$, the Fisher information $\boldsymbol{I}_1(\boldsymbol{\theta})$ is finite and invertible.

The main result is the next theorem. Proofs of this and similar theorems can be found in Rao (1973) and Lehmann (1983).

Theorem 3.1.1. *Suppose the regularity conditions 1 through 6 hold. Then as* $n \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}(\boldsymbol{K}) - \boldsymbol{\theta}) \longrightarrow N(\mathbf{0}, \boldsymbol{I}_1^{-1}(\boldsymbol{\theta})),$$

where $\widehat{\boldsymbol{\theta}}$ *is the MLE when it exists, and any arbitrary value when it does not. Also,*

$$\frac{\widehat{\boldsymbol{I}}_n(\widehat{\boldsymbol{\theta}})}{n} \longrightarrow \boldsymbol{I}_1(\boldsymbol{\theta}) \text{ in probability.}$$

The value of the theorem lies in the fact that for any regular model, the distribution of the MLE can be approximated by a normal with covariance matrix being the inverse of the observed Fisher information, i.e.,

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{K}) \approx N(\boldsymbol{\theta}, \widehat{\boldsymbol{I}}_n^{-1}(\widehat{\boldsymbol{\theta}})).$$

If you think about this result, it is amazing.

3.1.4 Likelihood ratio tests

It often (always?) happens that one has more than one model in mind, and wishes to find which is best. One model may be the same as another but for a restriction on the parameter, such as $\theta_1 = 0$, or it may be that the models have different parametrizations. Here, we present a result for testing two nested models, nested in the sense that the set of allowable $\boldsymbol{p}$'s for one is strictly contained in that for the other. Specifically, we assume that we have two regular models, both satisfying conditions 1 through 6. Let the smaller one have parametrization $\boldsymbol{\tau} \in \mathcal{T}$, where $\boldsymbol{\tau}$ is $q \times 1$, and corresponding probabilities $\boldsymbol{q}(\boldsymbol{\tau})$, and the larger one be as in (1.3.1), where $q < p$. The containment is

$$\{\boldsymbol{q}(\boldsymbol{\tau}) \mid \boldsymbol{\tau} \in \mathcal{T}\} \subset \{\boldsymbol{p}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \Theta\}.$$

The hypothesis test of interest tests

$$H_0 : \text{The } \boldsymbol{\tau}\text{-model holds} \quad \textit{versus} \quad H_A : \text{The } \boldsymbol{\theta}\text{-model holds.} \quad (3.1.6)$$

Let l_0 and l_A be the respective loglikelihoods (3.1.2) for the two hypothesized models. The likelihood ratio test compares the best likelihood under the null model to the best under the alternative model. The latter will always be at least as large than the former since there is a larger family of distributions. The question is how much larger must it be for us to reject the null model in favor of the alternative? It is convenient to consider the **likelihood ratio statistic**

$$\mathrm{LRS}_n(\boldsymbol{k}) \equiv 2(l_A(\widehat{\boldsymbol{\theta}};\, \boldsymbol{k}) - l_0(\widehat{\boldsymbol{\tau}};\, \boldsymbol{k})).$$

Theorem 3.1.2. *Under the above assumptions, if* H_0 *in* (3.1.6) *holds, then as* $n \to \infty$,

$$LRS_n(\boldsymbol{K}) \;\longrightarrow\; \chi^2_{p-q} \textit{ in distribution.}$$

From (3.1.2), it is easy to show that

$$\begin{aligned} \mathrm{LRS}_n(\boldsymbol{k}) &= 2 \sum_{a \in \mathcal{A}_m} k_a \log\left(\frac{p_a(\widehat{\boldsymbol{\theta}})}{q_a(\widehat{\boldsymbol{\tau}})}\right) \\ &= 2 \sum \mathrm{OBS} \cdot \log\left(\frac{\mathrm{EXP}_A}{\mathrm{EXP}_0}\right). \end{aligned} \tag{3.1.7}$$

The last expression is a common mnemonic, where OBS refers to the observed counts, and EXP_0 and EXP_A to the expected counts under the null and alternative hypotheses, respectively.

Even if one has only one model to consider, there are two particular hypothesis tests one usually performs. Suppose the model has the $\boldsymbol{\theta}$-parametrization. The first is to see whether there is "anything going on," that is, to see if it is possible that the uniform model holds. Then the H_0 model in (3.1.6) has q=0, the only allowable distribution being the uniform, $q_a = 1/M$ for all a. Thus the statistic (3.1.7) is

$$\mathrm{LRS}_n(\boldsymbol{k}) = 2 \sum \mathrm{OBS} \cdot \log\left(\frac{\mathrm{EXP}_A}{\mathrm{EXP}_0}\right), \; \mathrm{EXP}_0 \equiv \frac{n}{M}.$$

If it is not sufficiently large ($> \chi^2_{p,\alpha}$) then one cannot conclude that the model is any better than the uniform. The other important test is the goodness-of-fit test. That is, is there anything going on that the model does not detect? Now the $\boldsymbol{\theta}$-model is the H_0 model, and the H_A model is the saturated model. The saturated model has dimension $M - 1$ since it can be parametrized by $p_1, \ldots, p_{M-1}$. It can be shown that the MLE of p_a under the saturated model is k_a/n. Thus the goodness-of-fit test statistic is asymptotically χ^2_{M-1-p} under H_0, and can be written

$$\mathrm{LRS}_n(\boldsymbol{k}) = 2 \sum \mathrm{OBS} \cdot \log\left(\frac{\mathrm{OBS}}{\mathrm{EXP}}\right).$$

An alternative statistic, which has the same asymptotic null distribution, is the Pearson chi-squared $\sum(\mathrm{OBS}-\mathrm{EXP})^2/\mathrm{EXP}$, which may be more familiar. They are equally fine.

Nested sequences of models are important, especially when using hierarchical systems of models. For example, one might consider the sequence of models

$$\begin{aligned}&\text{Uniform} \subset \text{Mallows}' \ \theta \subset \text{Bradley-Terry}\\ &\qquad\qquad \subset \text{Babington Smith} \subset \text{Saturated}. \end{aligned} \tag{3.1.8}$$

The number of parameters for these are 0, 1, $m-1$, $\binom{m}{2}$ and $m!-1$, respectively. A nice property of the LRSs for testing pairs of models in a sequence is additivity: If $\mathrm{LRS}(M_i, M_j)$ is the LRS for testing model M_i versus model M_j, then

$$\mathrm{LRS}(M_1, M_3) = \mathrm{LRS}(M_1, M_2) + \mathrm{LRS}(M_2, M_3).$$

Thus for the sequence in (3.1.8), one can decompose the overall lack of uniformity in the data, i.e., the LRS for testing uniform versus saturated, into pieces which measure the additional amount of fit explained as one moves along the sequence:

$$\mathrm{LRS}(U, S) = \mathrm{LRS}(U, M) + \mathrm{LRS}(M, BT) + \mathrm{LRS}(BT, BS) + \mathrm{LRS}(BS, S).$$

Of course, when one performs a number of tests, one has to worry about the multiple comparisons problem.

3.2 Exponential families

A regular exponential family as in Section 2.6 automatically satisfies all the regularity conditions in Section 3.1.1. Here, a p-dimensional exponential family will have sufficient statistic $\boldsymbol{t}(w_i)$ for each observation, so that the density is

$$f_{\boldsymbol{\theta}}(w_i) = f_0(w_i)e^{\boldsymbol{\theta}'\boldsymbol{t}(w_i)-\psi(\boldsymbol{\theta})}, \tag{3.2.1}$$

where $\boldsymbol{\theta}'\boldsymbol{t} = \sum \theta_i t_i$, f_0 is a (null) density for W_i, and ψ is whatever is needed for the density to sum to 1, i.e.,

$$\psi(\boldsymbol{\theta}) = \log\left(\sum_{a\in\mathcal{A}_m} f_0(a)e^{\boldsymbol{\theta}'\boldsymbol{t}(a)}\right).$$

Differentiating ψ shows that

$$\boldsymbol{\beta}(\boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}}(\boldsymbol{t}(W_i)) = \left\{\frac{\partial\psi(\boldsymbol{\theta})}{\partial\theta_j}\right\}_{j=1}^{p}$$

and

$$\Sigma(\boldsymbol{\theta}) \equiv \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{t}(W_i)) = \left\{ \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right\}_{j,k=1}^{p}. \tag{3.2.2}$$

Thus the mean and the covariance matrix of the sufficient statistic vector can be found by differentiating rather than integrating.

Now from (3.1.1) and (3.2.1), the loglikelihood for the data is

$$\begin{aligned} l(\boldsymbol{\theta}\,;\,\boldsymbol{k}) &= \sum_{i=1}^{n} \boldsymbol{\theta}' \boldsymbol{t}(w_i) - n\psi(\boldsymbol{\theta}) + C \\ &= \boldsymbol{\theta}' \boldsymbol{s}(\boldsymbol{k}) - n\psi(\boldsymbol{\theta}) + C, \end{aligned} \tag{3.2.3}$$

where $\boldsymbol{s}$ is the vector of sufficient statistics for the sample, $\boldsymbol{s}(\boldsymbol{k}) = \sum \boldsymbol{t}(w_i)$, and $C(= \sum \log(f_0(w_i)))$ is a constant independent of $\boldsymbol{\theta}$. The likelihood equations (3.1.3) are thus $0 = \nabla l(\widehat{\boldsymbol{\theta}}, \boldsymbol{k}) = \boldsymbol{s} - n\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}})$, or

$$\boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}) = \frac{\boldsymbol{s}}{n}. \tag{3.2.4}$$

That is, the MLE of $\boldsymbol{\theta}$ is that value for which the population mean of the sufficient statistic equals the sample mean of the $\boldsymbol{t}(w_i)$'s. In a way this result should not be too surprising since in Section 1.8 it was noted that the exponential family is chosen so as to satisfy (3.2.4).

Next, to find the observed and expected Fisher information, use (3.1.4) and (3.1.5) on (3.2.3). The first term drops out since it is linear in $\boldsymbol{\theta}$, and what are left are second derivatives of $-\psi$, so that from (3.1.4) and (3.2.2),

$$\widehat{\boldsymbol{I}}_n(\widehat{\boldsymbol{\theta}}) = n\Sigma(\widehat{\boldsymbol{\theta}}) \text{ and } \boldsymbol{I}_1(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta}).$$

The dual central limit theorems for exponential families are

$$\sqrt{n}\left(\frac{\boldsymbol{S}}{n} - \boldsymbol{\beta}(\boldsymbol{\theta})\right) \longrightarrow N(\boldsymbol{0}, \Sigma(\boldsymbol{\theta})) \text{ and } \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \longrightarrow N(\boldsymbol{0}, \Sigma^{-1}(\boldsymbol{\theta})).$$

Those familiar with the Δ-method will realize these are the same theorem.

3.3 Finding the MLE

It is typical that the likelihood equations (3.1.3) cannot be solved in closed form, so that some numerical methods must be employed. A very appropriate method for current purposes is the Newton-Raphson method. In one dimension, it is derived as follows. Suppose the objective is to find $\widehat{\theta}$ so that $h(\widehat{\theta} = 0$ for a given function h. Expand h around $\widehat{\theta} = \theta$ in a Taylor Series:

$$h(\widehat{\theta} = h(\theta) + (\widehat{\theta} - \theta)h'(\theta) + \text{Remainder}. \tag{3.3.1}$$

If the remainder is small, then since $h(\widehat{\theta} = 0$, solving for $\widehat{\theta}$ gives

$$\widehat{\theta} \approx \theta - \frac{h(\theta)}{h'(\theta)}.$$

The Newton-Raphson method starts by guessing a value for $\widehat{\theta}$, say θ_1. If $h(\theta_1) = 0$ then $\widehat{\theta} = \theta_1$. If not, use (3.3.1) with $\theta = \theta_1$ to find a better guess for $\widehat{\theta}$. Call it θ_2. One continues until $h(\theta_i)$ is close enough to 0, where for $i > 2$, $\theta_i = \theta_{i-1} - h(\theta_{i-1})/h'(\theta_{i-1})$. It could be that the sequence θ_i does not converge, or it converges to a root of h other than $\widehat{\theta}$. Books on numerical analysis contain numerous conditions for convergence to be guaranteed. The multivariate version has $\boldsymbol{\theta}$ and h being p-dimensional, so that the h' becomes a matrix of derivatives, and the equation is

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - [h'(\boldsymbol{\theta}_{i-1})]^{-1} h(\boldsymbol{\theta}_{i-1}).$$

In finding MLE's, the h is $\nabla l(\boldsymbol{\theta}, \boldsymbol{k})$, the vector of derivatives of the likelihood function (3.1.3). Note that the matrix $[h'(\widehat{\boldsymbol{\theta}})]^{-1}$ is $-\widehat{\boldsymbol{I}}_n(\widehat{\boldsymbol{\theta}})$ of (3.1.4). The MLE $\widehat{\boldsymbol{\theta}}$ is found by iterating

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \widehat{\boldsymbol{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}) \nabla l(\boldsymbol{\theta}_{i-1}, \boldsymbol{k}).$$

In the exponential family case, this simplifies further to

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \Sigma^{-1}(\boldsymbol{\theta}_{i-1}) \left(\frac{\boldsymbol{S}}{n} - \boldsymbol{\beta}(\boldsymbol{\theta}_{i-1}) \right).$$

A possible drawback to the Newton-Raphson method occurs when p is large, so that $\widehat{\boldsymbol{I}}_n$ may be difficult to invert. There are many modifications and other possible methods. In the next four sections we give some methods which are useful in specialized contexts.

3.4 Iterative proportional fitting

Iterative proportion fitting is a method to find the MLE's of $\boldsymbol{p}(\boldsymbol{\theta})$ rather than the $\boldsymbol{\theta}$ directly. It avoids any matrix inversions. It works only in exponential family models for which the sufficient statistics are categorical in nature.

A **categorization** of the sample space $\mathcal{A}_m$ is a partition $\boldsymbol{I} = (\mathcal{J}_1, \dots, \mathcal{J}_K)$ of the elements $a \in \mathcal{A}_m$. That is, the subsets $\mathcal{J}_k$'s are nonempty and disjoint, and their union is the set $\mathcal{A}_m$. (These $\mathcal{J}_k$'s are different than those for orthogonal contrast models in Section 1.6.3. There, the $\mathcal{J}_k$'s were groups of objects; here, they are groups of rankings.) The statistic $\boldsymbol{S}_{\boldsymbol{I}}(\boldsymbol{k})$ corresponding to the categorization $\boldsymbol{I}$ is defined by

$$\boldsymbol{S}_{\boldsymbol{I}}(\boldsymbol{k}) = \left(\sum_{a \in \mathcal{J}_1} k_a, \cdots, \sum_{a \in \mathcal{J}_K} k_a \right),$$

which are the numbers of observations in each of the categories $\mathcal{J}_k$.

Categorization assumption. We assume we have L categorizations, $\boldsymbol{I}_1, \dots, \boldsymbol{I}_L$. The model is a regular exponential family with sufficient statistic being the **reduced set** $\boldsymbol{S}$ of counts $(\boldsymbol{S}_{\boldsymbol{I}_1}, \dots, \boldsymbol{S}_{\boldsymbol{I}_L})$.

By "reduced set" we mean the smallest subset of the counts such that all the counts can be obtained from the smaller set and the number n by addition or subtraction. The reason is that without reduction, the exponential family will not be identifiable.

Example. In a 2×3 contingency table, we have $M = 6$. Number the cells:

1	2	3
4	5	6

Consider the two categorizations

$$\boldsymbol{I}_{row} = (\{1,2,3\},\{4,5,6\} \text{ and } \boldsymbol{I}_{column} = (\{1,4\},\{2,5\},\{3,6\}.$$

The total set of counts is

$$(n_1 + n_2 + n_3,\ n_4 + n_5 + n_6,\ n_1 + n_4,\ n_2 + n_5,\ n_3 + n_6).$$

However, we can obtain all those counts from the reduced set $\boldsymbol{S} = (n_1 + n_2 + n_3, n_1 + n_4, n_2 + n_5)$. The model resulting from $\boldsymbol{S}$ is the model with row and column independence.

Suppose the categorization assumption holds. The MLE of $\boldsymbol{p}$ can be found by starting with an $M \times 1$ vector of 1's and iteratively changing the vector so that it conforms to the sufficient statistics. More specifically, let $\boldsymbol{\nu}$ be any $M \times 1$ vector. A new vector $\boldsymbol{\nu}^*$ is obtained using the categorization $\boldsymbol{I}$ as follows:

$$\boldsymbol{\nu}^* = \text{IP}(\boldsymbol{\nu}; \boldsymbol{I})$$

where for each $\mathcal{J}_k$ in $\boldsymbol{I}$,

$$\boldsymbol{k}^*_a = \boldsymbol{k}_a \frac{\sum_{j\in\mathcal{J}_k} k_j}{\sum_{j\in\mathcal{J}_k} \nu_j} = \boldsymbol{k}_a \frac{[\boldsymbol{S_I}(\boldsymbol{k})]_k}{[\boldsymbol{S_I}(\boldsymbol{\nu})]_k} \text{ for } a \in \mathcal{J}_k.$$

Thus $\boldsymbol{S_I}(\boldsymbol{\nu}^*) = \boldsymbol{S_I}(\boldsymbol{k})$, i.e., the counts of $\boldsymbol{\nu}^*$ for the categorization $\boldsymbol{I}$ are equal to those of $\boldsymbol{k}$. The iterative proportional fitting algorithm starts with $\boldsymbol{\nu}_0 = (1, 1, \ldots, 1)$. Then

$$\boldsymbol{\nu}_k = \mathrm{IP}(\boldsymbol{\nu}_{k-1}\,;\, \boldsymbol{I}_k),\ k = 1, 2, \ldots, K. \tag{3.4.1}$$

Next, set $\boldsymbol{\nu}_0$ to be the final vector $\boldsymbol{\nu}_K$, and repeat (3.4.1). This process is repeated until the applications of (3.4.1) no longer change the vector appreciably. Let $\boldsymbol{\nu}$ be the final vector. Then the MLE's are given by

$$\boldsymbol{k}_a = np_a(\widehat{\boldsymbol{\theta}}),\ a \in \mathcal{A}_m.$$

If one is interested in the MLE $\widehat{\boldsymbol{\theta}}$ and the Fisher information, one has to perform further computations. However, if only testing is of interest, the LRS's can be found directly from the $\boldsymbol{\nu}$'s. See (3.1.7).

3.5 The EM algorithm

Dempster, Laird, and Rubin (1977) present the general EM algorithm is. Here we will give just the relevant results for fitting latent class models. See Croon (1989), too. We assume we have a mixture of exponential family models, that is, there are G groups in the population, and

$$\begin{aligned} f(a\,;\, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{q}) &= \sum_{g=1}^{G} P[O_g] \times P[W = a \,|\, O_g] \\ &= f_0(a) \sum_{g=1}^{G} q_g e^{\boldsymbol{\theta}_g' \boldsymbol{\tau}_g(a) - \psi_g(\boldsymbol{\theta}_g)}, \end{aligned}$$

where

$$P[O_g] = q_g \ \ (O_g \text{ means ``Observation is in group } g\text{''})$$

and

$$P[W = a \,|\, O_g] = f_0(a) e^{\boldsymbol{\theta}_g' \boldsymbol{\tau}_g(a) - \psi_g(\boldsymbol{\theta}_g)}.$$

The likelihood function is

$$L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \boldsymbol{q}\,;\, \boldsymbol{k}) = \prod_{a\in\mathcal{A}_m} \left(f_0(a) \sum_{g=1}^{G} q_g e^{\boldsymbol{\theta}_g' \boldsymbol{\tau}_g(a) - \psi_g(\boldsymbol{\theta}_g)} \right)^{k_a}. \tag{3.5.1}$$

Because the summation is inside the product, the nice summing which occurs in the exponent of an exponential family does not occur here. The number of parameters can be large, $pG + G - 1$ if each $\boldsymbol{\theta}_g$ is of dimension p, and the formulas for the first and second derivatives can be reasonably complicated.

The idea of the algorithm is that if we knew which observations were from which group, we could just fit the individual models using maximum likelihood (the "M" step). On the other hand, if we knew the values of the parameters $\boldsymbol{\theta}_k$, we could estimate how many observations come from each group (the "E" step). Details follow.

For $a \in \mathcal{A}_m$ and $g = 1, \ldots, G$, let

$$\begin{aligned} h(g\,;\,a) &= P[O_g \,|\, W = a] \\ &= \frac{P[W = a \,|\, O_g]P[O_g]}{\sum_{l=1}^{G} P[W = a \,|\, O_l]P[O_l]} \\ &= \frac{e^{\boldsymbol{\theta}_g' \boldsymbol{\tau}_g(a) - \psi_g(\boldsymbol{\theta}_g)} q_g}{\sum_{l=1}^{G} e^{\boldsymbol{\theta}_l' \boldsymbol{\tau}_l(a) - \psi_l(\boldsymbol{\theta}_l)} q_l}. \end{aligned} \tag{3.5.2}$$

Now perform the steps:

0. Guess values of the $\boldsymbol{\theta}_g$'s and $\boldsymbol{q}$.

1. (E) Use (3.5.2) to estimate $h(g\,;\,a)$ for all $g = 1, \ldots, G$ and a from the current estimates of the $\boldsymbol{\theta}_g$'s and $\boldsymbol{q}$.

2. (M) For each g, find the new estimate of $\boldsymbol{\theta}_g$ by using the MLE from the g^{th} exponential family model with sufficient statistic

$$\begin{aligned} \boldsymbol{S}_g &= \sum_{i=1}^{n} h(g\,;\,w_i)\boldsymbol{\tau}_g(w_i) \\ &= \sum_{a \in \mathcal{A}_m} h(g\,;\,a)k_a \boldsymbol{\tau}_g(a). \end{aligned}$$

   This is like the usual sufficient statistic, but the observations are weighted by the probability they are in the particular group.

3. Find the new estimate of $\boldsymbol{q}$ by using

$$q_g = P[O_g] = \sum_{a \in \mathcal{A}_m} P[O_g \,|\, W = a]P[W = a] \approx \sum_{a \in \mathcal{A}_m} h(g\,;\,a)\frac{k_a}{n}.$$

4. Check (3.5.1) (or its log) using the current values of the parameters. If it is high enough, stop. Otherwise, go back to step 1 with the new estimates.

Each iteration produces a value of the likelihood. These values will increase, meaning each new set of estimates is more likely than the previous. One stops in step 4 if the increase is small enough. It can happen that the increases are small, but do not decrease very quickly. The algorithm may have to be repeated many (200, 500, $\ldots$) times. Thus it is valuable when the individual steps are relatively easy, while a full-blown Newton-Raphson is very difficult. One of the most attractive properties of the EM algorithm is that it is reasonably easy to program.

3.6 Approximation for Thurstonian models

In order to find MLE's in a Thurstonian model, one must be able to find the probability for each particular ordering,

$$P[Z_{x_1} < \cdots < Z_{x_m}], \;\; \boldsymbol{x} \in \mathcal{L}_m.$$

This probability in principle requires an m-fold integration, although by subtraction one can reduce it to an $(m-1)$-fold integral. The only distribution for which these probabilities can be obtained easily is the Gumbel distribution as in Section 2.2. In particular, the normal presents a rather daunting problem. There are algorithms for such normal integrals. The length of time it takes to perform the integrals can prove prohibitive if m is at all large (10?).

An alternative estimation scheme, which is not equivalent to MLE, uses only the paired-comparison data. For a sample, let $N_{l_i l_j}$ be the number of observations in which object l_i is preferred to object l_j. Then for a Thurstonian model (1.4.2), from (2.2.4) we have that

$$E\left[\frac{N_{l_i l_j}}{n}\right] = D(-(\mu_{l_i} - \mu_{l_j})),$$

where D is the distribution function of $U_1 - U_2$, U_1 and U_2 being iid with density g. Critchlow et al. (1991) note that therefore

$$E\left[D^{-1}\left(\frac{N_{l_i l_j}}{n}\right)\right] \approx -(\mu_{l_i} - \mu_{l_j}) \tag{3.6.1}$$

and the μ_{l_i}'s can be estimated directly from the statistics in the expectation in (3.6.1).

Let $\boldsymbol{K}^P$ be the $\binom{m}{2} \times 1$ vector of $N_{l_i l_j}$'s for $i < j$, $\boldsymbol{U}$ be the corresponding vector of $D^{-1}(N_{l_i l_j}/n)$'s, $\boldsymbol{\delta}$ be the corresponding vector of $\mu_{l_i} - \mu_{l_j}$'s, and $D(-\boldsymbol{\delta})$ the vector of $D(-(\mu_{l_i} - \mu_{l_j}))$'s. The central limit theorem shows that

$$\sqrt{n}\left(\frac{\boldsymbol{K}^P}{n} - D(-\boldsymbol{\delta})\right) \longrightarrow N(0, \Omega(\boldsymbol{\mu}))$$

for some covariance matrix Ω to be given below. The Δ-method then can be used to show that, as long as D has a continuous derivative d,

$$\sqrt{n}(\boldsymbol{U}+\boldsymbol{\delta}) \longrightarrow N(0,\Sigma(\boldsymbol{\mu}));\ \Sigma(\boldsymbol{\mu}) = \mathrm{Diag}(1/d(\boldsymbol{\delta}))\Omega(\boldsymbol{\mu})\mathrm{Diag}(1/d(\boldsymbol{\delta})), \tag{3.6.2}$$

where $\mathrm{Diag}(1/d(\boldsymbol{\delta}))$ is the diagonal matrix with diagonal elements

$$1/d(-(\mu_{l_i}-\mu_{l_j})).$$

Now (3.6.2) can be used to set up a normal linear model which approximates the distribution of $\boldsymbol{U}$:

$$\boldsymbol{U} \approx \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{E};\ \boldsymbol{E} \sim N(0,(1/n)\Sigma(\boldsymbol{\mu})),$$

where $\boldsymbol{\mu} = (\mu_{l_1},\ldots,\mu_{l_m})'$, and X is a $\binom{m}{2} \times m$ matrix of 0's and ±1's representing the differences. E.g., the matrix for $m = 4$ is next.

Pair ij	$\boldsymbol{X}$ − matrix			
12	−1	1	0	0
13	−1	0	1	0
14	−1	0	0	1
23	0	−1	1	0
24	0	−1	0	1
34	0	0	−1	1

Now estimates of $\boldsymbol{\mu}$ can be found using least squares or weighted least squares. The matrix $\boldsymbol{X}$ is not of full rank since the columns sum to $\mathbf{0}$. This property is due to the indeterminacy of the means μ_{l_i} in that adding the same constant to all gives exactly the same model. For estimation purposes one must place a constraint on the parameters. The easiest is to set $\mu_{l_m} = 0$, so that the final column in $\boldsymbol{X}$ can be dropped. Let $\boldsymbol{\mu}^* = (\mu_{l_1},\ldots,\mu_{l_{m-1}})$ and $\boldsymbol{X}^*$ be $\boldsymbol{X}$ without the final column. Then the unweighted estimate of $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}}^* = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{U}.$$

It is consistent, but not necessarily very efficient. Note that computationally this estimate is easy as long as the D function can be inverted. The estimate can be improved by using the weighted least squares estimate; however, the weights depend on $\Sigma(\boldsymbol{\mu})$, which is a function of the unknown $\boldsymbol{\mu}$. Iteratively reweighted least squares (IRWLS) is appropriate. An initial estimate $\boldsymbol{\mu}_0$ is needed, say from unweighted least squares, or just $\boldsymbol{\mu}_0 = \mathbf{0}$. Then $\boldsymbol{\mu}_i$ is obtained from $\boldsymbol{\mu}_{i-1}$ by setting

$$\boldsymbol{\mu}_i^* = (\boldsymbol{X}^{*\prime}\Sigma(\boldsymbol{\mu}_{i-1})^{-1}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\Sigma(\boldsymbol{\mu}_{i-1})^{-1}\boldsymbol{U}. \tag{3.6.3}$$

Objects	Covariance
r, s, t, u all distinct	0
$r = t, s \neq u$	$P[U_1 - U_2 < -\delta_{rs} \ \& \ U_1 - U_3 < -\delta_{ru}] - p_{rs}p_{ru}$
$r \neq t, s = u$	$P[U_1 - U_3 < -\delta_{rs} \ \& \ U_2 - U_3 < -\delta_{t}s] - p_{rs}p_{ts}$
$r \neq u, s = t$	$P[U_1 - U_2 < -\delta_{rs} \ \& \ U_2 - U_3 < -\delta_{su}] - p_{rs}p_{su}$
$r = t, s = u$	$p_{rs}(1 - p_{rs})$

Table 3.1: Covariances (3.6.4) for paired comparisons in the Thurstonian model.

Equation (3.6.3) is iterated until the estimates no longer change much. Then the final value is $\widehat{\boldsymbol{\mu}}$. The approximation to the distribution of $\widehat{\boldsymbol{\mu}}^*$ is

$$\widehat{\boldsymbol{\mu}}^* \approx N(\boldsymbol{\mu}^*, (1/n)(\boldsymbol{X}^{*\prime}\Sigma(\widehat{\boldsymbol{\mu}})^{-1}\boldsymbol{X}^*)^{-1}.$$

The iteratively reweighted least squares estimate should be better than the unweighted version, but may not be as efficient as the MLE. The reason is that the MLE uses all the information in the rankings, while the IRWLS estimate only uses the paired comparison data. It is an open problem exactly how much efficiency is lost.

3.6.1 The covariance matrices Ω and Σ

The matrix Ω contains the variances and covariances of indicator variables of paired comparisons, i.e., the $(rs, tu)^{\text{th}}$ element is

$$\mathrm{Cov}(I[Z_r < Z_s], I[Z_t < Z_u]), \tag{3.6.4}$$

for not necessarily distinct objects r, s, t, u. The covariances differ depending on which objects among r, s, t, u are the same. Table 3.1 gives the necessary cases, where

$$p_{rs} = P[Z_r < Z_s] = P[r \text{ is preferred to } s] = D(-\delta_{rs}); \ \delta_{rs} = \mu_r - \mu_s. \tag{3.6.5}$$

The U_1, U_2 and U_3 are independent with density g. The probabilities for the normal and Gumbel distributions are next.

Normal. Now the U_i's are iid $N(0,1)$, so that we need univariate and bivariate normal probabilities. In (3.6.5), D is the $N(0,2)$ distribution, hence $D(-\delta_{rs}) = \Phi(-\delta_{rs}/\sqrt{2})$, where Φ is the N(0,1) distribution function. For a bivariate normal (Z_1, Z_2) with means 0, variances 1, and correlation ρ, let $BVN(z_1, z_2\,;\,\rho)$ denote $P[Z_1 \leqslant z_1, Z_2 \leqslant z_2]$. For the

"$r = t, s \neq u$" probabilities, we have that $(U_1 - U_2, U_1 - U_3)$ is bivariate normal with means 0, variances 2, and correlation $\frac{1}{2}$. Hence

$$P[U_1 - U_2 < -\delta_{rs} \ \& \ U_1 - U_3 < -\delta_{ru}] = \mathrm{BVN}(-\tfrac{1}{\sqrt{2}}\delta_{rs}, -\tfrac{1}{\sqrt{2}}\delta_{ru} \,;\, \tfrac{1}{2}).$$

In the same way, for $r \neq t, s = u$,

$$P[U_1 - U_3 < -\delta_{rs} \ \& \ U_2 - U_3 < -\delta_{ts}] = \mathrm{BVN}(-\tfrac{1}{\sqrt{2}}\delta_{rs}, -\tfrac{1}{\sqrt{2}}\delta_{ts} \,;\, \tfrac{1}{2}).$$

For $r \neq u, s = t$, the correlation between $U_1 - U_2$ and $U_2 - U_3$ is $-\frac{1}{2}$, so that

$$P[U_1 - U_2 < -\delta_{rs} \ \& \ U_2 - U_3 < -\delta_{su}] = \mathrm{BVN}(-\tfrac{1}{\sqrt{2}}\delta_{rs}, -\tfrac{1}{\sqrt{2}}\delta_{su} \,;\, -\tfrac{1}{2}).$$

It is fairly easy to find computer routines which will calculate such bivariate normal probabilities.

Gumbel. As in (2.2.5), $p_{rs} = u_s/(u_r + u_s)$, where $u_l = e^{\mu_l}$. Thus

$$P[r \text{ is chosen as the worst from } \mathcal{S}] = Q_{\mathcal{S}}(r) = \frac{u_r}{\sum_{s \in \mathcal{S}} u_s}.$$

Bivariate probabilities can be found by using Lemma 2.3.1, which shows the ranking model derived from the Gumbel model is the same as the backwards Luce model. The probabilities for the table are for $r = t, s \neq u$,

$$\begin{aligned} P[Z_r < Z_s \ \& \ Z_r < Z_t] &= P[\boldsymbol{X} = (r,s,t) \text{ or } (r,t,s)] \\ &= Q_{\{r,s,t\}}(t) Q_{\{r,s\}}(s) + Q_{\{r,s,t\}}(s) Q_{r,t}(t) \\ &= \frac{u_t u_s}{u_r + u_s + u_t} \times \left(\frac{1}{u_r + u_s} + \frac{1}{u_r + u_t} \right); \end{aligned}$$

for $r \neq t, s = u$,

$$P[Z_r < Z_s \ \& \ Z_t < Z_s] = Q_{\{r,s,t\}}(s) = \frac{u_s}{u_r + u_s + u_t};$$

and for $r \neq u, s = t$,

$$P[Z_r < Z_s \ \& \ Z_s < Z_t] = P[\boldsymbol{X} = (r,s,t)] = \frac{u_t}{u_r + u_s + u_t} \times \frac{u_s}{u_r + u_s}.$$

These probabilities are easy to compute. The ones for the forwards Luce model, in which the $-Z_r$'s are Gumbel, are similar.

Other Thurstone models may or may not be easy. For example, if the g's are exponential or uniform, the bivariate probabilities can written in closed form. If g is logistic or Student's t, they cannot be, and in fact need trivariate computations. If we believe that all reasonable g's yield similar models, then the Gumbel model is the easiest to work with. Its advantage may disappear, however, if the Z_i's are not independent or homoscedastic.

To find the Σ from Ω, all that is needed is the $d(-\delta_{rs})$'s. In the normal case, D is $N(0,2)$, so that $d(-\delta_{rs}) = e^{-\delta_{rs}^2/4}/(2\sqrt{\pi})$. For the Gumbel, $d(-\delta_{rs}) = e^{\delta_{rs}}/(1+e^{\delta_{rs}})^2$.

3.7 ϕ models

The ϕ models include Mallows' ϕ model, the ϕ component models, and the ϕ orthogonal contrast models. They are all Babington Smith models and regular exponential family models, so the techniques in Section 3 work well. As mentioned, if m is large, the general Babington Smith model can be challenging to fit computationally since the normalizing constant (1.5.1) in the density (1.5.2) is the product of $m!$ terms, each a sum of $\binom{m}{2}$ terms. The ϕ models greatly simplify the calculations, in part because the natural sufficient statistics are independent.

The most general of the ϕ models is the orthogonal contrast model in Section 1.6.3. Let $(\boldsymbol{C}_1,\ldots,\boldsymbol{C}_q)$ be a set of orthogonal contrasts. For the ϕ model (1.6.9), let $f_0(\boldsymbol{y}) = 1/m!$ in (3.2.1), so that f_0 is the Uniform$(\mathcal{P}_m)$ density. Then the ψ function is

$$\begin{aligned}\psi(\boldsymbol{\theta}) &= \log\left(\frac{1}{m!}\sum_{\boldsymbol{y}\in\mathcal{P}_m} e^{\sum_{i=1}^q \theta_i d(\boldsymbol{C}_i(\boldsymbol{y}))}\right)\\ &= \log(E[e^{\sum \theta_i d(\boldsymbol{C}_i(\boldsymbol{Y}))}])\,;\ \boldsymbol{Y}\sim \text{Uniform}(\mathcal{P}_m).\end{aligned} \tag{3.7.1}$$

By Lemma 1.6.5, the expectation can be written as a product of expectations, one for each $\boldsymbol{C}_i$, since under Uniform$(\mathcal{P}_m)$ the $\boldsymbol{C}_i$'s are independent. Thus

$$\psi(\boldsymbol{\theta}) = \sum_{i=1}^q \log(E[e^{\theta_i d(\boldsymbol{C}_i(\boldsymbol{y}))}])) \equiv \sum_{i=1}^q \psi(\theta_i\,;\,\boldsymbol{C}_i), \tag{3.7.2}$$

that is, the ψ is a sum of terms, one for each contrast $\boldsymbol{C}_i$. Thus we need only the expectation for each individual contrast. Note that Lemma 1.6.5 also says that the $\boldsymbol{C}_i$ are uniform. Letting $\mathcal{C}_i$ be the set of possible values that contrast $\boldsymbol{C}_i$ can have,

$$\begin{aligned}\psi(\theta_i\,;\,\boldsymbol{C}_i) &= \log\left(E[e^{\theta_i d(\boldsymbol{C}_i(\boldsymbol{Y}))}]\right)\\ &= \log\left(\frac{1}{\#\mathcal{C}_i}\sum_{\boldsymbol{C}_i\in\mathcal{C}_i} e^{\theta_i d(\boldsymbol{C}_i)}\right).\end{aligned} \tag{3.7.3}$$

The summations in (3.7.3) are over reasonably small sets. For example, when $m = 10$ the number of elements in $\mathcal{P}_m$ in (3.7.1) is 3,628,800. On the other hand, the largest $\#\mathcal{C}_i$ could possibly be is $C(10, 5) = 252$, and I conjecture that the largest $\sum_{i=1}^{q} \#\mathcal{C}_i$ could be is 292. It turns out that the calculations can be made even simpler.

First consider the ϕ component model (1.6.6). This is an orthogonal contrast model with $\boldsymbol{C}_i = (\{l_i\}, \{l_{i+1}, \ldots, l_m\})$, so that $d(\boldsymbol{C}_i(\boldsymbol{y})) = \overline{y}_i - 1$ by (1.6.8). Under Uniform$(\mathcal{P}_m)$, then, $d(\boldsymbol{C}_i(\boldsymbol{y}))$ is Uniform$(\{0, 1, \ldots, m - i\}$, so that $\#\mathcal{C}_i = m - i + 1$, and (3.7.3) becomes

$$\begin{aligned}\psi(\theta_i\,;\,(\{l_i\}, \{l_{i+1}, \ldots, l_m\}) &= \log\left(\frac{1}{m-i+1}\sum_{j=0}^{m-i} e^{j\theta_i}\right)\\ &= \log\left(\frac{1}{m-i+1}\frac{1-e^{(m-i+1)\theta_i}}{1-e^{\theta_i}}\right)\\ &\equiv \phi_{m-i+1}(\theta_i), \qquad (3.7.4)\end{aligned}$$

where

$$\phi_k(\gamma) = \log(1 - e^{k\gamma}) - \log(1 - e^{\gamma}) - \log(k). \qquad (3.7.5)$$

Thus by (3.7.2), the ψ for the ϕ component model is

$$\psi_{\phi-\text{component}}(\boldsymbol{\theta}) = \sum_{i=1}^{m-1} \phi_{m-i+1}(\theta_i). \qquad (3.7.6)$$

As noted below (1.6.6), Mallows' ϕ model (1.5.5) can be obtained from the ϕ component model by equating the θ_i's. Thus from (3.7.6), the ψ for Mallows' ϕ model is

$$\psi_m(\theta) \equiv \sum_{i=1}^{m-1} \phi_{m-i+1}(\theta). \qquad (3.7.7)$$

Now there is a trick to obtain the (3.7.3) for any contrast $\boldsymbol{C} = (\mathcal{I}, \mathcal{J})$. Let $I = \#\mathcal{I}$ and $J = \#\mathcal{J}$, and relabel the objects so that $\boldsymbol{C} = (\{l_1, \ldots, l_I\}, \{l_{I+1}, \ldots, l_{I+J}\})$. Create the new contrasts

$$\boldsymbol{C}_i = (\{l_i\}, \{l_{i+1}, \ldots, l_I\}) \ \text{ for } \ i = 1, \ldots, I-1,$$

and

$$\boldsymbol{C}_{I+j} = (\{l_{I+j}\}, \{l_{I+j+1}, \ldots, l_{I+J}\}) \ \text{ for } \ j = 1, \ldots, J-1.$$

Then $(\boldsymbol{C}, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_{I-1}, \boldsymbol{C}_{I+1}, \ldots, \boldsymbol{C}_{I+J-1})$ is a set of orthogonal contrasts. Consider the ϕ model with all θ_i's equal to θ. It can be shown to be Mallows' ϕ model (with $I+J$ objects), which has ψ being $\psi_{I+J}(\theta)$ of (3.7.7). Since we know from (3.7.4) the ψ's for contrasts like the $\boldsymbol{C}_i$'s here, we have immediately that

$$\psi_{I+J}(\theta) = \psi(\theta\,;\boldsymbol{C}) + \sum_{i=1}^{I-1} \phi_{I-i+1}(\theta) + \sum_{j=1}^{J-1} \phi_{J-j+1}(\theta). \qquad (3.7.8)$$

Using (3.7.7) twice in (3.7.8), we have that

$$\psi(\theta\,;\boldsymbol{C}) = \psi_{I+J}(\theta) - \psi_I(\theta) - \psi_J(\theta).$$

The above can make finding MLE's quite easy. The loglikelihood function (3.2.3) is now

$$l(\boldsymbol{\theta}\,;\boldsymbol{k}) = \sum_{i=1}^{q} \left(\theta_i \left(\sum_{j=1}^{n} d(\boldsymbol{C}_i(\boldsymbol{y}_j)) \right) - n\psi(\theta_i\,;\boldsymbol{C}_i) \right) + C.$$

Thus the maximum over $\boldsymbol{\theta}$ can be found by maximizing each term over θ_i, which is a univariate maximization. Then Σ, hence $\widehat{\boldsymbol{I}}_n$ and $\boldsymbol{I}_1$, are diagonal. Note that we have changed a p-dimensional maximization of a sum of $m!$ terms into p one-dimensional maximizations of sums of at most m terms.

Chapter 4

Ties, Incomplete Rankings, and Partial Rankings

4.1 Introduction

There are many ways in which a judge could partially rank a set of objects. Some examples:

Example 1: APA election

In the American Psychological Association 1980 presidential election, people were supposed to rank the five candidates from 1 to 5. A total of 5738 people ranked all five, but 5141 only gave their top candidate, 2462 gave their first and second choice, and 2108 gave their top three.

Example 2: Desirable qualities

In the NORC General Social Survey, one question presented 13 desirable qualities for a child to have. The respondent was to give the top three choices, unordered, the bottom three choices, unordered, the top choice of the top three, and the bottom choice of the bottom three. Thus the objects are separated into five groups: #1, #2&3, #5-10, #11&12, and #13. See Alwin and Jackson (1982).

Example 3: Soft drinks

Böckenholt (1992) presents a study in which judges were to rank eight soft drinks. However, instead of ranking all eight together, each judge was presented with two groups of four soft drinks. The judges only ranked the drinks within each group of drinks so that preferences between drinks which were in two different groups were not obtained. The groups of four were chosen according to an incomplete block design. For example, some people were asked to rank Coke, Pepsi, Diet Coke and Diet Pepsi, then to rank 7-up, Sprite, Diet 7-up and Diet Sprite. The purpose of using blocks of four was to make ranking easier, as well as to reduce the number of possible rankings from 8! to 4!.

Example 4: Sports voting

In sports, there are often many judges who must give their top choices out of a large number. For example, in the college football and basketball polls, coaches give their top 20 (?) out of the hundreds of teams. In choosing the Most Valuable Player or Cy Young Award winner in baseball, or the Heisman Trophy winner in college football, the judges give their top k choices out of the hundreds or thousands of players eligible.

Example 5: Board candidates

The Mayor of a large Midwestern city asked a blue-ribbon committee of 13 people to interview and rank 10 candidates for the Board of Commissions. The Mayor would then choose the top three for the Board. Since a chosen candidate might turn down the appointment, it was necessary to do more than just find the top three. After the interviews, the members of the committee presented their assessments of the candidates in the form of rankings, but with ties. The pattern of ties varied with the member. Some examples of the rankings: $(1,1,1,1,2,2,2,2,3)$, $(4,1,3,6,2,7,5,8,9)$, and $(2,1,2,3,2,3,4,5,6)$. (One candidate was so bad that he is being ignored.) The first judge separated the candidates into the top four, next four, and last, while the second gave a complete ranking, and the third gave the top, next three, next two, and then seventh, eighth and ninth.

Example 6: Draft lottery

In the 1970 draft lottery, numbers from 1 to 366 were randomly assigned to the days of the year. Thus there are two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{P}_m$, where $\boldsymbol{x} = (1, 2, \ldots, 366)$ indicates the days of the year, and $\boldsymbol{y} = (305, 159, \ldots, 100)$ were the assigned numbers. Analysis of the data and the randomization procedure suggests that days from earlier months tended to have higher lottery numbers than later months, but that there are no consistent trends within months. Thus a possible model for the procedure could be based on the distance between the $\boldsymbol{y}$ and a vector in which days within months are tied, i.e., the vector $\boldsymbol{z} \equiv (1, 1, \ldots, 1, 2, 2, \ldots, 2, \ldots.., 12, 12, \ldots, 12)$, where there are 31 1's, 29 2's, etc. Note that $m = 366$ and $n = 1$, so that there is little chance of fitting many of the usual models.

Example 7: Educational testing

In educational testing, such as GRE's and ACT's, one finds data consisting of the performance of m test takers on a test with n questions. The objective may be to rank the examinees, or to analyze the properties of the individual questions. Looking at this from a ranking point of view, one has that each question is a "judge" which ranks the examinees (who in this case are, unfortunately, "objects"). Typically, each question can only give a crude ranking, especially if there is only right and wrong and no partial credit, in which case each question can only divide the examinees into a top set and bottom set.

Example 3 is of a different type than the other examples, where there are no ties, but many comparisons are not made. In the others, all objects are compared, but in some cases two or more objects are equally preferred. One can also imagine mixtures of situations, such as if someone preferred A to B and B to D and C to D, but did not compare C to either A or B. Thus C cannot be tied with A nor B.

The next three sections contain general approaches to modeling rank data with incomplete rankings. An incomplete ranking has to be represented somehow. In any specific case, it should be generally easy to invent a reasonable representation. We will represent a generic incomplete ranking by "$\boldsymbol{w}$," which will have to be given explicitly in special cases, but it general will be left vague. Section 4.5 discusses Babington Smith.

4.2 The censored data approach

For any incomplete ranking $\boldsymbol{w}$, there are a number of complete rankings $\boldsymbol{y} \in \mathcal{P}_m$ which are consistent with $\boldsymbol{w}$. For example, the rank vector with ties, $(2, 1, 3, 2)$, is consistent with $(2, 1, 4, 3)$ and $(3, 1, 4, 2)$. The set of partial rankings: "A preferred to B" and "C preferred to D" is consistent with the following rankings of (A, B, C, D):

$$(1, 2, 3, 4),\ (1, 3, 2, 4),\ (1, 4, 2, 3),\ (2, 3, 1, 4),\ (2, 4, 1, 3),\ \text{and}\ (3, 4, 1, 2).$$

The censored data approach assumes that there is a latent complete ranking $\boldsymbol{y}$, but for some reason it is only partially observed. It may be that one has a complete ranking, but is only asked for the top three choices or to make a few paired comparisons; or one does not have the time to perform a detailed ranking; or ranks are based on some real-valued variables, but roundoff error creates some ties. The model assumes that there are two random quantities: $\boldsymbol{Y}$, the complete rank vector, and Δ, which specifies the pattern of ties or incomplete rankings. Then the observed incomplete ranking is a function of $\boldsymbol{Y}$ and Δ,

$$\boldsymbol{W} = \mathrm{g}(\boldsymbol{Y}\,;\,\Delta).$$

The model makes the following assumptions:

1. $\boldsymbol{Y} \sim \mathrm{f}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$, for some density $\mathrm{f}_{\boldsymbol{\theta}}$;

2. Δ is independent of $\boldsymbol{Y}$;

3. The distribution of Δ does not depend on $\boldsymbol{\theta}$.

In Examples 2, 3, 4, and 6, the pattern of ties or incomplete rankings is fixed by the experimenter, hence Δ satisfies Assumptions 2 and 3 automatically. In Examples 1, 5, and 7, each judge decides on what pattern of ties to present, hence Assumption 2 in particular may be suspect. In Example 1, in fact, one can prove statistically that Assumption 2 fails.

One observes $\boldsymbol{W}$, and from that can infer Δ, but is generally only interested in $\boldsymbol{\theta}$. We need the distribution of $\boldsymbol{W}$, but note that by Assumption 3, Δ is an ancillary statistic. (An ancillary statistic is one whose distribution does not depend on $\boldsymbol{\theta}$.) Thus instead of the full likelihood for $\boldsymbol{W}$, we will look at just the partial likelihood, which is the likelihood of $\boldsymbol{W}$ conditional on Δ. Now

$$\begin{aligned}
\mathrm{h}_{\boldsymbol{\theta}}(\boldsymbol{w}\,|\,\delta) &= \mathrm{P}_{\boldsymbol{\theta}}[\boldsymbol{W} = \boldsymbol{w}\,|\,\Delta = \delta] \\
&= \mathrm{P}_{\boldsymbol{\theta}}[\boldsymbol{W} = \boldsymbol{w}\ \&\ \Delta = \delta]/\mathrm{P}_{\boldsymbol{\theta}}[\Delta = \delta] \\
&= \sum_{\{\boldsymbol{y}\,|\,\mathrm{g}(\boldsymbol{y}\,;\,\delta)=\boldsymbol{w}\}} \mathrm{P}_{\boldsymbol{\theta}}[\boldsymbol{Y} = \boldsymbol{y}\ \&\ \Delta = \delta]/\mathrm{P}_{\boldsymbol{\theta}}[\Delta = \delta] \\
&= \sum_{\{\boldsymbol{y}\,|\,\mathrm{g}(\boldsymbol{y}\,;\,\delta)=\boldsymbol{w}\}} \mathrm{P}_{\boldsymbol{\theta}}[\boldsymbol{Y} = \boldsymbol{y}]\mathrm{P}_{\boldsymbol{\theta}}[\Delta = \delta]/\mathrm{P}_{\boldsymbol{\theta}}[\Delta = \delta] \\
&= \sum_{\{\boldsymbol{y}\,|\,\mathrm{g}(\boldsymbol{y}\,;\,\delta)=\boldsymbol{w}\}} \mathrm{P}_{\boldsymbol{\theta}}[\boldsymbol{Y} = \boldsymbol{y}].
\end{aligned}$$

That is, the conditional likelihood for $\boldsymbol{W}$ is just the sum of the likelihoods for the $\boldsymbol{Y}$'s that could have produced $\boldsymbol{W}$. Given a sample of independent $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$, the joint conditional likelihood is

$$\prod_{i=1}^{n} h_{\boldsymbol{\theta}}(\boldsymbol{w}_i \,|\, \delta_i). \tag{4.2.1}$$

In the next three subsections, we will present applications of this approach to specific models and types of incomplete data which yield compact densities $h_{\boldsymbol{\theta}}$. For other models, the calculations may or may not be difficult, but in principle any model can be extended. Often, the EM algorithm can be of use for fitting, where the E-step estimates what the statistics would be if the rankings were complete.

4.2.1 Thurstonian models

In general, the density (4.2.1) is very complicated since each $h_{\boldsymbol{\theta}}$ term is a sum of several multidimensional integrals. However, if the incomplete rankings are constituted as in Example 3, then each $h_{\boldsymbol{\theta}}$ is just the product of the Thurstonian ranking probabilities of the smaller groups of objects. That is, suppose one is to rank {A, B, C, D} and {E, F, G, H}. Then, e.g., as long as the Z_l's corresponding to the two sets of objects are independent,

$$\begin{aligned} &P[(1,3,4,2) \;\&\; (3,2,4,1)] \\ &\qquad\qquad = P[Z_A < Z_D < Z_B < Z_C]P[Z_H < Z_F < Z_E < Z_G]. \end{aligned}$$

The pairwise analysis can also be performed, but one must modify the covariance matrices Ω and Σ appropriately.

4.2.2 Thurstone-Gumbel-Plackett-Luce

Suppose the forward Plackett model holds, and one only ranks the top q choices. Silverberg (1980) calls such partial rankings "q-permutations." Then

$$h_{\boldsymbol{\theta}}(\boldsymbol{w}) = \nu_{l_1} \times \frac{\nu_{l_2}}{\nu_{l_2} + \cdots + \nu_{l_m}} \times \cdots \times \frac{\nu_{l_q}}{\nu_{l_q} + \cdots + \nu_{l_m}},$$

where the observed order of the top q choices is $(l_1, \ldots, l_q)$. A similar formula works for the backwards Plackett model if one only ranks the bottom q objects. Note that q can be different for different judges.

4.2.3 Orthogonal contrast ϕ models

For these models, we assume the vector $\boldsymbol{W}$ consists of possibly tied ranks, but all objects are compared. First, a definition for patterns of ties must be made.

Definition 4.2.1. A **pattern** of ties is a partition $\boldsymbol{s} \equiv (s_1, s_2, \ldots, s_t)$ of the integer m. (That is, the s_i's are positive integers which sum to m.) Such a pattern describes a ranking with ties in which the objects are divided into t groups, with s_1 objects in the top group, s_2 in the second group, $\ldots$, and s_t in the bottom group.

Thus if $m = 5$ and one gives only the top choice, the pattern is (1,4). In Example 2, the pattern is $(1, 2, 7, 2, 1)$. If the ranking is complete, the pattern is $(1, 1, \ldots, 1)$, with m 1's. A ranking with ties, $\boldsymbol{w}$, that has pattern $\boldsymbol{s}$ will be given as an $m \times 1$ vector with s_i values being the integer i, $i = 1, \ldots, t$. Example 5 has such vectors corresponding to patterns $(4, 4, 1)$, $(1, 1, 1, 1, 1, 1, 1, 1, 1)$, and $(1, 3, 3, 1, 1, 1)$, respectively. Note that a complete ranking vector $\boldsymbol{y} \in \mathcal{P}_m$ is consistent with the $\boldsymbol{w}$ if and only if $d_K(\boldsymbol{w}, \boldsymbol{y}) = 0$, where d_K is Kendall's τ distance in (1.8.3).

Now suppose the underlying vector $\boldsymbol{Y}$ is distributed according to an orthogonal contrast ϕ model. The function $h_{\boldsymbol{\theta}}$ has a fairly nice form. First, define the "d" of a contrast $\boldsymbol{C} \equiv (\mathcal{I}_1, \mathcal{I}_2)$ at $\boldsymbol{w}$ by

$$d(\boldsymbol{C}(\boldsymbol{w})) = \sum_{i \in \mathcal{I}_1} \sum_{j \in \mathcal{I}_2} I[w_i > w_j].$$

This definition is actually the same as for a complete ranking $\boldsymbol{y}$, except that now the indicator function will be 0 if either object i is preferred to object j or they are tied. From Marden and Chung (1991), the constant in the exponent of the density for this contrast with parameter θ is

$$\psi(\theta\,; \boldsymbol{C}, \boldsymbol{w}) = \psi_{\#\mathcal{I}_1 + \#\mathcal{I}_2}(\theta_i) - \sum_{j=1}^{2} \psi_{\#\mathcal{I}_j}(\theta) \\ - \sum_{k=1}^{t} \psi_{s_k}(\theta) + \sum_{j=1}^{2} \sum_{k=1}^{t_j} \psi_{u_{jk}}(\theta), \quad (4.2.2)$$

where $\boldsymbol{s} = (s_1, \ldots, s_t)$ is the pattern of ties for the vector $\boldsymbol{w}$ restricted to the objects in $\mathcal{I}_1 \cap \mathcal{I}_2$, and $\boldsymbol{u}_j = (u_{j1}, \ldots, u_{jt_j})$ is that for $\boldsymbol{w}$ restricted to the objects in $\mathcal{I}_j$ $(j = 1, 2)$. Although the ψ looks a bit formidable, it is actually easy to compute since it is just a linear combination of the ϕ's in (3.7.5). The right-hand side of (4.2.2) is analogous to inclusion-exclusion formulas.

For a general orthogonal contrast ϕ model with contrasts $(\boldsymbol{C}_1, \ldots, \boldsymbol{C}_q)$, the density is then

$$h_{\boldsymbol{\theta}}(\boldsymbol{w}) = f_0^*(\boldsymbol{w}) e^{\sum_{i=1}^{q} (\theta_i d(\boldsymbol{C}_i(\boldsymbol{w})) - \psi(\theta_i\,; \boldsymbol{C}_i, \boldsymbol{w}))}, \quad (4.2.3)$$

where f_0^* is the uniform distribution on $\mathcal{W}(\boldsymbol{s})$, the space of tied rankings with pattern $\boldsymbol{s}$. Mallows ϕ model is a special case of the orthogonal contrast ϕ models, so the corresponding model with ties should be straightforward from (4.2.3). It is. The density is

$$h_\theta(\boldsymbol{w}) = f_0^*(\boldsymbol{w}) e^{\theta\, d(\boldsymbol{w},\boldsymbol{\mu}) - \psi(\theta\,;\boldsymbol{s})},$$

where now

$$\psi(\theta\,;\,\boldsymbol{s}) = \psi_m(\theta) - \sum_{j=1}^{t} \psi_{s_j}(\theta).$$

4.3 Distance models

Critchlow (1985) has a general method for extending metrics on $\mathcal{P}_m$ to metrics on spaces of tied rankings. See also Diaconis (1988). He uses the Hausdorff metric. The model assumes all the observations have the same pattern of ties. For any metric d on $\mathcal{P}_m$, the Hausdorff extension to a metric d^* on $\mathcal{W}(\boldsymbol{s})$ is defined by

$$d^*(\boldsymbol{w},\boldsymbol{v}) \equiv \max\{ \max_{\{\boldsymbol{x} \text{ cw } \boldsymbol{w}\}} \min_{\{\boldsymbol{y} \text{ cw } \boldsymbol{v}\}} d(\boldsymbol{x},\boldsymbol{y}),\ \max_{\{\boldsymbol{y} \text{ cw } \boldsymbol{v}\}} \min_{\{\boldsymbol{x} \text{ cw } \boldsymbol{w}\}} d(\boldsymbol{x},\boldsymbol{y})\},$$

where $\boldsymbol{v}$ and $\boldsymbol{w} \in \mathcal{W}(\boldsymbol{s})$, $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathcal{P}_m$, and "cw" means "is consistent with." Now d^* can be used to define a distance model on $\mathcal{W}(\boldsymbol{s})$, where the modal ranking is one with ties in the pattern $\boldsymbol{s}$.

4.4 Exponential family models

The idea here is the same as for complete rankings. First one must decide on what the sample space $\mathcal{W}$ should be, and then must derive the null density $f_0^*(\boldsymbol{w})$. In the complete ranking case, or when the rankings have a fixed pattern of ties, we had the uniform distribution. With an arbitrary set of incomplete rankings, the uniform distribution may not be appropriate. For example, suppose $m = 3$ and $\mathcal{W} = \{(1,2,2),(2,1,2),(3,2,1),(2,3,1)\}$. Here, the first two elements are consistent with two complete rankings, while the last two are consistent only with themselves. Thus rather than giving each element probability $\frac{1}{4}$, the distribution induced by the uniform on $\mathcal{P}_m$ is $(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{6})$.

Once $\mathcal{W}$ and f_0^* are set, all the exponential model needs is a set of sufficient statistics, $\boldsymbol{t} = (t_1,\ldots,t_p)$. The density is then as in (3.2.1), i.e.,

$$h_{\boldsymbol{\theta}}(\boldsymbol{w}) = f_0^*(\boldsymbol{w}) e^{\boldsymbol{\theta}'\boldsymbol{t}(\boldsymbol{w}) - \psi(\boldsymbol{\theta})}.$$

As before, with a sample from $\mathcal{W}$, $\boldsymbol{S}$ is the vector of sums of $\boldsymbol{t}$.

Silverberg (1980) describes his first-order model for q-permutations, which are rankings with pattern of ties $\boldsymbol{s} = (1, 2, \ldots, q, m - q)$, to have $\mathcal{W} = \mathcal{W}(\boldsymbol{s})$, $f_0(\boldsymbol{w})$ be a constant, and sufficient statistics for a sample being

$$S_{ij} \equiv \#\{\text{Object } l_i \text{ is ranked } j\},\ i = 1, \ldots, m,\ j = 1, \ldots, q.$$

Second-order models depend on the S_{ij}'s as well as the numbers of times each pair of objects is ranked j_1 and j_2 for $1 \leqslant j_1, j_2 \leqslant q$, and similarly for higher-(but-no-higher-than-q-)order models.

An analog of the Bradley-Terry model would again restrict to $\mathcal{W}(\boldsymbol{s})$ and use the average ranks as sufficient statistics, but one must decide how to average ranks with ties. The most common way is to use midranks, which for tied objects uses for their ranks the average of the ranks they would have if there were no ties. More specifically,

$$\text{midrank}(\boldsymbol{w}) \equiv \text{Average}\{\boldsymbol{y}\text{'s consistent with } \boldsymbol{w}\}.$$

For example, midrank$((2, 1, 2, 2, 3)) = (3, 1, 3, 3, 5)$.

4.5 Babington Smith

There are several ways one might imagine extending Babington Smith models to incomplete rankings. One way is to use the censored data approach as in Section 2. That may be a bit suspect, however. Suppose the observed $\boldsymbol{w}$ is $(2, 1, 2, 2, 3)$. The censored data model posits that first the judge made all 10 paired comparisons, perhaps obtaining a complete ranking, perhaps not. If not, the judge repeated the paired comparisons. When a complete ranking was finally found, the judge then ignored the A-C-D comparisons. Why go to all the bother of making sure all comparisons are consistent when some are ignored?

Another approach is to assume for each pair three possibilities: prefer the first, prefer the second, it's a tie. This model is Davidson's (1970) for paired comparisons. Thus each pair of objects will have a little triple of probabilities attached. The ranking proceeds by making all comparisons until a consistent ranking, possibly with ties, results.

If the usual Babington Smith experiment is performed but with only a subset of the possible paired comparisons being made, incomplete rankings are likely to result.

Bibliography

Alvo, M. and Yu, P. (2014). *Statistical Methods for Ranking Data*. Springer, New York.

Alwin, D. F. and Jackson, D. (1982). Adult values for children: An application of factor analysis to ranked preference data. In Hauser, R., Mechanic, D., Haller, A., and Hauser, T., editors, *Social Structure and Behavior: Essays in Honor of William Hamilton Sewell*, pages 311–329. Academic Press, New York.

Babington Smith, B. (1950). *Journal of the Royal Statistical Association: Series B*, 12:53–56. Discussion of Ross (1950).

Böckenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, 45:31–49.

Böckenholt, U. (1993). Thurstonian models for ranking data. In Fligner, M. A. and Verducci, J. S., editors, *Probability Models and Statistical Analysis for Ranking Data*, pages 157–172. Springer-Verlag, New York.

Bradley, R. and Terry, M. (1952). Rank analysis of incomplete block designs. I. *Biometrika*, 39:324–345.

Chung, L. and Marden, J. (1991). Use of nonnull models for rank statistics in bivariate, two-sample, and analysis-of-variance problems. *Journal of the American Statistical Association*, 86:188–200.

Coombs, C. (1964). *A Theory of Data*. Wiley, New York.

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

Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology*, 35(3):294–318.

Croon, M. A. (1989). Latent class models for the analysis of rankings. In Geert de Soete, H. F. and Klauer, K. C., editors, *New Developments in Psychological Choice Modeling*, pages 99 – 121. North-Holland.

Daniels, H. (1950). Rank correlation and population models. *Biometrika*, 33:129–135.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65:317–328.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Association: Series B*, 39:1–38.

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

Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *Annals of Statistics*, 17:949–979.

Fligner, M. and Verducci, J. (1993). *Probability Models and Statistical Analyses for Ranking Data*. Springer-Verlag.

Fligner, M. A. and Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Association: Series B*, 48:359–369.

Fligner, M. A. and Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical Association*, 83:892–901.

Henery, R. (1981). Permutation probabilities as models for horse races. *Journal of the Royal Statistical Society: Series B*, 43:86–91.

Henery, R. (1983). Permutation probabilities for gamma random variables. *Applied Probability*, 20:822–834.

Kendall, M. G. and Babington Smith, B. (1940). On the method of paired comparisons. *Biometrika*, 31(3-4):324–345.

Lamport, L. (1994). *LaTeX: A Document Preparation System*. Addison-Wesley, second edition.

Luce, R. (1959). *Individual Choice Behavior*. Wiley, New York.

Madsen, L. and Wilson, P. R. (2015). *memoir — Typeset Fiction, Nonfiction and Mathematical Books*. `https://www.ctan.org/pkg/memoir`.

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

Marden, J. (1992). Use of nested orthogonal contrasts in analyzing rank data. *Journal of the American Statistical Association*, 87:307–318.

Marden, J. (1996). *Analyzing and Modeling Rank Data*. Taylor & Francis.

McCullagh, P. (1993). Permutations and regression models. In Fligner, M. A. and Verducci, J. S., editors, *Probability Models and Statistical Analysis for Ranking Data*, pages 196–215. Springer-Verlag, New York.

Plackett, R. (1975). The analysis of permutations. *Applied Statistics*, 24:193–202.

Ross, A. S. C. (1950). Philological probability problems. *Journal of the Royal Statistical Association: Series B*, 12:19–59. With discussion.

Silverberg, A. (1980). *Statistical Models for* q*-permutations*. PhD thesis, Department of Statistics, Princeton University.

Silverberg, A. (1984). Statistical models for q-permutations. In *Proceedings of the Biopharmaceutical Section*, pages 107–112. American Statistical Association.

Stern, H. (1987). Gamma processes, paired comparisons and ranking. Technical Report #64, Department of Statistics, Stanford University.

Thurstone, L. (1927). A law of comparative judgement. *Psychological Reviews*, 34:273–286.

van Blokland-Vogelesang, R. A. (1989). Unfolding and consensus ranking: A prestige ladder for technical occupations. In Geert de Soete, H. F. and Klauer, K. C., editors, *New Developments in Psychological Choice Modeling*, pages 237–258. North-Holland.

Verducci, J. (1982). *Discriminating between two Probabilities on the Basis of Ranked Preferences*. PhD thesis, Department of Statistics, Stanford University.

Wikipedia (2017). Condorcet criterion — Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/w/index.php?title=Condorcet_criterion&oldid=780916852`.

Yellott, J. (1977). 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.