Quaderni di Statistica
Vol. 12, 2010

Some methodological issues on multivariate modelling of rank data

Angela D'Elia

Dipartimento di Scienze Statistiche, Università di Napoli Federico II

Summary: In this paper, we develop a parametric model for multivariate rank data, expressing individual preferences. Our aim is to propose a procedure for the ranking process of m items, based on a criterion of sequential comparisons among the objects. We discuss, then, some issues related to the probability distribution implied by the ordered choices process, and a numerical example is shown in order to confirm the consistency of the approach. Finally, some inferential aspects are highlighted, especially with regard to the relationship between the univariate and the multivariate analysis of ranks.

Keywords: Choice Criteria, Order Vector, Permutations, Preferences, Rank Vector.

1. Introduction

There are many situations in which a set of judges rates a set of items, as happens, for instance, in athletic competitions or in product evaluations. In general, in several contexts it is worth to study the issue of the choice and the preference between two or more items: the consumers tastes in Marketing, the voters preferences in Politics, the agreement toward different ideas or actions in Psychology, the customers satisfaction in Quality Management, the graduation among different diagnoses in Medicine, etc.

Quite obviously, the variety of fields of interest and of possible applications has yielded heterogeneous statistical tools aimed at analyzing the data (often, rank data) expressing the choice made by the raters, as consumers, voters, etc. (for an extensive review, see: Fligner and Verducci, 1993; Marden, 1995; Taplin, 1997).

Generally speaking, we can assume that the choice between two or more items is, mainly, based on two alternative procedures:

- *paired comparisons criterion*, when the items are few, so that each of them can be compared with each other;

- *sequential comparisons criterion*, when the rater elicits his/her preferred object, then the second best among the remaining ones, and so on, up to the least preferred one. This model seems more suitable than the former for situations in which there are several items to be ranked (more than 5/6 items, say).

These two approaches, introduced by Kendall and Babington-Smith (1940) and Plackett (1975), respectively, have been extended in several ways and from many points of view, focussing on the measure of agreement among the raters (Daniels, 1950; Agresti, 1984; Tanner and Young, 1985; Sadooghi-Alvandi, 1992), or, more specifically, on the preferences modelling (Bradley and Terry, 1952; Davidson and Bradley, 1969; McCullagh, 1980; Fligner and Verducci, 1988; Henery, 1991; Agresti, 1992). However, in spite of the wideness of developments and applications, there is a lack of proposals concerning specific distribution models for the rankings.

Recently, both the choice criteria have been considered in order to develop statistical models for rankings, based on the Shifted Binomial and on the Inverse Hypergeometric random variables, respectively (D'Elia, 1999; 2000a). Both the approaches performed well in terms of explanatory capabilities, and resulted to be also consistent tools for the univariate analysis of the preferences in a generalized linear models framework. In fact, the *sequential choices* model turned out to have a better goodness of fit to several data sets and, of course, it seems to be conceptually more convincing (D'Elia, 2000b; 2000c).

More specifically, these approaches were devoted to a univariate analysis of the ranks, separately for each item. In this work, instead, we want to consider the complete ranking of all the items, taking into account that the allotment of ranks to m items is a choice process intrinsically multivariate. Thus, our aim is to highlight a procedure on which the ranking process of m different items could be based, and to propose a new statistical model for studying multivariate rank data.

We will assume that the choice among more items happens on the basis of a *sequential comparisons* criterion. With respect to this point, it is important to notice that the ranks are just the result of an ordered choices process: thus, in order to properly analyze rankings, the order data, expressing the choices sequence, will be considered too.

Finally, we will assume that ties are not permitted inside the ranking; indeed, it has been noticed that the possibility of tied ranks encourages lazy behaviours of the judges in rating the items: for this reason, most of ranking/classifying experiments does not allow ties (Quandt, 1998).

The structure of the paper is the following. In section 2, the notation and some basic ideas are introduced with regard to the distinction between rank data and order data. Then, our proposal of a multivariate model is shown in section 3, and subsequently developed in detail for the case of three items, in section 4. In section 5 some estimation and computational issues are addressed, and the relation with the univariate approach is also discussed. Section 6 is devoted to further developments concerning the proposed model and to final remarks.

2. Rank and order vectors

Let us consider the problem of ranking m items. For a statistical analysis, it is irrelevant to state the nature of the items, because the problem of the choice can arise in several situations and, then, the items could be almost everything (food or car brands, political parties, football teams, singers, actors, places where to live, colours, actions, ideas, jobs, etc.).

Let $(R_1, R_2, ..., R_m)$ be the rank random variables assigned by the choice process to the ordered m items $(\mathcal{O}_1, \mathcal{O}_2, ..., \mathcal{O}_m)$ of a fixed list. Let us assume that, $R_j = 1$ means best and $R_j = m$ means worst, $\forall j = 1, 2, ..., m$.

Of course, the observed ranks $(r_1, r_2, ..., r_m)$ are just permutations of the first m integers: $(1, 2, ..., m)$. Each permutation belongs, thus, to the class of permutations $\mathcal{P}_m$, whose cardinality is: $card(\mathcal{P}_m) = m!$. Usually (e.g. Marden, 1995), the probability distribution functions used to describe rankings are defined with respect to $\mathcal{P}_m$, so that, for instance, in the case of indifference, each element of the class has a constant probability mass equal to $1/m!$.

This approach has three main drawbacks:

- there is no criterion for ordering the elements of $\mathcal{P}_m$;
- the enumeration of all the $m!$ permutations becomes, rapidly, cumbersome as m increases;
- the observed ranks can give some insight about the probability distribution function only when then number of raters n is extremely large with respect to the number of items m (as noticed, for example, by Sadooghi-Alvandi, 1992).

For these reasons, we think that a more useful approach should be based on the definition of a parametric model for the ranks.

With regard to the distribution of the ranks of a single and fixed item $\mathcal{O}_j$ $(j = 1, 2, ..., m)$, we proposed (D'Elia, 1999) that the observed rank r_j can be thought of as a realization of an Inverse Hypergeometric random variable, $R_j \sim IHG(m, B_j)$, whose probability mass function (pmf), for $r = 1, 2, ..., m$, is:

$$Pr(R_j = r) = \frac{\binom{B_j+m-1-r}{m-r}}{\binom{B_j+m-1}{m-1}} = \frac{(\frac{B_j-1}{m-r}+1)\,(\frac{B_j-1}{m-r-1}+1)\ldots(\frac{B_j-1}{1}+1)}{(\frac{B_j}{m-1}+1)\,(\frac{B_j}{m-2}+1)\ldots(\frac{B_j}{1}+1)}\,.$$

The IHG random variable used in our proposal describes a drawing without replacement from an urn, with B_j white balls and $m-1$ not white balls, until the first white ball is drawn (Guenther, 1975); thus, this scheme is consistent with the values assumed by the ranks assigned to a given object $\mathcal{O}_j$.

A different, and often more useful, parametrization can be obtained by letting:

$$\theta_j = \frac{B_j}{B_j + m - 1}\,;$$

thus, after some algebra, the $R_j \sim IHG(m, \theta_j)$ random variable has the following pmf:

$$Pr(R_j\!=\!r)\!=\!\begin{cases} \theta_j, & r = 1, \\ c_r\, \theta_j(1-\theta_j)^{r-1} \prod_{s=1}^{r-1}(m-s-1+ s\theta_j)^{-1}, & r = 2,\dots,m, \end{cases}$$

where $c_r = \prod_{s=1}^{r-1}(m-s) = (m-1)!/(m-r)!,\ r = 2,\dots,m$, and the parameter θ_j represents a liking/agreement measure for the j-th item.

More specifically, we have

$$\begin{aligned} Pr(R_j = 1) &= \theta_j, \\ Pr(R_j = 2) &= \theta_j(1-\theta_j)\frac{m-1}{m-2+\theta_j}, \\ Pr(R_j = 3) &= \theta_j(1-\theta_j)^2\frac{(m-1)(m-2)}{(m-2+\theta_j)(m-3+2\theta_j)}, \\ \dots\dots\dots\dots\dots &\dots\dots \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \end{aligned}$$

These probabilities can be computed in an effective manner by noting that $Pr(R_j = 1) = \theta_j$ and the recursive formula, for $r = 1, 2, ..., m-1$:

$$Pr(R_j = r+1) = Pr(R_j = r)(1-\theta_j)\frac{m-r}{m-1-r+r\theta_j}.$$

Let us consider, now, a ranking as a multivariate random variable. In previous works, we have analyzed its components, separately, especially focussing on the relation between the raters' features and ranks assigned to a prespecified item j (D'Elia, 2000b). A problem with this kind of univariate analysis is that each single component of a ranking is considered independently from the others, while the multivariate random variable $(R_1, R_2, ..., R_m)$ lies in a $(m-1)$-dimensional space, with:

$$R_1 + R_2 + ... + R_m = \frac{m(m+1)}{2}\,.$$

Thus, we need to develop a multivariate structure that takes into account both the relation between the ranks of the items and the nature of each single ranking element itself.

More difficulties come out when we consider the whole ranking by a multivariate point of view.

The main problem is that we cannot state a unique order among the components of a ranking, since the natural order $(R_1, R_2, ..., R_m)$ not necessarily corresponds to the order of choice performed by each rater.

In particular, two issues must be considered.

- If no ties are allowed in the ranking of m items, the value observed for R_j cannot be observed for R_h, $\forall h \neq j$: this means, for example, that if $R_1 = 3$, no other rank random variable R_h $(h \neq 1)$ can assume the value 3. This leads to conditional distributions that are not always consistent with the temporal sequence of choices. Indeed, if we use a multivariate distribution where R_2 cannot take the value observed for R_1, R_3 cannot take the values assumed by R_1 and R_2, and so on. Then, the admissible values domain for each element would be restricted following only the natural order $(1, 2, \dots, m)$, while it should logically follow the sequential order of choice.

  Moreover, for example, if we have $(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4)$ and we observe the ranking (3,1,4,2), it is not temporal consistent to ask for the probability of ($R_2 = 1$) given that ($R_1 = 3$), since, on the basis of the *sequential comparison criterion*, this question would imply a conditioning of a past event (the choice of the most preferred item) on future events (the choice of the third preferred item).

- On the other hand, if we use a multivariate distribution which follows the order of choice, the sequence of its components will vary with the raters: for instance, if in correspondence of $(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4)$ we observe the rankings (3,2,1,4) and (2,4,3,1), the ordered sequences of the components for the two raters will be (R_3, R_2, R_1, R_4) and (R_4, R_1, R_3, R_2), respectively. Thus, it turns out that each rater will have its own pmf, making impossible any likelihood based inference.

For these reasons, we are proposing to transform the multivariate rank random variable $(R_1, R_2, ..., R_m)$ in a new set of *index* random variable

$(\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_m)$, by means of the following relation:

$$(R_j = k) \Leftrightarrow (\mathcal{W}_k = j),\ \forall (k, j) \in \{1, 2, ..., m\} \times \{1, 2, ..., m\}.$$

In this way, the index[1] random variable ($\mathcal{W}_1$, $\mathcal{W}_2, ..., \mathcal{W}_m$) represent, in their natural order, *where* is located the item most preferred, *where* is the second best, and so on, with respect to the fixed and prespecified order of the submitted list $(\mathcal{O}_1, \mathcal{O}_2, ..., \mathcal{O}_m)$.

For example, if there are $m = 4$ items $(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4)$, and the preference order of a rater is $(\mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_1, \mathcal{O}_4)$ – that is, $\mathcal{O}_2$ is the most preferred item, and $\mathcal{O}_4$ is the least preferred item – the observed *rank vector* is (3, 1, 2, 4)′, while the observed *order vector* (or *index vector*) is $(w_1 = 2, w_2 = 3, w_3 = 1, w_4 = 4)'$. It represents the places in the list where are located the most preferred item, the second best, and so on.

Thus, the value assumed by the random variable $\mathcal{W}_1$, that is $w_1 = 2$, can be interpreted as the answer to the question: “*Where* is, in the ordered list, the most preferred item ? It is immediate, in this example, to notice the correspondence between the rank random variable and the index random variable:

$$(R_1 = 3) \Leftrightarrow (\mathcal{W}_3 = 1); \quad (R_2 = 1) \Leftrightarrow (\mathcal{W}_1 = 2);$$

$$(R_3 = 2) \Leftrightarrow (\mathcal{W}_2 = 3); \quad (R_4 = 4) \Leftrightarrow (\mathcal{W}_4 = 4).$$

These relations show, for example, that since the first item of the list ($\mathcal{O}_1$) has rank R_1=3, the random variable $\mathcal{W}_3$ must take value 1, and so on.

The previous example emphasizes that, even if the rank and the order vectors are both permutations of the same set $\mathbf{e} = (1, 2, \ldots, m)'$, they are conceptually quite different. Indeed,

- the *rank vector* represents the position in a preference list of the items listed in a given order;

[1] Sometimes different notations and motivations are introduced in the literature for this transformation: for instance, Marden (1995) speaks of *order vector*, while Sadooghi-Alvandi (1992) calls this a *preference vector*. In the following, we call ($\mathcal{W}_1$, $\mathcal{W}_2, ..., \mathcal{W}_m$), indifferently, *order vector* or *index vector*.

- the *order vector* represents the position in the fixed list of the items in a preference order.

Thus, R_j is the preference position of the j-th item in the list, while $\mathcal{W}_k$ is the list position of the k-th preferred item.

Let $\boldsymbol{r}_i$=$(r_{i1}, r_{i2}, ..., r_{im})'$ be the observed rank vector for a given i-th rater, and $\mathbf{w}_i$=$(w_{i1}, w_{i2}, ..., w_{im})'$ be the corresponding order vector. Then, the following relations result:

$$\boldsymbol{r}_i = \mathbf{Pe}; \qquad \mathbf{w}_i = \mathbf{Pr}_i = \mathbf{P^2 e} = \mathbf{Qe};$$

where $\boldsymbol{P}$ is an orthogonal *permutation matrix* $(m \times m)$, whose elements are defined as:

$$p_{j,k} = \begin{cases} 1, \text{ if } r_{ij} = k; \\ 0, \text{ otherwise;} \end{cases}$$

and $\boldsymbol{Q}$ is an orthogonal *permutation matrix* $(m \times m)$ too, whose elements are:

$$q_{k,j} = \begin{cases} 1, \text{ if} w_{ik} = j; \\ 0, \text{ otherwise.} \end{cases}$$

A typical **P** (or **Q**) is a null matrix, except for single units values in any row and in any column, as the following one:

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 & ... & 0 \\ 1 & 0 & ... & 0 \\ 0 & 0 & ... & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Usually, the empirical analysis of rankings is focused on the matrix of the observed ranks **R**:

$$\boldsymbol{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,m} \end{bmatrix}$$

where:

$$r_{i,j} = \{rank\ assigned\ by\ rater\ i\ to\ item\,\mathcal{O}_j\,\},$$

for $i = 1, 2, ...n; j = 1, 2, ..., m.$

Using the previously introduced relation, we can transform each row of $\mathbf{R}$ in a row of index random variable, obtaining the matrix of observed indexes:

$$\mathcal{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,m} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n,1} & w_{n,2} & \cdots & w_{n,m} \end{bmatrix}$$

where:

$$w_{i,k} = \{location\ in\ the\ list\ of\ the\ k{-}th\ preferred\ item\,by\,rater\,i\,\},$$

for $i = 1, 2, ...n; k = 1, 2, ..., m.$

For example, let us consider 3 items $(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3)$ and the ranking $(2, 3, 1)'$ expressed by a rater. Then, using the previous definitions, we have:

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix};$$

$$\boldsymbol{Q} = \boldsymbol{P}^2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Thus, $\mathrm{w}_i = \mathbf{Q}\mathrm{e} = (3, 1, 2)'$ and it represents the places in the list where are located the most preferred item, the second best, and the least preferred item, respectively.

3. A multivariate model for the choice process

In order to develop a multivariate model for the rankings of m items as expressed by n raters, we can think of the choice process as a four stages procedure, based on the *sequential comparison criterion.*

1. Every rater acts by assigning a rank to each of the m items. If we consider this action separately for a given item $\mathcal{O}_j$, the rater's behaviour can be assimilated to a drawing without replacement from an urn, and modeled by an IHG random variable Thus, if we let S_j be the rank random variable assigned by the raters to the item $\mathcal{O}_j$, we can assume that: $S_j \sim IHG(m, \theta_j)$, $j = 1, 2, ..., m$.

2. During the process of ranking the m items, every rater starts from the most preferred one, then selects the second best, and so on, up to the worst. Thus, we have to model the index random variable $(\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_m)$, that represent the positions of the items in the list, from the most preferred to the worst. Since each single component has a conditioned (restricted) domain of definition, on the basis of the values taken by the previous components, a consistent multivariate pmf could be:

$$Pr\left(\mathcal{W}_1 = w_1, \mathcal{W}_2 = w_2, ..., \mathcal{W}_m = w_m\right) =$$

$$= Pr(S_1 = w_1)\,\frac{Pr(S_2 = w_2)}{1 - Pr(S_2 = w_1)} \cdots \frac{Pr(S_{m-1} = w_{m-1})}{1 - \sum_{j=1}^{m-2} Pr(S_{m-1} = w_j)}.$$

   In this way:

   - we obtain a multivariate $(m-1)$-dimensional distribution, since
   $$Pr(\mathcal{W}_m = w_m \mid \text{past } (m-1) \text{ choices}) = 1;$$
   - the $\mathcal{W}_k$ random variable cannot take the values observed for the previous index random variable;
   - the univariate distribution of the preferred item is IHG, since its elicitation is not conditioned by the other choices.

3. Exploiting the relation $(R_j = k) \Leftrightarrow (\mathcal{W}_k = j)$, we have a one-to-one correspondence between the multivariate $(\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_m)$ and $(R_1, R_2, ..., R_m)$ random variable. Indeed:

$$\begin{gathered} Pr\left(\mathcal{W}_1 = w_1, \mathcal{W}_2 = w_2, ..., \mathcal{W}_m = w_m\right) = \\ = Pr\left(R_{w_1} = 1, R_{w_2} = 2, ..., R_{w_m} = m\right); \end{gathered}$$

or, alike:

$$Pr\,(R_1 = r_1, R_2 = r_2, ..., R_m = r_m) =$$
$$= Pr\,(\mathcal{W}_{r_1} = 1, \mathcal{W}_{r_2} = 2, ..., \mathcal{W}_{r_m} = m)\,.$$

4. The multivariate distribution of $(R_1, R_2, ..., R_m)$ and that of the univariate R_j, $j = 1, 2, ..., m$, are strictly implied by the previous steps. Moreover, only $\mathcal{W}_1 \sim IHG(m, \theta_1)$, while the pmf of each other component must be computed as a marginal distribution of the multivariate random variable

The previous scheme is developed in some detail in the next section for the case of m=3 items.

4. The choice among m = 3 items

Let us consider m = 3 items, which we call for simplicity {ABC}, whose indexes in the ordered list are {123}.

On the basis of the previous assumptions, the allotment of the ranks to A, B, or C, separately, is modeled as:

$$S_1 \sim IHG(3, \theta_1);\;\; S_2 \sim IHG(3, \theta_2);\;\; S_3 \sim IHG(3, \theta_3);$$

where the preference for each item is measured by $\theta_1, \theta_2, \theta_3$, respectively.

As a matter of fact:

$$\theta_1 = Pr(S_1 = 1); \quad \theta_2 = Pr(S_2 = 1); \quad \theta_3 = Pr(S_3 = 1).$$

For a fixed item, the pmf of S turns out to be:

$$Pr\,(S = 1) = \theta;\;\; Pr\,(S = 2) = \frac{2\theta(1-\theta)}{1+\theta};\;\; Pr\,(S = 3) = \frac{(1-\theta)^2}{1+\theta}.$$

The linkage between the univariate choices and the multivariate process is possible by means of the relation:

$$Pr\,(\mathcal{W}_1 = w_1, \mathcal{W}_2 = w_2, \mathcal{W}_3 = w_3) = Pr(S_1 = w_1)\,\frac{Pr(S_2 = w_2)}{1 - Pr(S_2 = w_1)},$$

which is defined for any triplet (w_1, w_2, w_3) such that $w_j = 1, 2, 3, \forall j = 1, 2, 3;\ w_1 + w_2 + w_3 = 6.$

In fact, the rater selects, first of all, his/her most preferred item, whose pmf is $S_1 \sim IHG(3, \theta_1);$ then he/she chooses between the (two) remaining items, so that the second selection is represented by a random variable that can take all the values $\{1, 2, 3\}$ except the one taken by the first random variable This means that the second component of the multivariate distribution is conditioned by S_1 in terms of a restriction of its range, leading to a censored pmf.

Notice that both $\mathcal{W}_1$ and S_1 are identical and belong to the IHG random variable family; the random variable S_2 is IHG, while $\mathcal{W}_2$ is not.

It is also important to stress that, in our scheme, we are modelling the indexes and not the ranks. This implies that the rater selects as most preferred item the one located in the place w_1 of the ordered list; then, between the two remaining, he/she selects the one located in the place w_2; the last choice is determined by the relation: $\mathcal{W}_3 = 6 - \mathcal{W}_1 - \mathcal{W}_2$. Thus the trivariate random variable $(\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3)$ is a degenerate random variable and lies in a 2-dimensional space, since the third component is univocally determined.

Formally, for m = 3, there are $card(\mathcal{P}_3) = 3! = 6$ permutations, with the following probabilities:

$$\begin{aligned}\{123\} \to \quad & Pr\left(\mathcal{W}_1 = 1, \mathcal{W}_2 = 2, \mathcal{W}_3 = 3\right) = \\ & = Pr\left(S_1 = 1\right) \tfrac{Pr(S_2=2)}{1-Pr(S_2=1)} = \tfrac{2\,\theta_1\theta_2}{1+\theta_2};\end{aligned}$$

$$\begin{aligned}\{132\} \to \quad & Pr\left(\mathcal{W}_1 = 1, \mathcal{W}_2 = 3, \mathcal{W}_3 = 2\right) = \\ & = Pr\left(S_1 = 1\right) \tfrac{Pr(S_2=3)}{1-Pr(S_2=1)} = \tfrac{\theta_1(1-\theta_2)}{1+\theta_2};\end{aligned}$$

$$\begin{aligned}\{213\} \to \quad & Pr\left(\mathcal{W}_1 = 2, \mathcal{W}_2 = 1, \mathcal{W}_3 = 3\right) = \\ & = Pr\left(S_1 = 2\right) \tfrac{Pr(S_2=1)}{1-Pr(S_2=2)} = \tfrac{2\theta_1\theta_2(1-\theta_1)(1+\theta_2)}{(1+\theta_1)\left(1-\theta_2+2\theta_2^2\right)};\end{aligned}$$

$$\{231\} \rightarrow \qquad Pr\left(\mathcal{W}_1 = 2, \mathcal{W}_2 = 3, \mathcal{W}_3 = 1\right) =$$
$$= Pr\left(S_1 = 2\right) \tfrac{Pr(S_2=3)}{1-Pr(S_2=2)} = \tfrac{2\theta_1(1-\theta_1)(1-\theta_2)^2}{(1+\theta_1)\left(1-\theta_2+2\theta_2^2\right)};$$

$$\{312\} \rightarrow \qquad Pr\left(\mathcal{W}_1 = 3, \mathcal{W}_2 = 1, \mathcal{W}_3 = 2\right) =$$
$$= Pr\left(S_1 = 3\right) \tfrac{Pr(S_2=1)}{1-Pr(S_2=3)} = \tfrac{(1-\theta_1)^2(1+\theta_2)}{(1+\theta_1)(3-\theta_2)};$$

$$\{321\} \rightarrow \qquad Pr\left(\mathcal{W}_1 = 3, \mathcal{W}_2 = 2, \mathcal{W}_3 = 1\right) =$$
$$= Pr\left(S_1 = 3\right) \tfrac{Pr(S_2=2)}{1-Pr(S_2=3)} = \tfrac{2(1-\theta_1)^2(1-\theta_2)}{(1+\theta_1)(3-\theta_2)}.$$

The previous results can be summarized by the probability distribution of the $(\mathcal{W}_1, \mathcal{W}_2)$ bivariate random variable, that -with the respective marginals induced by $(\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3)$- is shown in the following table.

$\mathcal{W}_1\downarrow\ \mathcal{W}_2\rightarrow$	*1*	*2*	*3*	*Marginals*
1	0	$\frac{2\theta_1\theta_2}{1+\theta_2}$	$\frac{\theta_1(1-\theta_2)}{1+\theta_2}$	θ_1
2	$\frac{2\theta_1\theta_2(1-\theta_1)(1+\theta_2)}{(1+\theta_1)\left(1-\theta_2+2\theta_2^2\right)}$	0	$\frac{2\theta_1(1-\theta_1)(1-\theta_2)^2}{(1+\theta_1)\left(1-\theta_2+2\theta_2^2\right)}$	$\frac{2\theta_1(1-\theta_1)}{1+\theta_1}$
3	$\frac{(1-\theta_1)^2(1+\theta_2)}{(1+\theta_1)(3-\theta_2)}$	$\frac{2(1-\theta_1)^2(1-\theta_2)}{(1+\theta_1)(3-\theta_2)}$	0	$\frac{(1-\theta_1)^2}{1+\theta_1}$
Marginals	$p_1(\theta_1, \theta_2)$	$p_2(\theta_1, \theta_2)$	$p_3(\theta_1, \theta_2)$	1

Here, for simplicity, we let:

$$\begin{aligned}
p_1(\theta_1, \theta_2) &= \frac{(1-\theta_1)(1+\theta_2)\left(7\theta_1\theta_2 - 4\theta_1\theta_2^2 + 1 - \theta_2 + 2\theta_2^2 - \theta_1\right)}{(1+\theta_1)(3-\theta_2)\left(1-\theta_2+2\theta_2^2\right)};\\
p_2(\theta_1, \theta_2) &= 2\,\frac{3\theta_1\theta_2 + \theta_1\theta_2^2 + 3\theta_1^2\theta_2 - 2\theta_1^2\theta_2^2 + 1 - \theta_2^2 - 2\,\theta_1 + \theta_1^2}{(1+\theta_1)(1+\theta_2)(3-\theta_2)};\\
p_3(\theta_1, \theta_2) &= \frac{\theta_1(1-\theta_2)\left(3 - \theta_1 - \theta_2 - \theta_1\theta_2 + 4\theta_1\theta_2^2\right)}{(1+\theta_1)(1+\theta_2)\left(1-\theta_2+2\theta_2^2\right)}.
\end{aligned}$$

By a cumbersome algebra, we checked[2] that: $\sum_{j=1}^{3} p_j(\theta_1,\theta_2) = 1$.

[2] In this work, all the algebric developments have been checked by the symbolic language Maple V© (Waterloo Maple Inc., 1998).

The correspondence between $(\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3)$ and (R_1, R_2, R_3) is based on the relation above discussed and, in this case, it results:

$(\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3) \to$	$\{123\}$	$\{132\}$	$\{213\}$	$\{231\}$	$\{312\}$	$\{321\}$
$(R_1, R_2, R_3) \to$	$\{123\}$	$\{132\}$	$\{213\}$	$\{312\}$	$\{231\}$	$\{321\}$

Consequently, we can deduce a new table for the bivariate distribution of (R_1, R_2).

$R_1 \downarrow R_2 \to$	*1*	*2*	*3*	*Marginals*
1	0	$\dfrac{2\theta_1\theta_2}{1+\theta_2}$	$\dfrac{\theta_1(1-\theta_2)}{1+\theta_2}$	θ_1
2	$\dfrac{2\theta_1\theta_2(1-\theta_1)(1+\theta_2)}{(1+\theta_1)\left(1-\theta_2+2\theta_2^2\right)}$	0	$\dfrac{(1-\theta_1)^2(1+\theta_2)}{(1+\theta_1)\,(3-\theta_2)}$	$t_2(\theta_1,\theta_2)$
3	$\dfrac{2\theta_1\,(1-\theta_1)\,(1-\theta_2)^2}{(1+\theta_1)\left(1-\theta_2+2\theta_2^2\right)}$	$\dfrac{2(1-\theta_1)^2(1-\theta_2)}{(1+\theta_1)\,(3-\theta_2)}$	0	$t_3(\theta_1,\theta_2)$
Marginals	$q_1(\theta_1,\theta_2)$	$q_2(\theta_1,\theta_2)$	$q_3(\theta_1,\theta_2)$	1

Here, for simplicity, we let:

$$\begin{aligned}
q_1(\theta_1,\theta_2) &= 2\,\frac{\theta_1\,(1-\theta_1)}{(1+\theta_1)} \\
q_2(\theta_1,\theta_2) &= 2\frac{3\theta_1\theta_2+\theta_1\theta_2^2+3\theta_1^2\theta_2-2\theta_1^2\theta_2^2+1-\theta_2^2-2\theta_1+\theta_1^2}{(1+\theta_1)\,(1+\theta_2)\,(3-\theta_2)} \\
q_3(\theta_1,\theta_2) &= \frac{8\theta_1\theta_2-\theta_1-4\theta_1^2+2\theta_1^2\theta_2+\theta_1\theta_2^2-2\theta_1^2\theta_2^2-1-2-2\theta_2^2}{(1+\theta_1)\,(1+\theta_2)\,(3-\theta_2)} \\
t_2(\theta_1,\theta_2) &= \frac{(1-\theta_1)\,(1+\theta_2)\,(7\theta_1\theta_2-4\theta_1\theta_2^2+1-\theta_2+2\theta_2^2-\theta_1)}{(1+\theta_1)\,(3-\theta_2)\,(1-\theta_2+2\theta_2^2)}; \\
t_3(\theta_1,\theta_2) &= 2\frac{(1-\theta_1)\,(1-\theta_2)\,(2\theta_1-3\theta_1\theta_2-\theta_1\theta_2^2+1-\theta_2+2\theta_2^2)}{(1+\theta_1)\,(3-\theta_2)\,(1-\theta_2+2\theta_2^2)}.
\end{aligned}$$

Again we checked that:

$$\sum_{j=1}^{3} q_j(\theta_1,\theta_2) = 1; \quad \theta_1 + t_2(\theta_1,\theta_2) + t_3(\theta_1,\theta_2) = 1\,.$$

Since from the previous tables $p_2(\theta_1,\theta_2) \equiv q_2(\theta_1,\theta_2)$, it results:

$$Pr\,(\mathcal{W}_2 = 2) \equiv Pr(R_2 = 2);$$

and it is confirmed that:

$Pr(\mathcal{W}_1\!=\!1)\!\equiv\!Pr(R_1\!=\!1);\ Pr(\mathcal{W}_1\!=\!2)\!\equiv\!Pr(R_2\!=\!1);\ Pr(\mathcal{W}_2\!=\!1)\!\equiv\!Pr(R_1\!=\!2).$

Of course, we can also express the bivariate distribution of $(\mathcal{W}_k, \mathcal{W}_3)$, $k = 1, 2$, and of (R_j, R_3), $j = 1, 2$; moreover, it is possible to obtain the marginal distributions of $\mathcal{W}_3$ and of R_3, respectively. Indeed, we have:

$$\begin{gathered} Pr\left(\mathcal{W}_3 = 1\right) = \\ = Pr\{[(\mathcal{W}_1 = 2) \cap (\mathcal{W}_2 = 3)] \cup [(\mathcal{W}_1 = 3) \cap (\mathcal{W}_2 = 2)]\} \\ = 2\,\tfrac{(1-\theta_1)(1-\theta_2)\left(2\theta_1-3\theta_1\theta_2-\theta_1\theta_2^2+1-\theta_2+2\theta_2^2\right)}{(1+\theta_1)(3-\theta_2)\left(1-\theta_2+2\theta_2^2\right)}\ ; \end{gathered}$$

and similarly:

$$Pr\left(\mathcal{W}_3 = 2\right) = \frac{8\theta_1\theta_2 - \theta_1 - 4\theta_1^2 + 2\theta_1^2\theta_2 + \theta_1\theta_2^2 - 2\theta_1^2\theta_2^2 - 1 - 2 - 2\theta_2^2}{\left(1+\theta_1\right)\left(1+\theta_2\right)\left(3-\theta_2\right)};$$

$$Pr\left(\mathcal{W}_3 = 3\right) = 2\theta_1\theta_2 \frac{2 + \theta_2 + 3\theta_2^2 - 3\theta_1\theta_2 + \theta_1\theta_2^2}{\left(1+\theta_1\right)\left(1+\theta_2\right)\left(1-\theta_2+2\theta_2^2\right)}.$$

In the same way, we get the distribution of R_3:

$$\begin{aligned} Pr\left(R_3\!=\!1\right) &= Pr\{(R_1\!=\!2) \cap (R_2\!=\!3)] \cup [(R_1\!=\!3) \cap (R_2\!=\!2)]\} = \\ &= \frac{\left(1-\theta_1\right)^2}{1+\theta_1}; \\ Pr\left(R_3 = 2\right) &= \frac{\theta_1\left(1-\theta_2\right)\left(3-\theta_1-\theta_2-\theta_1\theta_2+4\theta_1\theta_2^2\right)}{\left(1+\theta_1\right)\left(1+\theta_2\right)\left(1-\theta_2+2\theta_2^2\right)}; \\ Pr\left(R_3 = 3\right) &= 2\theta_1\theta_2 \frac{2+\theta_2+3\theta_2^2-3\theta_1\theta_2+\theta_1\theta_2^2}{\left(1+\theta_1\right)\left(1+\theta_2\right)\left(1-\theta_2+2\theta_2^2\right)}. \end{aligned}$$

Finally, in summary, we obtain the marginal distribution of the random variable R_j, $j = 1, 2, 3$:

	R_1	R_2	R_3
1	θ_1	$2\,\frac{\theta_1(1-\theta_1)}{(1+\theta_1)}$	$\frac{(1-\theta_1)^2}{1+\theta_1}$
2	$t_2(\theta_1, \theta_2)$	$q_2(\theta_1, \theta_2)$	$\frac{\theta_1(1-\theta_2)\left(3-\theta_1-\theta_2-\theta_1\theta_2+4\theta_1\theta_2^2\right)}{(1+\theta_1)(1+\theta_2)\left(1-\theta_2+2\theta_2^2\right)}$
3	$t_3(\theta_1, \theta_2)$	$q_3(\theta_1, \theta_2)$	$2\theta_1\theta_2\frac{2+\theta_2+3\theta_2^2-3\theta_1\theta_2+\theta_1\theta_2^2}{(1+\theta_1)(1+\theta_2)\left(1-\theta_2+2\theta_2^2\right)}$

The developments discussed above can be shown by the help of a numerical example. For instance, let $\theta_1 = 1/2; \theta_2 = 1/2$; that is: $S_1 \sim IHG(3, 1/2)$, $S_2 \sim IHG(3, 1/2)$.

Then, the following bivariate distributions for $(\mathcal{W}_1, \mathcal{W}_2)$ and (R_1, R_2) result from te previous tables:

$\mathcal{W}_1 \downarrow \mathcal{W}_2 \rightarrow$	1	2	3	*Marginals*
1	0	$1/3$	$1/6$	$1/2$
2	$1/4$	0	$1/12$	$1/3$
3	$1/10$	$1/15$	0	$1/6$
Marginals	$7/20$	$2/5$	$1/4$	1

$R_1 \downarrow R_2 \rightarrow$	1	2	3	*Marginals*
1	0	$1/3$	$1/6$	$1/2$
2	$1/4$	0	$1/10$	$7/20$
3	$1/12$	$1/15$	0	$3/20$
Marginals	$1/3$	$2/5$	$4/15$	1

On this basis and exploiting the fact that:

$$\mathcal{W}_3 = 6 - \mathcal{W}_1 - \mathcal{W}_2; \ \ R_3 = 6 - R_1 - R_2,$$

we can obtain the marginal distribution of $\mathcal{W}_3$ and R_3. More explicitly, it results:

$$Pr\,(\mathcal{W}_3 = 1) = Pr\{[(\mathcal{W}_1 = 2) \cap (\mathcal{W}_2 = 3)] \cup [(\mathcal{W}_1 = 3) \cap (\mathcal{W}_2 = 2)]\} = \frac{3}{20};$$

$$Pr\,(\mathcal{W}_3 = 2) = \frac{4}{15}; \qquad Pr\,(\mathcal{W}_3 = 3) = \frac{7}{12}.$$

In the same way:

$$Pr\,(R_3 = 1) = Pr\{[(R_1 = 2) \cap (R_2 = 3)] \cup [(R_1 = 3) \cap (R_2 = 2)]\} = \frac{1}{6};$$

$$Pr\,(R_3 = 2) = \frac{1}{4}; \qquad Pr\,(R_3 = 3) = \frac{7}{12}.$$

Moreover, the correspondence between ranks and indexes is confirmed:

$$Pr\,(R_3 = 1) = Pr\,(\mathcal{W}_1 = 3) = \frac{1}{6};$$
$$Pr\,(R_3 = 2) = Pr\,(\mathcal{W}_2 = 3) = \frac{1}{4};$$
$$Pr\,(R_3 = 3) = Pr\,(\mathcal{W}_3 = 3) = \frac{7}{12}.$$

In this way, the marginal distributions of the univariate rank random variable are:

	R_1	R_2	R_3
1	1/3	1/2	1/6
2	2/5	7/20	1/4
3	4/15	3/20	7/12

5. Inferential and computational issues

Let $(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n)$ be the observed rankings expressed by a sample of n raters, where $\mathbf{r}_i = (r_{i1}, r_{i2}, ..., r_{ij}, ..., r_{im})'$, $i = 1, 2, ..., n$, represents the ranks assigned to the m items by the i-th rater.

On the basis of the previous defined correspondence, we can consider, alike, a sample of indexes vectors $(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n)$, where each vector, $\mathbf{w}_i = (w_{i1}, w_{i2}, ..., w_{ij}, ..., w_{im})'$, represents the observed order of choice for the i-th rater.

Thus, the parameters $\theta = (\theta_1, \theta_2, ..., \theta_{m-1})'$ of the multivariate distribution can be estimated maximizing the log-likelihood function:

$$log\mathcal{L}(\theta; \mathbf{w}) = \sum_{i=1}^{n} log\mathcal{L}_i(\theta; \mathbf{w}_i),$$

where

$$log\mathcal{L}_i(\theta; \mathbf{w}_i) = \sum_{j=1}^{m-1} l_j,$$

and

$$\begin{aligned} l_1 &= log[Pr(S_1 = w_1)], \\ l_{j \neq 1} &= log[Pr(S_j = w_j)] - log\left[1 - \sum_{jj=1}^{j-1} Pr(S_j = w_{jj})\right]. \end{aligned}$$

Then, in the case of m = 3 items, the i-th unit contribution to the log-likelihood function reduces to:

$$log\mathcal{L}_i(\theta; \mathbf{w}_i) = l_1 + l_2,$$

where

$$l_2 = log[Pr(S_2 = w_2)] - log[1 - Pr(S_2 = w_1)].$$

From a computational point of view, it is worth noticing that the log-likelihood function, shown above, can be quite easily maximized by means of numerical optimization algorithms (e.g. Newton-Raphson, etc.), which are often available in most of the statistical software and/or programming language (e.g. Gauss©, by Aptech, 1995). Thus, the only task is to compute the log-likelihood function. In our case, this can be easily accomplished by means of the recursive formula:

$$Pr(S_j = s + 1) = Pr(S_j = s)\frac{(1 - \theta_j)(m - s)}{m - 1 - s + s\theta_j}, \quad s = 1, 2, ..., m - 1,$$

with $Pr(S_j = 1) = \theta_j$, which makes really fast the computation of the probabilities involved in the likelihood function.

However, with regard to the speedy of convergence of the optimization algorithm, a proper choice of starting values for the parameters plays an important role, too. Some preliminary experiences on real data sets showed us that the univariate estimates of θ_1, ..., θ_{m-1}, might be effectively used to initialize the maximization routine.

Finally, some care is needed as far as concerns the meaning of the estimated parameters. The case of m=3 items, $\{$ABC$\}$, can be considered in order to stress some aspects of the problem, although they are still valid for $m > 3$.

In fact, the estimate $\hat{\theta}_1$ represents a liking/agreement measure for the item A, since it is the estimated probability that the rank of A is 1, that is $Pr(R_1 = 1)$. This is true both in a univariate and in a multivariate approach, since the allotment of the rank to the most preferred item is not conditioned on previous choices.

Instead, this is not the case for the estimate $\hat{\theta}_2$. We could consider $\hat{\theta}_2$ a liking measure for the item B from a univariate point of view, but we cannot state the same in a multivariate approach. Indeed, as it is evident from the table on page 22, the estimated $Pr(R_2 = 1)$ is not a function of $\hat{\theta}_2$, but only of $\hat{\theta}_1$, and the same happens for the estimated $Pr(R_3 = 1)$. For this reason, it does not matter that we cannot estimate $\hat{\theta}_3$ (the third component of the multivariate distribution is a degenerate r.v), since we infer the liking measure for C directly from $\hat{\theta}_1$.

On the other hand, $\hat{\theta}_1$ and $\hat{\theta}_2$ are both necessary to estimate the whole bivariate distribution of (R_1, R_2) as well as the marginal random variables R_1, R_2, R_3. In this way we can obtain the expected frequencies of each of the $m!$ rankings. Of course, these frequencies can be compared to the observed frequencies in the sample by means of a chi-square test, in order to check if the proposed model adequately represents the generating process of the rank data.

6. Final remarks and further developments

In this paper we have proposed a parametric multivariate model for studying the whole ranking of m items. This framework is mainly based on the distinction between rank and order data, which allows for overcoming some difficulties of the multivariate approach.

Moreover, this structure stands on a choice criterion -the *sequential comparisons*- which seems consistent with the real behaviour of the raters in several situations. This makes the model a useful tool in analyzing preferences, ratings, evaluations, and so on.

For sake of simplicity, we have just focused on the information inside the rank matrix, while it would be interesting to study also the characteristics of the raters in relation to the ranks they express. Of course, this means to develop a multivariate generalized linear model for the rankings, extending some previous results on some regression models for ranks (D'Elia, 2000b). With regard to this point, it would be also important to introduce a suitable measure of goodness of fit of the model and to develop some diagnostic tools.

Acknowledgements: This work has been financially supported by Italian MURST, University of Naples Federico II and University of Salerno funds. This paper is a marginally modified version of D'Elia (2001) which circulated in a very limited number of copies as a Working Paper. In order to give larger diffusion to the results of that work with permission of the Author, the original contribution is now reprinted in this journal.

References

Agresti A. (1984), *Analysis of ordinal categorical data*, John Wiley & Sons, New York.

Agresti A. (1992), Analysis of ordinal paired comparison data, *Applied Statistics*, 41, 287–297.

Aptech Systems, Inc. (1995), *Gauss-386i, Version 3.2.13*, Maple Valley, WA.

Bradley R. A. and Terry M. E. (1952), Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39, 324–345.

Daniels H.E. (1950), Rank correlation and population models (with discussion), *Journal of the Royal Statistical Society*, Series B, 1, 171–191.

Davidson R. R. and Bradley R. A. (1969), Multivariate paired comparisons: the extension of a univariate model and associated estimation and test procedures, *Biometrika*, 56, 81–95.

D'Elia A. (1999), A proposal for ranks statistical modelling, *Statistical Modelling* (Friedl, H., Berghold, A., Kauermann, G. eds.), Graz - Austria, 468–471.

D'Elia, A. (2001), Some methodological issues on multivariate modelling of rank data, *Working Paper* n. 3.99, Dipartimento di Scienze Economiche, Università degli Studi di Salerno.

D'Elia A. (2000a), A shifted binomial model for rankings, *Statistical Modelling*, (Nunez-Anton, V., Ferreira E. eds.), Servicio Editorial de la Universidad del Pais Vasco, 412–416.

D'Elia A. (2000b), Un modello lineare generalizzato per i ranghi: aspetti statistici, problemi computazionali e verifiche empiriche, *Italian Journal of Applied Statistics*, 12, 205–227.

D'Elia A. (2000c), Il meccanismo dei confronti appaiati nella modellistica per graduatorie: sviluppi statistici ed aspetti critici, *Quaderni di Statistica*, 2, 173–203.

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

Fligner M.A. and Verducci J.S. (1993), *Probability models and statistical analyses for ranking data*, Springer-Verlag, New York.

Guenther W.C. (1975), The Inverse Hypergeometric - a Useful Model, *Statistica Neerlandica*, 29, 129–144.

Henery R. J. (1991), The covariance of rank scores in order-statistics models, *Statistica Sinica*, 1, 301–308.

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

Marden J. I. (1995), *Analyzing and modeling rank data*, Chapman & Hall, London.

McCullagh P. (1980), Regression models for ordinal data, *Journal of the Royal Statistical Society*, Series B, 42, 109–127.

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

Quandt R. E. (1998), Measurement and inference in wine tasting, *Meetings of the Vineyard Data Quantification Society*, Corsica.

Sadooghi-Alvandi S. M. (1992), Testing for agreement among several groups of raters: a contingency-table approach, *Statistica Sinica*, 2, 285–296.

Tanner M. A. and Young M. A. (1985), Modeling agreement among raters, *Journal of the American Statistical Association*, 80, 175–180.

Taplin R. H. (1997), The statistical analysis of preference data, *Applied Statistics*, 46, 49–512.

Waterloo Maple Inc. (1998), *Maple V release 5.1.*