# Chapter 8
# Probability Models for Ranking Data

Probability modeling for ranking data is an efficient way to understand people's perception and preference on different objects. Various probability models for ranking data have been developed, particularly in the last decade where many new problems involving a large number of objects emerged. In their review paper on probability models for ranking data, Critchlow et al. (1991) broadly categorized these models into four classes: (1) order statistics models, (2) paired comparison models, (3) distance-based models, and (4) multistage models. Since their publication in 1991, variants of these models and new models have been developed. In this chapter, we will introduce these four classes of models and describe their properties.

Before introducing these models, we would like to describe several distinctive features of these models, which may affect the choice of models to be considered in our study:

(a) *Some models allow for the presence of covariates*
In collecting data on rankings of a set of objects from a sample of judges, we may also obtain information on some covariates from the judges (judge-specific covariates) and covariates of the objects (object-specific covariates). Some covariates may even be judge-object-specific. For example, in collecting customers' preferences on a list of mobile phones, the judge-specific covariates could be age, gender, and income, and the object-specific covariates could be prices, weights, and brands, and the judge-object-specific covariates could be some personal experience on using each phone or brand. Most models except for the distance-based models and multistage models can allow for the presence of covariates.

(b) *Some models are predictive*
If we want to build a model to predict a ranking assigned by an individual, we need to have a predictive model for ranking data. In this case, the presence of covariates is a must and it is expected that the population is heterogeneous and different covariates may lead to different ranking of objects predicted from the



M. Alvo, P.L.H. Yu, *Statistical Methods for Ranking Data*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-1-4939-1471-5_8

fitted model. However when the population is homogeneous, the rankings given by judges can be assumed to be generated from a probability model on rankings. A distance-based model is a typical example.

(c) *Some models can handle big ranking data with a large number of objects*
Most ranking models should work well for a small number of objects, say less than 10 or 15. Some may become computationally demanding or even infeasible to use for a large number of objects, examples of which will be the Thurstone order statistics models as its likelihood requires the computation of a high-dimensional integration. Recently, the development of social networks and the competitive pressure to provide customized services motivated many new ranking problems on hundreds or thousands of objects. Recommendations on products such as movies, books, and songs are typical examples in which the number of objects is extraordinarily large. In recent years, many researchers in statistics and computer science have developed models to handle such big data.

## 8.1 Order Statistics Models

Among the above four classes of probability models for ranking data, the class of order statistics models has the longest history in the statistical and psychological literature. Dating back to 1927, Thurstone published his/her famous paper *A law of comparative judgment* in *Psychological Review* in which the ranking of two objects was considered. The basic idea behind this approach is that a judge may have tastes that fluctuate from one instant to another according to the perception of each object which is not perfectly predictable and hence is a random variable. The ordering of these random variables then determines the judge's ranking of the objects. Thurstone (1927) proposed a ranking process where the ranking $\boldsymbol{\pi}_j$ of $t$ objects given by a random sample of judge $j$ $(j = 1, 2, \ldots, n)$ is determined by the relative ordering of $t$ random utilities $y_{1j}, y_{2j}, \cdots, y_{tj}$, where $\boldsymbol{y}_j = (y_{1j}, \cdots, y_{tj})', j = 1, \cdots, n$ are independent.

The probability of observing a ranking $\boldsymbol{\pi}_j$ under the class of order statistics models is

$$P(\boldsymbol{\pi}_j) = P(y_{[1]_j j} > y_{[2]_j j} > \ldots > y_{[t]_j j}), \qquad \boldsymbol{\pi}_n \in \mathcal{P} \tag{8.1}$$

where $< [1]_j, [2]_j, \cdots, [t]_j >$ is the ordering of objects corresponding to ranking $\boldsymbol{\pi}_j$ such that judge $j$ assigns rank $i$ to object $[i]_j$ (i.e., $\pi_j([i]_j) = i$ or $\pi_j^{-1}(i) = [i]_j$) and $\mathcal{P}$ is the set of all $t!$ possible rankings. It should be noted that the order statistics model (8.1) is invariant under any strictly increasing transformation of the $y_{ij}$'s for which the ordering of the $y_{ij}$'s is preserved.

Critchlow et al. (1991) observed that if the utilities $y_{1j}, \cdots, y_{tj}$ are allowed to have arbitrary dependencies, any probability distribution on rankings can be expressed as in (8.1). To simplify the model, some probabilistic structures on $y$'s

are assumed. The most common one is to assume that the $y_{ij}$'s are independent with cumulative distribution function. $F_i(y) = F(y - \alpha_{ij})$ or equivalently

$$y_{ij} = \alpha_{ij} + \varepsilon_{ij}, \tag{8.2}$$

where $\alpha_{ij}$ is the expected utility determined by judge $j$ to object $i$ and $\boldsymbol{\varepsilon}_j = (\varepsilon_{1j}, \cdots, \varepsilon_{tj})'$, $j = 1, \cdots, n$ are i.i.d. random vectors with cumulative distribution function $F$. Such models are referred to as *Thurstone order statistics models* (see Yellot 1977; Critchlow et al. 1991). Two famous Thurstone models studied extensively in the literature are

- Thurstone model (Thurstone 1927; Daniels 1950; Mosteller 1951):

$$F \text{ is the standard normal.}$$

- Luce model (Bradley and Terry 1952; Luce 1959):

$$F \text{ is Gumbel (type I extreme value)}^{1}, \text{ i.e., } F(\varepsilon) = \exp(-\exp(-\varepsilon)).$$

Since the Luce model leads to a closed form,

$$P(\boldsymbol{\pi}_j) = \prod_{i=1}^{t-1} \frac{\exp(\alpha_{[i]_j j})}{\sum_{l=i}^{t} \exp(\alpha_{[l]_j j})}, \tag{8.3}$$

most applications and extensions are based on the Luce model. As the exponential distribution satisfies the memoryless property, it may not be appropriate in modeling the running times in many track competitions. Henery (1983) and Stern (1990a,b) thus extended the Luce model to the Thurstone model with error $\varepsilon_{ij} = \ln(u_{ij})$, where $u_{ij}$ follows a Gamma distribution with shape $r$ and scale 1. Properties of the Thurstone order statistics model can be found in Henery (1981) and Critchlow et al. (1991).

### *8.1.1 Luce Model*

The Luce model can be viewed as an extension of the *multinomial (conditional) logit model* for top choice (McFadden 1974). For example, in examining 3 objects by judge $j$, object 2 is selected as the top-choice, i.e., the ordering is $< 2, \_, \_ >$, with the following top choice probability:

$$P(< 2, \_, \_ >) = P(y_{2j} > y_{1j}, y_{3j}) = \frac{\exp(\alpha_{2j})}{\exp(\alpha_{1j}) + \exp(\alpha_{2j}) + \exp(\alpha_{3j})}.$$

[1] Note $e^{-\varepsilon}$ follows an exponential distribution with mean 1.

Also, the probability of observing the ranking of 3 objects (3, 1, 2) (i.e., ordering: $< 2, 3, 1 >$) under the Luce model is

$$P(y_{2j} > y_{3j} > y_{1j}) = \frac{e^{\alpha_{2j}}}{e^{\alpha_{1j}} + e^{\alpha_{2j}} + e^{\alpha_{3j}}} \cdot \frac{e^{\alpha_{3j}}}{e^{\alpha_{1j}} + e^{\alpha_{3j}}}.$$

It is not difficult to see that the ranking probability under the Luce model can be expressed as a function of top-choice probabilities only.

**Theorem 8.1.** *Let $p_{ij}$ be the probability that object $i$ is ranked first by judge $j$ among the full list of $t$ objects. That is, $p_{ij} = P(y_{ij} > y_{kj}\ \forall k \neq i)$. Then the probability of ranking $\boldsymbol{\pi}_j$ with ordering $< [1]_j, [2]_j, \cdots, [t]_j >$ under the Luce model is given by*

$$P(\boldsymbol{\pi}_j) = p_{[1]_j j} \frac{p_{[2]_j j}}{1 - p_{[1]_j j}} \frac{p_{[3]_j j}}{1 - p_{[1]_j j} - p_{[2]_j j}} \cdots \frac{p_{[t-1]_j j}}{1 - p_{[1]_j j} - p_{[2]_j j} - \cdots - p_{[t-2]_j j}}.$$

*Proof.* The proof follows by observing that under the Luce model,

$$p_{ij} = \frac{\exp(\alpha_{ij})}{\exp(\alpha_{1j}) + \cdots + \exp(\alpha_{tj})}.$$ □

**Definition 8.1 (Independence of Irrelevant Alternatives (IIA) Tversky 1972).** Let $P(a|S)$ be the probability of choosing an object $a$ from a choice set $S \subseteq \{1, 2, \cdots, t\}$. The independence of irrelevant alternatives asserts that object $a$ is preferred to object $b$, by the (top) choice probability, is independent of the choice set $S$.

From Definition 8.1, we have,

$$P(a|S) > P(b|S) \Longleftrightarrow P(a|\{a, b\}) > P(b|\{a, b\}) \Longleftrightarrow P(a|\{a, b\}) > \frac{1}{2}.$$

If object $a$ is preferred to object $b$ out of the choice set $\{a, b\}$, then introducing a third alternative object $c$, thus expanding the choice set to $\{a, b, c\}$, must not make object $b$ preferable to object $a$. In other words, the choices between $a$ and $b$ depend on the preferences between $a$ and $b$ only, i.e., it is irrelevant to $c$.

**Theorem 8.2 (Luce 1959).** *The Luce model satisfies the IIA.*

*Proof.* Under the Luce model, it is easy to show that

$$P(a|S) = \frac{\exp(\alpha_a)}{\sum_{i \in S} \exp(\alpha_i)}$$

and thus $\frac{P(a|S)}{P(b|S)} > 1 \Longleftrightarrow \frac{\exp(\alpha_a)}{\exp(\alpha_b)} > 1 \Longleftrightarrow \frac{\exp(\alpha_a)/(\exp(\alpha_a)+\exp(\alpha_b))}{\exp(\alpha_b)/(\exp(\alpha_a)+\exp(\alpha_b))} > 1 \Longleftrightarrow \frac{P(a|\{a,b\})}{P(b|\{a,b\})} > 1.$ □

*Example 8.1.* Is IIA a good property? No. Let us consider the problem of selecting a travel mode to work among a car (C), a blue bus (B), or a red bus (R). Initially a traveler has a choice of going to work by car or taking a blue bus with $P(C) = P(B) = \frac{1}{2}$ . Now a red bus is introduced and the traveler considers the red bus to be exactly like the blue bus (i.e., $P(R) = P(B)$). However, in the Luce model, the odds $P(C)/P(B)$ is the same whether or not another alternative exists. The only probabilities for which $P(C)/P(B) = 1$ and $P(R)/P(B) = 1$ are $P(C) = P(B) = P(R) = \frac{1}{3}$. In real life, we would expect $P(C) = \frac{1}{2}$ and $P(B) = P(R) = \frac{1}{4}$.

It is natural that our choice on an object (such as the blue bus) will depend on our preference on similar objects or even its substitutes (like the red bus). By ignoring such dependency, the estimation of choice/ranking probabilities of course will be biased. In other words, if the list of all travel modes contains many irrelevant objects such as walking, bicycling, and skateboarding, it might be acceptable to estimate the probability for choosing car/bus based on the subset {car, bus} instead of the full list {car, bus, walking, bicycling, skateboarding}. However the estimation in this case will be relatively less efficient.

### *8.1.2 Rank-Ordered Logit Models*

The Luce model can be extended to incorporate covariates as well. For example, we may include $M$ covariates of judge $j$, $x_{mj}$, $m = 1, 2, \ldots, M$, into the mean utility, i.e.,

$$\alpha_{ij} = \beta_{i0} + \sum_{m=1}^{M} \beta_{im} x_{mj}, \tag{8.4}$$

where $\beta_{im}$, $m = 0, 1, \ldots, M$ are the parameters specific to object $i$, and $P$ covariates of object $i$, $z_{pi}$, $p = 1, 2, \cdots, P$, into the mean utility, i.e.,

$$\alpha_{ij} = \beta_{i0} + \sum_{p=1}^{P} \gamma_p z_{pi}, \tag{8.5}$$

where $\gamma_p$, $p = 1, 2, \cdots, P$ are the parameters specific to all judges.

A further extension of the Luce model (specified in Allison and Christakis (1994)) includes judge-specific covariates, object-specific covariates, and their interactions or judge-object-specific covariates ($w_{qij}, q = 1, 2, \cdots, Q$) into the mean utility:

$$\alpha_{ij} = \beta_{i0} + \sum_{p=1}^{P} \gamma_p z_{pi} + \sum_{m=1}^{M} \beta_{im} x_{mj} + \sum_{q=1}^{Q} \theta_q w_{qij}, \tag{8.6}$$

where $\theta_q, q = 1, 2, \cdots, Q$ are the parameters specific to all judges and objects. These extensions of the Luce models are known as rank-ordered logit (ROL) model in the field of econometrics (see for example Chapman and Staelin 1982; Beggs et al. 1981; Hausman and Ruud 1987).

In the Luce and ROL models described above, the log-likelihood function is globally concave, and hence a global maximum exists (Beggs et al. 1981). The maximum likelihood estimates (MLE) of the model parameters can thus be obtained using standard methods, e.g., Newton-Raphson algorithm. Besides MLE, Koop and Poirier (1994) used a Bayesian method to estimate the parameters.

Both the Luce and ROL models can be built using the R package `mlogit`. Here, we use an example to demonstrate these two models.

*Example 8.2.* Consider a ranking data set for gaming platforms in which 91 Dutch students were asked to rank 6 gaming platforms: Xbox, PlayStation, GameCube, PlayStation Portable, Gameboy, or a personal computer (PC). The data set also contains information on whether the student currently owns each platform (own), the age of the student (age), and the number of hours spent on gaming per week (time). This data set was first studied in Fok et al. (2012) and can be accessed in the R package, `mlogit`.

First, we fit a Luce model (9.3) with PC as the reference level and the parameter estimates are shown in Table 8.1. It is noticed that students prefer Xbox for playing games the most and then PC, PlayStation, PlayStation Portable, GameCube, and Gameboy. However, playing games on Xbox and PlayStation are not significantly different from that on PC.

Now, we extend the Luce model by including object-specific covariate (own) and judge-specific covariate (time) into the model. This leads to the rank-ordered logit (ROL) model and the parameter estimates are shown in Table 8.1. It can be observed that owning a platform has a positive effect on the preference for the same platform and that students who spend more time playing games prefer a PC more than other gaming platforms. Applying the likelihood ratio test to compare the two models, it is clearly that the ROL model is substantially better. Notice that including age as another judge-specific covariate does not significantly improve the likelihood of the ROL (from $-517.37$ to $-516.55$), and hence the results are omitted here.

ROL models are popular for ranking data, and many extensions have been developed by different scholars. Koop and Poirier (1994) extended the use of ROL models to more general cases of ranking data. The number of objects ranked by $n$ judges can be different. The rank given by each judge is not necessarily complete. The objects that each judge is assigned to rank can be different as well. Fok et al. (2012) studied the mixtures of ROL models and found them to be useful in analyzing ranking capabilities.

**Table 8.1** Parameter estimates of the fitted Luce and ROL models for the gaming platform data

| Variable | Luce | ROL |
|---|---|---|
| *Intercept* | | |
| Xbox | 0.13 (0.18) | 1.40 (0.29) |
| PlayStation | −0.00 (0.18) | 0.94 (0.27) |
| PlayStation Portable | −0.65 (0.18) | 0.80 (0.28) |
| GameCube | −1.22 (0.20) | 0.05 (0.30) |
| Gameboy | −1.28 (0.19) | 0.09 (0.28) |
| *Platform ownership* | | 0.96 (0.19) |
| *Hours spent on gaming* | | |
| Xbox | | −0.17 (0.05) |
| PlayStation | | −0.13 (0.04) |
| PlayStation Portable | | −0.23 (0.05) |
| GameCube | | −0.19 (0.05) |
| Gameboy | | −0.24 (0.05) |
| Log-likelihood | −547.00 | −517.37 |

### *8.1.3 Some Non-IIA Order Statistics Models*

In spite of the fact that the ranking probability (8.1) under both the Luce and ROL models has a closed form, the unrealistic IIA property makes them fit some data not so well (see for example Brook and Upton 1974; Tallis and Dansie 1983; Bockenholt 1993). The main reason is that no correlation is assumed among the errors over the objects. This lack of correlation translates into an unrealistic substitution pattern among objects in some situations (see Example 8.1). Therefore, to overcome these problems, dependency structures other than those in the Thurstone order statistics model are required.

#### 8.1.3.1 Multivariate (Generalized) Extreme Value (GEV) Models

McFadden (1978) introduced the multivariate (or generalized) extreme value model which provides closed-form top-choice probabilities without the IIA restriction. The GEV assumes that the error terms in (8.2) follows a generalized extreme value distribution with cumulative distribution function

$$F(\varepsilon_1, \cdots, \varepsilon_t) = \exp[-H(e^{-\varepsilon_1}, \cdots, e^{-\varepsilon_t})],$$

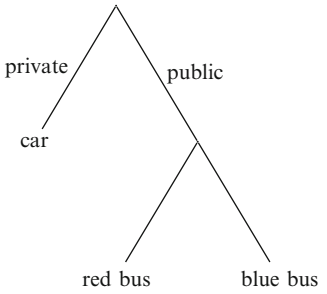
where $G = \exp(-H)$ is a $t$-dimensional copula and all the univariate marginals are Gumbel distributed. Of course, when $H(x_1, \cdots, x_I) = \sum_{i=1}^{I} x_i$, the model degenerates to the Luce model. The GEV model is very flexible and Joe (2001) showed that the GEV model can fit various types of ranking data. Note that this result does not provide a way to construct the function $H$. In fact, the popular GEV

model used in the literature is the nested logit model in which the function $H$ is expressed in a hierarchical form:

$$H(x_1, \cdots, x_I) = \sum_{k=1}^{K} (\sum_{j \in B_k} x_j^{1/\lambda_k})^{\lambda_k},$$

where $B_1, \cdots, B_K$ are $K$ nonoverlapping subsets (called nests) formed from a partition of all objects.

Under the nested logit model, the $\varepsilon_i$'s are correlated within nests but uncorrelated between nests. For example, suppose $K = 2$, $B_1 = \{\text{car}\}$, and $B_2 = \{\text{red bus, blue bus}\}$, it is reasonable that one who prefers traveling with the red bus may also prefer traveling with the blue bus and vice versa, but one's preference on car may not depend on his/her preference on the two buses. Such dependency structure can be represented by the following hierarchical form:




#### 8.1.3.2 Mixed Logit Models

Note that the rank-ordered logit model assumes that the utility for each object follows the linear model:

$$y_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{\beta} + \varepsilon_{ij}$$

where the error terms $\varepsilon_{ij}$'s are independent and identically (type I) extreme value distributed. To allow dependency among the utilities, mixed logit models assume that the beta coefficients are judge-specific:

$$y_{ij} = \boldsymbol{x}_{ij}' \boldsymbol{\beta}_j + \varepsilon_{ij}$$

and further assume that $\boldsymbol{\beta}_j$'s are random and independent identically distributed with density $f(\boldsymbol{\beta}|\theta)$, where $\theta$ are some unknown parameters. A typical choice of $f$ is the normal density with mean $\boldsymbol{\beta}_0$ and covariance matrix $\boldsymbol{\Omega}$. Such randomness in

$\boldsymbol{\beta}_j$ allows unexplainable variation of covariates' impacts over judges and correlation of utilities across objects. McFadden and Train (2000) showed that any discrete choice model can be well approximated by a mixed logit model with appropriate specification of the distribution of $\boldsymbol{\beta}_j$ and the covariates $\boldsymbol{x}$.

Conditional on $\boldsymbol{\beta}_j$, the probability of observing $\boldsymbol{\pi}_j$ by judge $j$ is given in (8.3) with $\alpha_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j$. Integrating it over the density of $\boldsymbol{\beta}_j$ then gives the unconditional probability under the mixed logit model:

$$P(\boldsymbol{\pi}_j) = \int \prod_{i=1}^{t-1} \left( \frac{e^{\boldsymbol{x}'_{[i]j}\boldsymbol{\beta}}}{\sum_{l=i}^{t} e^{\boldsymbol{x}'_{[l]j}\boldsymbol{\beta}}} \right) f(\boldsymbol{\beta})\, d\boldsymbol{\beta}.$$

If the mixing distribution $f(\boldsymbol{\beta})$ is discrete, with $\boldsymbol{\beta}$ taking a finite set of distinct values, the mixed logit model becomes the **latent class model** and sometimes called the **finite mixture model**.

Both nested logit and mixed logit models can be built using the R package `mlogit` which provides maximum likelihood estimation and the numerical integration (if any) in the likelihood is estimated using simulation techniques such as quasi-Monte Carlo method.

*Example 8.3.* The R package `mlogit` contains a top-choice data set named `Electricity` in which 361 individuals were asked in a series of at most 12 choice experiments. In each experiment, each individual was asked to choose the best out of four hypothetical electricity suppliers with different combination of characteristics including electricity price ($pf$) (in cents per kWh) and length of contract ($cl$, in years) offered, whether a time-of-day rate ($tod$) is included, whether a seasonal rate ($seas$) is included, and whether the supplier is local ($loc$) or is well known ($wk$).

We first fit a multinomial logit (MNL) model (i.e., the Luce model with the top choice only) using `mlogit` and its parameter estimates are shown in Table 8.2. The significant negative coefficients for $pf$, $tod$, $seas$, and $cl$ and the significant positive coefficients for $loc$ and $wk$ indicate that individuals tend to prefer a local and well-known supplier which offers a shorter length of contract with a lower fee.

Note that `Electricity` is a clustered data set as each individual was involved in a number of choice experiments. The independence assumption of the choice responses used in the multinomial logit model is therefore invalid. To incorporate such clustered effect, we use the mixed logit model with utility $y_{ijk}$ for supplier $i$ given by individual $j$ in the $k$th experiment as follows:

$$y_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta}_j + \varepsilon_{ijk},$$

where $\boldsymbol{\beta}_j$'s are independent identically distributed. As the utilities made by individual $j$ share the same random $\boldsymbol{\beta}_j$, the utilities given by the same individual are correlated whereas the utilities given by different individuals are uncorrelated. This helps describe the clustered effect.

**Table 8.2** Parameter estimates of the fitted MNL and mixed logit models for the electricity supplier data

| Variable | MNL | Mixed logit |
|---|---|---|
| *Fixed effect* | | |
| Electricity price ($pf$) | −0.625 (0.023) | −0.933 (0.034) |
| Length of contract ($cl$) | −0.108 (0.008) | −0.196 (0.013) |
| Time-of-day rate ($tod$)? | −5.463 (0.184) | −8.838 (0.286) |
| Seasonal rate ($seas$)? | −5.840 (0.187) | −8.860 (0.287) |
| Local ($loc$)? | 1.442 (0.051) | 2.105 (0.080) |
| Well known ($wk$)? | 0.996 (0.045) | 1.493 (0.065) |
| *Random effect (standard deviation)* | | |
| Electricity price ($pf$) | | 0.200 (0.011) |
| Length of contract ($cl$) | | 0.357 (0.018) |
| Time-of-day rate ($tod$)? | | 2.489 (0.120) |
| Seasonal rate ($seas$)? | | 1.274 (0.107) |
| Local ($loc$)? | | 1.503 (0.089) |
| Well known ($wk$)? | | 0.885 (0.075) |
| Log-likelihood | −4958.6 | −3970.3 |

Using the independent normal assumption for the $\boldsymbol{\beta}_j$'s, the mixed logit model is fitted using `mlogit` and the parameter estimates are shown in Table 8.2. It can be seen from the log-likelihood that the mixed logit model significantly performs better than the multinomial logit model and, in fact, all random effects in the mixed logit model are highly significant. Based on the fitted model, it is easy to see that an individual with mean coefficients for $pf$ and $cl$ is willing to pay $0.196/0.933 = 0.21$ cent per kWh extra in order to shorten the contract length by one year.

#### 8.1.3.3 Multilevel Logit Models

Notice that the above mixed logit model is basically a mixed model with both fixed and random effects. If more sampling information and dependency structures are available, more structured mixed models can be considered. For instance, Skrondal and Rabe-Hesketh (2003) applied a three-level logit model to analyze ranking data collected from the 1987–1992 panel of the British Election Study for rankings on three political parties: Conservative, Labour, and Liberal (Alliance) (indexed by $a$), given by a sample of voters (indexed by $j$) casting votes at different elections (indexed by $i$) over different constituencies (indexed by $k$). Note that in this three-level model, elections are nested within voters and voters nested within constituencies. One model considered is the random intercepts model at voter and constituency levels:

$$y_{aijk} = \alpha_{aijk} + \gamma_{ajk} + \gamma_{ak} + \varepsilon_{aijk}$$

where $\alpha_{aijk} = z_{aijk}b + \boldsymbol{x}'_{aijk}\boldsymbol{\beta}_a$ represents the fixed effects while $\gamma_{ajk}$ and $\gamma_{ak}$ represent, respectively, the random intercepts at both voter and constituency levels, and constituency level only. This special kind of multilevel logit models for ranking data can be built using the Stata program `gllamm` (http://gllamm.org/examples.html) which provides maximum likelihood estimation with integration approximated by quadrature methods.

## 8.2 Paired Comparison Models

Motivated by the connection between a ranking of objects and all pairwise comparisons of objects, paired comparison models aim at combining models for paired comparisons to generate a probabilistic model for ranking data. Note that a ranking of $t$ objects can be indexed by $t(t-1)/2$ pairwise preferences $I_{ab}$, $a < b$, where $I_{ab} = 1$ means object $a$ is preferred to object $b$. Smith (1950) assumed that the ranking is deduced from a set of $t(t-1)/2$ arbitrary paired comparison probabilities $p_{ab}$, $a < b$, where $p_{ab}$ is the probability of object $a$ being preferred to object $b$. The model does not allow ties, so that $p_{ab} = 1 - p_{ba}$. Assuming mutual independence of these $t(t-1)/2$ paired comparisons under the Smith model, the probability of observing a ranking $\boldsymbol{\pi}_j$ is thus given by

$$P(\boldsymbol{\pi}_j) = C \prod_{\{(a,b):\pi_j(a)<\pi_j(b)\}} p_{ab}, \tag{8.7}$$

where the constant $C$ is chosen to make the probabilities sum to 1. Note that the Smith model is indexed by $t(t-1)/2$ parameters $\{p_{ab}\}$. Imposing additional constraints on the $\{p_{ab}\}$ proposed by Mallows (1957) leads to two important subclasses of the Smith model: the Mallows-Bradley-Terry model and the Mallows model.

*The Class of Mallows-Bradley-Terry (MBT) Models*. To reduce the number of parameters in (8.7), Bradley and Terry (1952) proposed to re-parametrize $p_{ab}$ as

$$p_{ab} = \frac{v_a}{v_a + v_b}$$

where $v_i$ is a positive value associated with object $i$ and the sum of all $v_{i's}$ is equal to 1. Mallows (1957) substituted this form into the Smith model, which leads to the following ranking model. For any ranking $\boldsymbol{\pi}_j$ with associated ordering $< [1]_j, [2]_j, \cdots, [t]_j >$,

$$P(\boldsymbol{\pi}_j) = C(\boldsymbol{v}) \prod_{s=1}^{t-1} (v_{[s]_j})^{t-s}$$

where $C(\boldsymbol{\nu})$ is the proportionality constant. Since the Bradley-Terry paired comparison probabilities are invariant when multiplying the $v_i$'s by a positive constant, the number of free parameters is reduced to $t - 1$. Larger values of $v_i$ correspond to more preferred objects, just as the Thurstone order statistics model.

*The Class of Mallows Models*. Before discussing details of the model, we first give the definition of *modal ranking*.

**Definition 8.2.** A probability model is said to be strongly unimodal with modal ranking $\boldsymbol{\pi}_0$, if its ranking probability has the unique maximum at $\boldsymbol{\pi} = \boldsymbol{\pi}_0$.

Mallows (1957) further simplified the MBT model by re-expressing $p_{ab}$ as

$$p_{ab} = \frac{1}{2} + \frac{1}{2}\tanh[(\pi(a) - \pi(b))\ln(\theta) + \ln(\phi)],$$

where $\theta, \phi \in (0, 1)$. Thus, the Mallows model is given by

$$P(\boldsymbol{\pi}_j) = c(\theta, \phi)\theta^{d_S(\pi, \pi_0)}\phi^{d_K(\pi, \pi_0)},$$

where $c(\theta, \phi)$ is chosen to make the probabilities sum to 1 and $d_S(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ and $d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ are the Spearman and Kendall distances between $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$ (c.f. Sect. 3.1). The Mallows model has the interpretation that the ranking probability decreases geometrically according to increasing distance from $\boldsymbol{\pi}$ to the modal ranking $\boldsymbol{\pi}_0$.

For a detailed review on paired comparison models, readers can refer to David (1988). Pendergrass and Bradley (1960) further extended the paired comparison models to triple comparison models.

## 8.3 Distance-Based Models

A distance function is useful in measuring the discrepancy between two rankings. The usual properties of a distance function between two rankings $\mu$ and $\boldsymbol{\nu}$ are: (1) reflexivity, $d(\mu, \mu) = 0$; (2) positivity, $d(\mu, \boldsymbol{\nu}) > 0$ if $\mu \neq \boldsymbol{\nu}$; and (3) symmetry, $d(\mu, \boldsymbol{\nu}) = d(\boldsymbol{\nu}, \mu)$. For ranking data, we require that the distance, apart from having these usual properties, must be right invariant,

$$d(\mu, \boldsymbol{\nu}) = d(\mu \circ \boldsymbol{\tau}, \boldsymbol{\nu} \circ \boldsymbol{\tau}), \text{ where } \mu \circ \boldsymbol{\tau}(i) = \mu(\boldsymbol{\tau}(i)).$$

This requirement ensures that a relabeling of the objects has no effect on the distance. If a distance function satisfies the triangle inequality $d(\mu, \boldsymbol{\nu}) \leq d(\mu, \boldsymbol{\sigma}) + d(\boldsymbol{\sigma}, \boldsymbol{\nu})$, the distance is said to be a *metric*.

Some popular right-invariant distances have been given in Chap. 3. Note that the Spearman Footrule and Kendall distance are metrics, but the Spearman distance

is not, just as the squared Euclidean distance is not. To produce a metric version of the Spearman distance, we may take the square root of the Spearman distance, as given by

$$\left(\sum_{i=1}^{t}[\mu(i)-\nu(i)]^2\right)^{0.5}. \tag{8.8}$$

Readers can refer to Critchlow et al. (1991) for further examples of distance functions.

It is reasonable to assume that there is a modal ranking $\boldsymbol{\pi}_0$, and we expect most of the judges to have rankings close to $\boldsymbol{\pi}_0$. According to this framework, Diaconis (1988) developed a class of distance-based models,

$$P(\boldsymbol{\pi}|\lambda,\boldsymbol{\pi}_0)=\frac{e^{-\lambda d(\boldsymbol{\pi},\boldsymbol{\pi}_0)}}{C(\lambda)}, \tag{8.9}$$

where $\lambda \geq 0$ is the dispersion parameter and $d(\boldsymbol{\pi},\boldsymbol{\sigma})$ is an arbitrary right-invariant distance. In the particular case where we use Kendall as the distance function, the model is called the Mallows' $\phi$-model (Mallows 1957). Note that Mallows' $\phi$-models also belong to the class of paired comparison models (Critchlow et al. 1991). Critchlow and Verducci (1992) and Feigin (1993) provided more details about the relationship between distance-based models and paired comparison models.

In distance-based models, the ranking probability is the greatest at the modal ranking $\boldsymbol{\pi}_0$ and the probability of a ranking will decay the further it is away from the modal ranking $\boldsymbol{\pi}_0$. The rate of the decay is governed by the parameter $\lambda$. For a small value of $\lambda$, the distribution of rankings will be more concentrated around $\boldsymbol{\pi}_0$. When $\lambda$ becomes very large, the distribution of rankings will look more uniform. The closed form for the proportionality constant $C(\lambda)$ only exists for some distances. In principle, it can be solved numerically by summing the value $e^{-\lambda d(\boldsymbol{\pi},\boldsymbol{\pi}_0)}$ over all possible $\boldsymbol{\pi}$ in $\mathcal{P}$. This numerical calculation could be time-consuming, as the computational time increases exponentially with the number of objects.

Given a ranking data set $\{\boldsymbol{\pi}_k, k=1,\ldots,n\}$ and a known modal ranking $\boldsymbol{\pi}_0$, the maximum likelihood estimator (MLE) $\hat{\lambda}$ of the distance-based model can be found by solving the following equation:

$$\frac{1}{n}\sum_{k=1}^{n} d(\boldsymbol{\pi}_k,\boldsymbol{\pi}_0)=E_{\hat{\lambda},\boldsymbol{\sigma}}[d(\boldsymbol{\pi},\boldsymbol{\pi}_0)], \tag{8.10}$$

which equates the observed mean distance with the expected distance under the distance-based model.

The MLE can be found numerically because the observed mean distance is a constant and the expected distance is a strictly decreasing function of $\hat{\lambda}$. For the ease of solving, we re-parametrize $\lambda$ with $\phi$ where $\phi = e^{-\lambda}$. The range of $\phi$ lies in

$(0, 1]$ and the value of $\hat{\phi}$ can be obtained using the method of bisection. Critchlow (1985) suggested applying the method with 15 iterations, which yields an error of less than $2^{-15}$. Also, the central limit theorem holds for the MLE $\hat{\lambda}$, which is shown in Marden (1995).

If the modal ranking $\boldsymbol{\pi}_0$ is unknown, it can be estimated by the MLE $\hat{\boldsymbol{\pi}}_0$ which minimizes the sum of distance over $\mathcal{P}$, that is,

$$\hat{\boldsymbol{\pi}}_0 = \underset{\boldsymbol{\pi}_0 \epsilon \mathcal{P}}{\operatorname{argmin}} \sum_{k=1}^{n} d(\boldsymbol{\pi}_k, \boldsymbol{\pi}_0). \tag{8.11}$$

For a large $t$, a global search algorithm for MLE $\hat{\boldsymbol{\pi}}_0$ is not practical because the number of possible rankings is too large. Instead, as suggested in Busse et al. (2007), a local search algorithm should be used. They suggested iteratively searching for the optimal model ranking with the smallest sum of distances $\sum_{k=1}^{n} d(\boldsymbol{\pi}_k, \boldsymbol{\pi}_0)$ over $\boldsymbol{\pi}_0 \in \Pi^{(m)}$, where $\Pi^{(m)}$ is the set of all rankings having a Cayley distance (Sect. 8.3.2) of 0 or 1 to the optimal modal ranking found in the $m$th iteration:

$$\hat{\boldsymbol{\pi}}_0^{(m+1)} = \underset{\boldsymbol{\pi}_0 \epsilon \Pi^{(m)}}{\operatorname{argmin}} \sum_{k=1}^{n} d(\boldsymbol{\pi}_k, \boldsymbol{\pi}_0).$$

Cayley's distance $d_C(\boldsymbol{\pi}, \boldsymbol{\sigma})$ is defined to be the minimal number of transpositions needed to transform $\boldsymbol{\pi}$ to $\boldsymbol{\sigma}$. A reasonable choice of the initial ranking $\hat{\boldsymbol{\pi}}_0^{(0)}$ can be formed by ordering the mean ranks.

Distance-based models can handle partial ranking, with some modifications in the distance measures. There are several ways to handle partially ranked data in distance-based models. Beckett (1993) estimated the model parameters using the EM algorithm. On the other hand, Adkins and Fligner (1998) offered a non-iterative maximum likelihood estimation procedure for Mallows' $\phi$-model without using the EM algorithm. Critchlow (1985) suggested replacing the distance metric $d$ by the Hausdorff metric $d^*$. The Hausdorff metric between two partial rankings $\boldsymbol{\pi}^*$ and $\boldsymbol{\sigma}^*$ equals

$$d^*(\boldsymbol{\pi}^*, \boldsymbol{\sigma}^*) = \max[\max_{\boldsymbol{\pi} \in C(\boldsymbol{\pi}^*)} \min_{\boldsymbol{\sigma} \in C(\boldsymbol{\sigma}^*)} d(\boldsymbol{\pi}, \boldsymbol{\sigma}), \max_{\boldsymbol{\sigma} \in C(\boldsymbol{\sigma}^*)} \min_{\boldsymbol{\pi} \in C(\boldsymbol{\pi}^*)} d(\boldsymbol{\pi}, \boldsymbol{\sigma})], \tag{8.12}$$

where $C(\mu^*)$ is the set of complete rankings compatible with $\mu^*$ (see Definition 3.1).

### 8.3.1 $\phi$-Component Models

Fligner and Verducci (1986) extended the distance-based models by decomposing the distance metric $d(\boldsymbol{\pi}, \boldsymbol{\sigma})$ into $t - 1$ distance metrics,

$$d(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{t-1} d_i(\boldsymbol{\pi}, \boldsymbol{\sigma}), \tag{8.13}$$

where $d_i(\boldsymbol{\pi}, \boldsymbol{\sigma})$'s are statistically independent. Kendall's distance can be decomposed in this form. Fligner and Verducci (1986) developed two new classes of ranking models, called $\phi$-component models and cyclic structure models, for the decomposition.

Fligner and Verducci (1986) showed that Kendall distance satisfies (8.13):

$$d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{i=1}^{t-1} V_i, \tag{8.14}$$

where

$$V_i = \sum_{j=i+1}^{t} I\{[\pi(\pi_0^{-1}(i)) - \pi(\pi_0^{-1}(j))] > 0\}. \tag{8.15}$$

Here, $V_1$ represents the number of adjacent transpositions required to place the best object in $\boldsymbol{\pi}_0$ in the first position and then remove this item in both $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$, and $V_2$ is the number of adjacent transpositions required to place the best remaining object in $\boldsymbol{\pi}_0$ in the first position of the remaining items, and so on. Therefore, the ranking can be described as $t-1$ stages, $V_1$ to $V_{t-1}$, where $V_i = m$ can be interpreted as $m$ mistakes made in stage $i$.

By applying dispersion parameter $\lambda_i$ at stage $V_i$, the Mallow's $\phi$-model is extended to

$$P(\boldsymbol{\pi}|\boldsymbol{\lambda}, \boldsymbol{\pi}_0) = \frac{e^{-\sum_{i=1}^{t-1} \lambda_i V_i}}{C(\boldsymbol{\lambda})}, \tag{8.16}$$

where $\boldsymbol{\lambda} = \{\lambda_i, i = 1, \ldots, t-1\}$ and $C(\boldsymbol{\lambda})$ is the proportionality constant, which equals

$$\prod_{i=1}^{t-1} \frac{1 - e^{-(t-i+1)\lambda_i}}{1 - e^{-\lambda_i}}. \tag{8.17}$$

These models were named $t-1$ parameter models in Fligner and Verducci (1986), but were also named $\phi$-component models in other papers (e.g., Critchlow et al. 1991). Mallow's $\phi$-models are special cases of $\phi$-component models when $\lambda_1 = \ldots = \lambda_{t-1}$.

Based on a ranking data set $\{\boldsymbol{\pi}_k, k = 1, \ldots, n\}$ and a given modal ranking $\boldsymbol{\pi}_0$, the maximum likelihood estimates $\hat{\lambda}_i$, $i = 1, 2, \ldots, t-1$ can be found by solving the equation

$$\frac{1}{n} \sum_{k=1}^{n} V_{k,i} = \frac{e^{-\hat{\lambda}_i}}{1 - e^{-\hat{\lambda}_i}} - \frac{(t-i+1)e^{-(t-i+1)\hat{\lambda}_i}}{1 - e^{-(t-i+1)\hat{\lambda}_i}}, \tag{8.18}$$

where

$$V_{k,i} = \sum_{j=i+1}^{t} I\{[\pi_k(\pi_0^{-1}(i)) - \pi_k(\pi_0^{-1}(j))] > 0\}. \tag{8.19}$$

The left- and right hand sides of (8.18) can be interpreted as the observed mean and theoretical mean of $V_i$, respectively.

The extension of distance-based models to $t-1$ parameters allows more flexibility in the model, but unfortunately, the symmetric property of distance is lost. Notice here that the so-called "distance" in $\phi$-component models can be expressed as

$$\sum_{i<j} \lambda_i I\{[\pi(\pi_0^{-1}(i)) - \pi(\pi_0^{-1}(j))] > 0\}, \tag{8.20}$$

which is obviously not symmetric, and hence it is not a proper distance measure. For example, in $\phi$-component model, let $\boldsymbol{\pi} = (2, 3, 4, 1), \boldsymbol{\pi}_0 = (4, 3, 1, 2)$:

$$\begin{aligned} d(\boldsymbol{\pi}, \boldsymbol{\pi}_0) &= \lambda_1 V_1 + \lambda_2 V_2 + \lambda_3 V_3 = 3\lambda_1 + 0\lambda_2 + 1\lambda_3 \\ &\neq 1\lambda_1 + 2\lambda_2 + 1\lambda_3 = d(\boldsymbol{\pi}_0, \boldsymbol{\pi}). \end{aligned}$$

The symmetric property of distance is thus not satisfied. Lee and Yu (2012) introduced new weighted distance measures which can retain the properties of a distance and also allow different weights for different ranks. For the details, read Chap. 11.

### 8.3.2 Cyclic Structure Models

Cayley's distance can also be decomposed into $t-1$ statistical independent metrics. Fligner and Verducci (1986) showed that $d_C(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ can be decomposed as

$$d_C(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{i=1}^{t-1} X_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0), \tag{8.21}$$

where $X_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = I(i \neq \max\{\sigma(i), \sigma(\sigma(i)), \ldots\})$ and $\sigma(i) = \pi(\pi_0^{-1}(i))$.

This generalization can be illustrated by an example found in Fligner and Verducci (1986). Suppose there are $t$ lockers and each locker has one key that can open it. The key for locker $j$ is placed in locker $\sigma(j)$. Without loss of generality, let the cost of breaking a locker be one. The minimum cost of opening all lockers will then be $C(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$, and it can be decomposed as the sum of costs of opening locker $\pi^{-1}(i), i = 1, 2, \ldots t-1$, which equals $X_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$.

If we relax the assumption that the costs of breaking every locker are equal, the total cost will become

$$\sum_{i=1}^{t-1} \theta_i X_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0), \tag{8.22}$$

where $\theta_i$ is the cost of opening locker $i$. This "total cost" can be interpreted as a weighted version of Cayley's distance. Similar to the extension of Mallow's $\phi$-models to $\phi$-component models, Fligner and Verducci (1986) developed the cyclic structure models using the weighted Cayley distance. Under this model assumption, the probability of observing a ranking $\boldsymbol{\pi}$ is

$$P(\boldsymbol{\pi}|\boldsymbol{\theta}, \boldsymbol{\pi}_0) = \frac{e^{-\sum_{i=1}^{t-1} \theta_i X_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}}{C(\boldsymbol{\theta})}, \tag{8.23}$$

where $\boldsymbol{\theta} = \{\theta_i, i = 1, \ldots, t-1\}$ and $C(\boldsymbol{\theta})$ is the proportionality constant, which equals

$$\prod_{i=1}^{t-1} \{1 + (t-i)e^{-\theta_i}\}. \tag{8.24}$$

For a ranking data set $\{\boldsymbol{\pi}_k, k = 1, \ldots, n\}$ with a given modal ranking $\boldsymbol{\pi}_0$, the MLEs $\hat{\theta}_i$, $i = 1, 2, \ldots, t-1$ can be found from the equation

$$\hat{\theta}_i = \log(t-i) - \log \frac{\bar{X}_i}{1 - \bar{X}_i}, \tag{8.25}$$

where

$$\bar{X}_i = \frac{\sum_{k=1}^{n} X_i(\boldsymbol{\pi}_k, \boldsymbol{\pi}_0)}{n}. \tag{8.26}$$

## 8.4 Multistage Models

The class of multistage models includes ranking data models that postulate the ranking process can be decomposed into a sequence of independent stages. For a ranking of $t$ objects, the ranking process can be decomposed into $t-1$ stages, where at stage $i$, the $i$th object is selected. In this respect, the Luce models and $\phi$-component models described above clearly belong to the class of multistage models.

Fligner and Verducci (1988) proposed the general multistage models with $\frac{t(t-1)}{2}$ parameters. They are

$$p(m, r) = \text{Prob}(V_r = m), \tag{8.27}$$

where

$$\sum_{m=0}^{t-r} p(m, r) = 1 \tag{8.28}$$

and $V$'s are defined as in the previous section.

A total of three multistage models are proposed in Fligner and Verducci (1988), namely the free model, the strongly unimodal model, and the exponential factor model. Under the free model, which is the most general (least constraints) multistage models, the probability of observing a ranking $\boldsymbol{\pi}$ is

$$\prod_{r=1}^{t-1} p(m, r). \tag{8.29}$$

Under the strongly unimodal model, the parameters will have additional constraints, which are

$$p(0, r) > p(1, r) \tag{8.30}$$

and

$$p(m, r) \text{ is a nonincreasing function of } m, \tag{8.31}$$

for both $m$ and $r = 1, 2, \ldots t$.

Under the exponential factor model, the parameters will be in the form of

$$p(m, r) = C(r)e^{-\lambda_r f(m)}, \tag{8.32}$$

where $f(\cdot)$ is a nonnegative and strictly increasing arbitrary function, and $C(r)$ is the proportionality constant. To avoid the identification problem, the convention that $f(0) = 0$ and $f(1) = 1$ is suggested. Note that if $f(x) = x$, the model will become the $\phi$-component model.

Besides the multistage model proposed by Fligner and Verducci (1988), Xu (2000) also proposed a multistage model with $(t-1)^2$ parameters $c_{ij}$, both $i$ and $j = 1, 2, \ldots t-1$. The parameters $c_{rj}$, $j = 1, 2, \ldots, t-1$ determine which object will be selected in stage $r$.

## 8.5 Properties of Ranking Models

As defined in Critchlow et al. (1991), some properties for ranking models are as follows:

(1) Label invariance
The relabeling of objects has no effect on the probability models.
(2) Reversibility
A reverse function $\gamma(\boldsymbol{\pi})$ for a ranking of $t$ objects is defined as

$$\gamma(i) = t + 1 - i. \tag{8.33}$$

Reversing the ranking $\boldsymbol{\pi}$ has no effect on the probability models.
(3) $L$-decomposability
The ranking of $t$ objects can be decomposed into $t-1$ stages. At stage $i$, where $i = 1, 2, \ldots, t-1$, the best among the objects remaining at that stage is selected, and then this object will be removed in the following stages.
(4) Strong unimodality (weak transposition property)
A transposition function $\tau_{ij}$ is defined to mean that $i$ and $j$ are interchanged as

$$\tau(i) = j, \tau(j) = i, \tau(m) = m \text{ for all } m \neq i, j. \tag{8.34}$$

With modal ranking $\boldsymbol{\pi}_0$, for every pair of objects $i$ and $j$ such that $\pi_0(i) < \pi_0(j)$ and every $\boldsymbol{\pi}$ such that $\pi(i) = \pi(j) - 1$,

$$P(\boldsymbol{\pi}) \geq P(\boldsymbol{\pi} \circ \tau_{ij}), \tag{8.35}$$

with equality attained at $\boldsymbol{\pi} = \boldsymbol{\pi}_0$. It guarantees the probability is nonincreasing as $\boldsymbol{\pi}$ moves one step away from $\boldsymbol{\pi}_0$, for objects having adjacent ranks.
(5) Complete consensus (transposition property)
As compared with the strong unimodality, complete consensus is an even stronger property which guarantees for every pair of objects $(i, j)$ such that $\pi_0(i) < \pi_0(j)$ and every $\boldsymbol{\pi}$ such that $\pi(i) < \pi(j)$, $P(\boldsymbol{\pi}) \geq P(\boldsymbol{\pi} \circ \tau_{ij})$. From this definition, we can see that complete consensus implies strong unimodality.

All four classes of models satisfy property (1). However, not all of them satisfy properties (2) to (5). We will discuss them in the following.

### *8.5.1 Properties of Order Statistics Models*

Critchlow et al. (1991) showed that, for order statistics models, if the random error distribution is symmetric, then the models will satisfy property (2).

Property (3) is difficult to verify for the order statistics model, because it involves a multiple integral which may not have a closed form, except for the special case of the Luce (1959) model, which can satisfy property (3).

Savage (1956, 1957), and Henery (1981) showed that, if for all $i$, $\mu_{ij}$ is distinct for $j = 1, 2, \ldots, t$ and

$$\frac{F'(y - \mu_{iu})}{F'(y - \mu_{iv})} \tag{8.36}$$

is a nonincreasing function of $x$ for $\mu_{iu} < \mu_{iv}$, where $F(\cdot)$ is the cumulative distribution function of the random error, the order statistics models will satisfy properties (4) and (5).

### 8.5.2 Properties of Paired Comparison Models

Marley (1968) showed that the class of paired comparison models satisfy properties (2) and (3), which can be easily verified from the definition of paired comparison models.

Critchlow et al. (1991) showed that paired comparison models will satisfy property (4) under the following conditions:

- $p_{ij} > 0.5$ and $p_{jm} > 0.5$ imply $p_{im} > 0.5$,
- $p_{ij} \neq 0.5$,

for all $i, j, m = 1, 2, \ldots, t$.

Property (5) will be satisfied under the following conditions:

- $p_{ij} > 0.5$ and $p_{jm} > 0.5$ imply $p_{im} > \max(p_{ij}, p_{jm})$,
- $p_{ij} \neq 0.5$,

for all $i, j, m = 1, 2, \ldots, t$.

### 8.5.3 Properties of Distance-Based and Multistage Models

Critchlow et al. (1991) showed that all distance-based models satisfy properties (1) and (2) and models with the four distances in Sect. 8.3 satisfy properties (3) to (5). The Hausdorff metric extension of Critchlow (1985) with the four distances in Sect. 8.3 also satisfies properties (1) to (5).

It is obvious that multistage models satisfy property (3) but not (2). Fligner and Verducci (1988) showed that the strongly unimodal model, but not the free model, satisfies property (3). Furthermore, the exponential factor model satisfies property (4), and hence the $\phi$-component model also satisfies property (4) as it is a special case of the exponential factor model.

## Chapter Notes

In this chapter, we have introduced several important probability models for ranking data. Extension of order statistics models and distance-based models will be discussed in Chaps. 9 and 11, respectively. Other models not considered here are a variety of exponential family models based on marginals (spectral decomposition of Diaconis (1988, 1989)) or pairwise and higher-way comparisons (inversion models of McCullagh (1993b)), nested orthogonal contrast models (Marden 1992), and models based on insertion sorting (Doignon et al. 2004; Biernacki and Jacques 2013).