

## A new approach to inter-rater agreement through stochastic orderings: the discrete case

Alessandra Giovagnoli · Johnny Marzioletti · Henry P. Wynn

Received: 28 December 2006 / Published online: 26 June 2007  
© Springer-Verlag 2007

**Abstract** We wish to study inter-rater agreement comparing groups of observers who express their ratings on a discrete or ordinal scale. The starting point is that of defining what we mean by “agreement”. Given  $d$  observers, let the scores they assign to a given statistical unit be expressed as a  $d$ -vector in the real space. We define a deterministic ordering among these vectors, which expresses the degree of the raters’ agreement. The overall scoring of the raters on the sample space will be a  $d$ -dimensional random vector. We then define an associated partial ordering among the random vectors of the ratings, illustrate a number of its properties, and look at order-preserving functions (agreement measures). In this paper we also show how to test the hypothesis of greater agreement against the unrestricted hypothesis, and the hypothesis of equal agreement against the hypothesis that an agreement ordering holds. The test is applied to real data on two medical observers rating clinical guidelines.

**Keywords** Raters’ agreement · Stochastic orderings · Chi-bar-squared distribution · Contingency tables · Discrete random variables

---

A. Giovagnoli (✉) · J. Marzioletti  
Department of Statistical Sciences, University of Bologna, Bologna, Italy  
e-mail: alessandra.giovagnoli@unibo.it

J. Marzioletti  
e-mail: johnny.marzioletti@unibo.it

H. P. Wynn  
London School of Economics and Political Science, London, UK  
e-mail: h.wynn@lse.ac.uk

## 1 Introduction

Classification and rating are basic in all scientific fields, and often there is the need to test the reliability of a classification process by assessing the level of agreement between two or more different observers (the raters) that classify the same group of statistical units (the subjects). In the literature the extent of inter-rater agreement is studied mainly by means of indicators: Cohen's Kappa is the most popular one, especially for two raters and a categorical scale of measurement (Cohen 1960). Another way is the use of statistical models to describe the "structure" of this relationship, mainly log-linear and latent class models (Banerjee et al. 1999).

Classification can take place on a set of nominal or ordered categories or on a discrete or continuous scale: our approach is general but it focuses on the discrete case. The approach of this article is the following: given the sample space (the subjects) to each observer there corresponds the probability distribution of the scores of his/her ratings. Thus an observer gives rise to a random variable with values in the set  $\mathcal{C}$  of all possible ratings and the agreement among a group of observers is a type of association among the jointly distributed random ratings of the various observers (Bishop et al. 1975). The question "when does one group show more agreement than another?" is answered defining a special order relation among multidimensional random variables (see also Giovagnoli 2002). Indicators of agreement will then be all the real-valued functions preserving the ordering under consideration.

Observe that this description fits three different situations, namely when two distinct groups of  $d$  raters each classify the same subjects, or the same raters classify two samples of subjects, or the same raters classify the same subjects at different times.

There are many fields of study where raters' agreement is applied: for example, in medicine it is used to assess the reliability of diagnoses; in psychiatry and in cognitive research it is applied in evaluating the quality of classification and coding systems; in multi-centre clinical trials it is used to improve the quality of data, by preceding the beginning of data collection by an inter-rater reliability study.

This article is structured as follows: in Sect. 2 we define an equivalence relation and a deterministic ordering for the  $d$ -dimensional vectors representing the ratings of the observers classifying just one subject. In Sect. 3 we define a stochastic agreement ordering among the  $d$ -dimensional random vectors representing the joint probability distributions of ratings of the observers. In Sects. 4 and 5 we concentrate our attention on the special cases of  $d = 2$  and  $d = 3$  raters, which are of wide use in practice. In Sect. 6 we explain how to deal with observed data, indicating some hypothesis testing problems in order to compare two  $d$ -way tables of joint rating frequencies to decide if the agreement ordering holds or not. Finally, in Sect. 7 we give an example to see how to apply this methodology to real data. There is an Appendix at the end of the paper with a review of order relations (Part A) and a review of chi-bar-squared random variables (Part B).

## 2 Agreement among observers rating just one subject

Assume the vector  $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$  in a set  $\mathcal{S}$  expresses numerically the ratings of the same statistical unit by  $d$  observers. If the scores are on a continuous scale, let

$\mathcal{S} = \mathbb{R}^d$ , if on a  $1 - to - m$  scale, let  $\mathcal{S} = (1, 2, \dots, m)^d$ , etc. Reasonable demands for an agreement ordering  $\geq_A$  are:

- A1 The maximum extent of agreement is reached when all the observers rate the subject in the same way.
- A2 Agreement does not change after permuting the observers.
- A3 Agreement increases if two of the observers change their judgments making them “closer” and leaving the sum of their scores unchanged.
- A4 The level of agreement does not change if all the observers increase (or decrease) their scores by the same amount.

From Axioms A2, A4 we can derive a definition of equivalence in terms of agreement among the vectors of ratings:

**Definition 1** Given two vectors  $\mathbf{x}$  and  $\mathbf{y}$  belonging to  $\mathcal{S}$ , we say that  $\mathbf{x}$  is equivalent to  $\mathbf{y}$  in terms of agreement, and write  $=_A$ , if and only if the following condition holds:

$$\mathbf{x} =_A \mathbf{y} \Leftrightarrow \mathbf{x} = \Pi(\mathbf{y} + k \cdot \mathbf{1}_d) = \Pi \cdot \mathbf{y} + k \cdot \mathbf{1}_d \tag{1}$$

with  $\Pi$  a permutation matrix,  $\mathbf{1}_d = (1, 1, \dots, 1)^T$ , and  $k \in \mathbb{R}$  such that  $(\mathbf{y} + k\mathbf{1}_d) \in \mathcal{S}$ .

For example, if there are  $d = 3$  observers rating two subjects on a  $1 - to - 5$  scale assigning the scores  $(3 \ 5 \ 5)^T$  and  $(3 \ 1 \ 3)^T$  respectively, then the two ratings express the same amount of agreement because the first vector can be obtained from the second by means of a permutation of its elements and adding two points to all the scores:

$$\begin{pmatrix} 3 \\ 5 \\ 5 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 3 \end{pmatrix} + 2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

We now show that Axioms A1–A4 uniquely define a pre-order relation  $\geq_A$  (see Appendix A). It is well-known (Marshall and Olkin 1979) that axioms A1, A2, A3 characterize the opposite of the so-called majorization ordering  $<_m$ :

**Definition 2** Given two vectors in  $\mathcal{S} = \mathbb{R}^d$ ,  $\mathbf{x} = (x_1, \dots, x_d)^T$  and  $\mathbf{y} = (y_1, \dots, y_d)^T$ ,  $\mathbf{x}$  is majorized by  $\mathbf{y}$  ( $\mathbf{x} <_m \mathbf{y}$ ) if and only if:

$$\left\{ \begin{array}{l} \sum_{i=1}^k \mathbf{x}_{[i]} \leq \sum_{i=1}^k \mathbf{y}_{[i]} \quad \text{where } k = 1, \dots, d - 1 \text{ and } \mathbf{x}_{[1]} \geq \mathbf{x}_{[2]} \geq \dots \geq \mathbf{x}_{[d]} \\ \sum_{i=1}^d \mathbf{x}_{[i]} = \sum_{i=1}^d \mathbf{y}_{[i]} \end{array} \right. \tag{2}$$

An equivalent condition is that  $\mathbf{x}$  is a convex combination of the vectors obtained by permuting the coordinates of  $\mathbf{y}$ :

$$\mathbf{x} <_m \mathbf{y} \Leftrightarrow \mathbf{x} = \sum_{i=1}^{d!} \alpha_i (\Pi_i \mathbf{y}) \tag{3}$$

with  $\alpha_i \geq 0$  for  $i = 1, \dots, d!$  and  $\sum_{i=1}^{d!} \alpha_i = 1$ .

So, for example, the vector  $(3\ 2\ 4)^T$  is majorized by  $(5\ 1\ 3)^T$  because, after ordering their coordinates in non increasing order, the conditions of Definition 2 are satisfied:

$$\begin{cases} 4 \leq 5 \\ 4 + 3 \leq 5 + 3 \\ 4 + 3 + 2 = 5 + 3 + 1 \end{cases}$$

or, equivalently, we can state that the vector  $(3\ 2\ 4)^T$  can be obtained by a convex combination of the vectors resulting from permutations of the entries of  $(5\ 1\ 3)^T$ :

$$\begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix} = \alpha_1 \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix} + \alpha_2 \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 5 \\ 3 \end{pmatrix} + \alpha_4 \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} + \alpha_5 \begin{pmatrix} 3 \\ 5 \\ 1 \end{pmatrix} + \alpha_6 \begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix}$$

with  $\alpha_1 = \alpha_4 = \frac{1}{2}$  and  $\alpha_2 = \alpha_3 = \alpha_5 = \alpha_6 = 0$  being one of the possible solutions. An interpretation of this ordering is that the coordinates of a certain vector are ‘more concentrated’ with respect to those of any other vector which is greater in terms of majorization. Therefore, the first of the two vectors will mean more agreement than the second one.

Given a group  $\mathcal{G}$  of linear transformations of  $S = \mathbb{R}^d$  one can define a majorization ordering with respect to it, called *G-majorization* (see Giovagnoli and Wynn 1985):

**Definition 3**

$$\mathbf{x} <_G \mathbf{y} \Leftrightarrow \mathbf{x} = \sum_{i=1}^r \alpha_i g_i(\mathbf{y}) \tag{4}$$

with  $g_i \in \mathcal{G}, \alpha_i \geq 0$  for  $i = 1, \dots, r$  and  $\sum_{i=1}^r \alpha_i = 1$ .

Let us specialize  $\mathcal{G}$  to be the group of permutations and simultaneous shifts of all the vector coordinates, then Definition 1 implies that  $\mathbf{x}$  is equivalent to  $\mathbf{y}$  if and only if  $\mathbf{x}$  is obtained from  $\mathbf{y}$  by means of a transformation in  $\mathcal{G}$  and

$$\mathbf{x} <_G \mathbf{y} \Leftrightarrow \mathbf{x} = \sum_{i=1}^{d!} \alpha_i (\Pi_i \mathbf{y} + k \mathbf{1}_d) = \sum_{i=1}^{d!} \alpha_i \Pi_i \mathbf{y} + k \mathbf{1}_d. \tag{5}$$

For example, the vector  $(2\ 3\ 1)^T$  is G-majorized by the vector  $(1\ 5\ 3)^T$  because

$$\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} = \alpha_1 \begin{pmatrix} 1 \\ 5 \\ 3 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix} + \alpha_3 \begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix} + \alpha_4 \begin{pmatrix} 3 \\ 5 \\ 1 \end{pmatrix} + \alpha_5 \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix} + \alpha_6 \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix} + k \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

with  $k = -1, \alpha_1 = \alpha_6 = \frac{1}{2}$  and  $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$  as one of the possible solutions. Therefore, the G-majorization ordering so defined enables us to compare vectors not having the same sum of coordinates.

Furthermore, it can be shown that if we set

$$\bar{x} = \frac{1}{d} \sum_{i=1}^d x_i \text{ and } \mathbf{x}^* = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_d - \bar{x})^T,$$

so that the vector  $\mathbf{x}^*$  is the projection of  $\mathbf{x}$  onto the  $(d - 1)$ -dimensional hyperplane through the origin orthogonal to  $\mathbf{1}_d = (1, 1, \dots, 1)^T$ , then we get:

**Proposition 1** *If  $\geq_A$  is an order relation satisfying A1–A4, then*

$$\mathbf{x} \geq_A \mathbf{y} \Leftrightarrow \mathbf{x}^* \prec_m \mathbf{y}^* \Leftrightarrow \mathbf{x} \prec_G \mathbf{y} \tag{6}$$

*Proof* Vector  $\mathbf{x}^*$  ( $\mathbf{y}^*$ ) is obtained from vector  $\mathbf{x}$  ( $\mathbf{y}$ ) via a shift of the coordinates, therefore they are equivalent in terms of agreement by Axiom A4 and

$$\mathbf{x} \geq_A \mathbf{y} \Leftrightarrow \mathbf{x}^* \geq_A \mathbf{y}^*. \tag{7}$$

Now  $\mathbf{x}^*$  and  $\mathbf{y}^*$  are vectors with the same coordinate sum and by Axiom A3 agreement increases by a transfer from one coordinate to another and is invariant under permutations (Axiom A1). This is the property that characterizes the inverse majorization ordering (see Marshall and Olkin 1979). □

Proposition 1 can be used as an operative definition of the ordering  $\geq_A$ .

A large class of indicators consistent with the agreement ordering  $\geq_A$  is given by all concave, permutation-and-shift-invariant functions  $\Psi$ . For this and related results see Marshall and Olkin (1979). In particular

$$\mathbf{x}^* \prec_m \mathbf{y}^* \Leftrightarrow \sum_{i=1}^d \phi(x_i^*) \leq \sum_{i=1}^d \phi(y_i^*) \tag{8}$$

for all convex functions  $\phi$ , so that examples of measures of agreement (i.e. order preserving functions w.r.t.  $\geq_A$ ) are for instance all the decreasing functions of the variance

$$\frac{1}{d} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_d - \bar{x})^2]$$

or of the mean absolute deviation

$$\frac{1}{d} (|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_d - \bar{x}|).$$

### 3 Agreement ordering of multidimensional r.v’s.

The relation  $\geq_A$  defined in the previous section is a deterministic pre-ordering for vectors of ratings. On that basis we define a stochastic ordering for the probability distributions of the ratings of the observers.

**Definition 4** Given the pre-ordering  $\geq_A$  on the set  $\mathcal{S}$  and the  $d$ -dimensional random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  with values in  $\mathcal{S}$ , let

$$\mathbf{X} \geq_{Agr} \mathbf{Y} \Leftrightarrow Pr\{\mathbf{X} \geq_A \mathbf{z}\} \geq Pr\{\mathbf{Y} \geq_A \mathbf{z}\} \quad \forall \mathbf{z} \in \mathcal{S}. \tag{9}$$

This definition is in analogy with the definition of upper-orthant stochastic ordering for  $d$ -dimensional random variables (see for instance [Shaked and Shanthikumar 2007](#)), where  $\geq_A$  replaces the entry-wise ordering of vectors in  $d$ -dimensional real space. A very special case is when  $X$  and  $Y$  are jointly distributed, e.g. refer to observers rating the same subjects, and  $Pr(X \geq_A Y) = 1$ , since if  $X \geq_A Y$  *a.s.* then

$$Y \geq_A z \text{ implies } X \geq_A z \quad \forall z \in \mathcal{S}.$$

Hence

$$Pr\{Y \geq_A z\} \leq Pr\{X \geq_A z\} \quad \forall z \in \mathcal{S}.$$

Observe that Proposition 1 implies

$$X \geq_{Agr} Y \Leftrightarrow Pr\{X^* <_m z^*\} \geq Pr\{Y^* <_m z^*\} \tag{10}$$

$\forall z^* = z - \bar{z} \cdot \mathbf{1}_d$  such that  $z \in \mathcal{S}$ .

If the equality holds we obtain an equivalence relation that we call  $\cong_{Agr}$ .

Note that the probability in (9) can be expressed as

$$Pr\{X \geq_A z\} = \int \int \int_{A_z} F_x(\mathbf{x}) d\mathbf{x} = E[I_{A_z}(X)] \tag{11}$$

where  $A_z = \{\mathbf{a} : \mathbf{a}^* <_m z^*\}$  and  $I_{A_z}$  is the indicator function of the set  $A_z$ . Therefore for all  $z$

$$X \geq_{Agr} Y \Leftrightarrow E[I_{A_z}(X)] \geq E[I_{A_z}(Y)]. \tag{12}$$

From this expression it is possible to derive functions which preserve the ordering  $\geq_{Agr}$  defined in (4), for instance by taking linear positive combinations of indicator functions of sets  $A_z$ .

We now show how to check that (4) holds in the special (but important) case of only two raters. We restrict our attention to the situation of a discrete measurement scale with values from 1 to  $m$ .

#### 4 The case of two raters ( $d = 2$ )

We remark that in the case of just two raters the ordering  $\geq_A$  is a total ordering, i.e. every pair of vectors are comparable, namely

$$(x_1, x_2) \geq_A (y_1, y_2) \Leftrightarrow |x_1 - x_2| \leq |y_1 - y_2|. \tag{13}$$

The joint rating probabilities for the two raters can be expressed by an  $m \times m$  table  $P$ , where the entry  $p_{ij}$  stands for the probability that a generic subject is scored  $i$  by the first rater and  $j$  by the second rater. Moreover, let  $Q$  be another  $m \times m$  table of joint probabilities relative to another pair of raters. Applying Definition 4 it is straightforward to prove that:

**Proposition 2** Given two tables  $P = [p_{ij}]$  and  $Q = [q_{ij}]$   $i, j = 1, \dots, m$ , the agreement of  $P$  is higher — in the sense of  $\geq_{Agr}$  — than that of  $Q$  (we write  $P \geq_{Agr} Q$ ) if and only if

$$\sum_{h=0}^k tr_h P \geq \sum_{h=0}^k tr_h Q \quad k = 0, \dots, m - 2 \tag{14}$$

where

$$tr_h P = \sum_{i=1}^m \sum_{j=1}^m \sum_{|i-j|=h} p_{ij} = \begin{cases} \sum_{i=1}^m p_{ii} & \text{for } h = 0 \\ \sum_{i=1}^{m-h} p_{i,i+h} + \sum_{i=1}^{m-h} p_{i+h,i} & h = 1, \dots, m - 1 \end{cases}$$

*Proof* Because  $\leq_A$  is a total pre-ordering when  $d = 2$ , Definition 4 is also equivalent to

$$\mathbf{X} \geq_{Agr} \mathbf{Y} \Leftrightarrow \Pr(\mathbf{X} \leq_A \mathbf{z}, \mathbf{X} \neq_A \mathbf{z}) \leq \Pr(\mathbf{Y} \leq_A \mathbf{z}, \mathbf{Y} \neq_A \mathbf{z}).$$

Two scores  $(i, j)$  and  $(i', j')$  are equivalent if and only if  $|i - j| = |i' - j'|$  and thus

$$A_{(i,j)} = \{(i', j') \text{ s.t. } k = |i - j| \leq |i' - j'|\} \text{ and } E(I_{A_{(i,j)}}) = \sum_{h=0}^k tr_h P.$$

We now want to check how this definition fits in with existing measures of agreement. The *Weighted Total Proportion of Agreement* is the index

$$TPA_w = \sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{ij} \tag{15}$$

where the “weights”  $w_{ij}$  (Schouten 1982) are such that  $0 \leq w_{ij} \leq 1$  when  $i \neq j$ ,  $w_{ij} = w_{ji}$  and  $w_{ii} = 1$ . The weights are usually a decreasing function of the distance from the main diagonal, i.e.

$$w_{ij} = f(|i - j|) \quad \text{and} \quad f(0) \geq f(1) \geq f(2) \geq \dots \geq f(m - 1)$$

For example, Cicchetti (1972) suggests the following weights:

$$w_{ij} = 1 - \frac{|i - j|}{m - 1} \quad i, j = 1, \dots, m. \tag{16}$$

When this measure is chance-corrected, namely the amount of agreement obtained for the effect of chance alone is subtracted, and the result is normalized with respect to the maximum value it can assume, it gives rise to the Weighted Kappa introduced by Cohen (1968):

$$\kappa_w = \frac{\sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{ij} - \sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{i+} p_{+j}}{1 - \sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{i+} p_{+j}}. \tag{17}$$

The ordinary (unweighted) Kappa is when  $w_{ij} = 0$  for  $i \neq j$ . We state the following result:

**Proposition 3** *The index  $TPA_w$  is order-preserving with respect to  $\geq_{Agr}$*

$$P \geq_{Agr} Q \Rightarrow TPA_w(P) \geq TPA_w(Q) \tag{18}$$

*Proof* We need to show:

$$\sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{ij} \geq \sum_{i=1}^m \sum_{j=1}^m w_{ij} q_{ij} \quad i.e. \quad \sum_{i=1}^m \sum_{j=1}^m w_{ij} (p_{ij} - q_{ij}) \geq 0 \quad i.e.$$

$$\sum_{i=1}^m \sum_{j=1}^m f(|i - j|) (p_{ij} - q_{ij}) \geq 0 \quad i.e.$$

$$f(0) (tr_0 P - tr_0 Q) + f(1) (tr_1 P - tr_1 Q) + \dots + f(m-1) (tr_{m-1} P - tr_{m-1} Q) \geq 0.$$

Since

$$\begin{aligned} f(0) &= f(1) + a_1 \\ f(1) &= f(2) + a_2 \\ &\dots \\ f(m-2) &= f(m-1) + a_{m-1} \end{aligned}$$

where  $a_1, a_2, \dots, a_{m-1} \geq 0$ ,

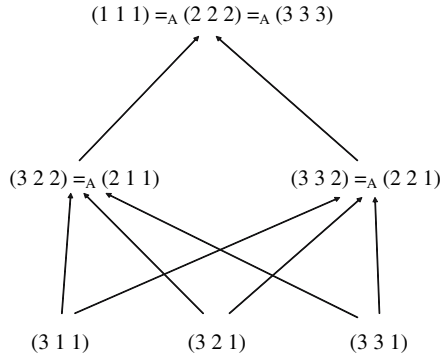
$$\begin{aligned} f(0) &= a_1 + a_2 + \dots + a_{m-1} + f(m-1) \\ f(1) &= a_2 + a_3 + \dots + a_{m-1} + f(m-1) \\ &\dots \\ f(m-2) &= a_{m-1} + f(m-1). \end{aligned}$$

Thus

$$\begin{aligned} &f(0) (tr_0 P - tr_0 Q) + f(1) (tr_1 P - tr_1 Q) + \dots + f(m-1) (tr_{m-1} P - tr_{m-1} Q) \\ &= f(m-1) \sum_{h=0}^{m-1} (tr_h P - tr_h Q) + a_{m-1} \sum_{h=0}^{m-2} (tr_h P - tr_h Q) \\ &\quad + a_{m-2} \sum_{h=0}^{m-3} (tr_h P - tr_h Q) + \dots + a_2 \sum_{h=0}^1 (tr_h P - tr_h Q) + a_1 (tr_0 P - tr_0 Q) \end{aligned}$$

Hence the result.

**Fig. 1** Deterministic ordering of the vectors in the case of  $d = 3$  raters



**Corollary 1** For tables with the same margins, the Weighted Kappa preserves the stochastic agreement ordering:

$$\begin{aligned}
 & \text{if } p_{i+} = q_{i+} \text{ and } p_{+i} = q_{+i} \forall i = 1, \dots, m \text{ then} \\
 & P \geq_{Agr} Q \Rightarrow \kappa_w(P) \geq \kappa_w(Q). \tag{19}
 \end{aligned}$$

**5 The case of three raters ( $d = 3$ )**

For simplicity consider the special case of  $d = 3$  raters and discrete random vectors defined on  $\mathcal{S} = (1, 2, 3)^3$ . The scheme presented in Fig. 1 shows the agreement deterministic ordering of the vectors of ratings deriving from Definition 1 and Proposition 1. This scheme has the following interpretation: for example, the vector (2 1 1) is smaller in terms of agreement than the vectors (1 1 1), (2 2 2), (3 3 3); it is equivalent in terms of agreement to (3 2 2), (2 3 2), (2 2 3), (1 2 1), (1 1 2); it is greater in terms of agreement than the vectors (3 1 1), (3 2 1), (3 3 1) and also the vectors obtained from them permuting their components. It is a partial ordering because, for example, the vectors (2 1 1) and (2 2 1) are not comparable.

Applying Definition 4 to  $3 \times 3 \times 3$  tables of joint probabilities referring to the ratings of three observers, we can easily show the following.

**Proposition 4** Given two tables  $P = [p_{ijk}]$  and  $Q = [q_{ijk}]$  with  $i, j, k = 1, 2, 3$ , define the following sums of probabilities

$$\begin{aligned}
 p_1 &= p_{111} + p_{222} + p_{333} \\
 p_2 &= p_{322} + p_{232} + p_{223} + p_{211} + p_{121} + p_{112} \\
 p_3 &= p_{332} + p_{323} + p_{233} + p_{221} + p_{212} + p_{122} \\
 p_4 &= p_{311} + p_{131} + p_{113} \\
 p_5 &= p_{321} + p_{312} + p_{231} + p_{213} + p_{132} + p_{123} \\
 p_6 &= p_{331} + p_{313} + p_{133}
 \end{aligned}$$

then the agreement of  $P$  is higher—in the sense of  $\geq_{Agr}$ —than that of  $Q$  (we write  $P \geq_{Agr} Q$ ) if and only if the following inequalities hold:

$$\begin{aligned} p_1 &\geq q_1 \\ p_1 + p_2 &\geq q_1 + q_2 \\ p_1 + p_3 &\geq q_1 + q_3 \\ p_1 + p_2 + p_3 + p_4 &\geq q_1 + q_2 + q_3 + q_4 \\ p_1 + p_2 + p_3 + p_5 &\geq q_1 + q_2 + q_3 + q_5 \\ p_1 + p_2 + p_3 + p_6 &\geq q_1 + q_2 + q_3 + q_6 \end{aligned}$$

*Proof* The parameters  $p_1, \dots, p_6$  are the probabilities under  $P$  of the six equivalence classes shown in Fig. 1;  $p_1 + p_2$  is the probability of being more in agreement than the scores (3 2 2),  $p_1 + p_3$  is the probability of being more in agreement than the scores (3 3 2), etc., which proves the statement of this Proposition.

There are no additional conceptual problems when the number of raters and/or of scores is greater than three, except that the calculations are more complex.

## 6 Testing hypotheses concerning the agreement ordering $\geq_{Agr}$

### 6.1 Introductory remarks

In Sects. 2 and 3 we have defined a mathematical model to compare agreement among multivariate probability distributions. Now we want to show how to use observed data to test the hypothesis that the order relation holds, i.e. one group of observers shows more agreement among themselves than another group. To this purpose we apply results from the theory of statistical inference under order constraints (Barlow et al. 1972; Robertson et al. 1988; Silvapulle and Sen 2005), dealing with hypothesis testing problems involving linear constraints on the parameters of a parametric model.

We consider three kinds of hypotheses:  $H_0$  assumes equality of agreement among the ratings of two groups of observers;  $H_S$  assumes that the ratings are ordered under the agreement ordering;  $H_2$  assumes no restrictions among the ratings. Likelihood Ratio Tests are used for these types of testing problems: under the null hypothesis the test statistics have an asymptotic Chi-bar-squared distribution, which is a mixture of Chi-squared distributions with different degrees of freedom (Bartholomew 1959; Kudo 1963; Shapiro 1985; Shapiro 1988): see also Appendix B.

### 6.2 General case ( $d$ raters)

Let  $N$  be an  $m \times m \times \dots \times m$  ( $d$  times) frequency table referring to the observed ratings of  $t$  subjects by a group of  $d$  raters, on the basis of a 1 – to –  $m$  scale of measurement. Assume there are two such groups of observers rating  $t_1$  and  $t_2$  subjects respectively and thus two tables  $N_1$  and  $N_2$ . Then  $\mathbf{n}_1 = \text{vec}(N_1)$  and  $\mathbf{n}_2 = \text{vec}(N_2)$

are two  $m^d \times 1$  vectors of observations of two multinomial r. v.  $Y_1$  and  $Y_2$  respectively

$$Y_1 \sim Mult(t_1; \pi_1) \quad Y_2 \sim Mult(t_2; \pi_2)$$

where  $t_1$  and  $t_2$  are the sample size respectively of the two samples, while  $\pi_1 = vec(P)$  and  $\pi_2 = vec(Q)$  and  $P$  and  $Q$  are the tables of rating joint probabilities. Moreover let:

$$\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}.$$

We can express the ordering  $\geq_{Agr}$  of agreement for discrete variables defined in (9) by means of linear constraints of the kind

$$K \pi \geq \mathbf{0}$$

where  $K$  is a matrix whose rows are contrasts.

Assume, to start with, that we want to test the hypothesis that the probability distribution  $P$  associated to the first group of raters is equivalent in terms of agreement to the distribution  $Q$  of the second group against the alternative that the first distribution is greater in terms of agreement than the second one, i.e.

$$H_0 : P \cong_{Agr} Q \quad vs \quad H_S : P \geq_{Agr} Q \tag{20}$$

or equivalently

$$H_0 : \pi \in \Theta_0 \quad vs \quad H_S : \pi \in \Theta_S, \pi \notin \Theta_0$$

where

$$\Theta_0 = \{\pi \in \Theta : K \pi = \mathbf{0}\}, \quad \Theta_S = \{\pi \in \Theta : K \pi \geq \mathbf{0}\},$$

$$\Theta = \{\pi \in \mathbb{R}^{2m^d} : \pi \geq \mathbf{0} \text{ and } \pi_{i,1} + \dots + \pi_{i,m^d} = 1 \quad \forall i = 1, 2\}$$

and therefore  $\Theta_0 \subset \Theta_S \subset \Theta \subset \mathbb{R}^{2m^d}$ . In this case the Likelihood Ratio Test statistic is  $T_{0S}$

$$T_{0S} = -2 \left[ \sup_{\pi \in \Theta_0} L(\pi) - \sup_{\pi \in \Theta_S} L(\pi) \right] \tag{21}$$

where  $L(\pi)$  is the log-likelihood function. Under  $H_0$   $T_{0S}$  has the following asymptotic distribution (Silvapulle and Sen 2005):

$$\lim_{n \rightarrow \infty} Pr(T_{0S} \geq t \mid \pi \in \Theta_0) = \sum_{j=0}^k w_j \left( k, K I(\pi)^{-1} K^T, \mathbb{R}_+^k \right) \cdot Pr(\chi_j^2 \geq t) \tag{22}$$

where  $w_j(\cdot, \cdot, \cdot, \cdot)$  are positive weights which sum to 1,  $k$  is the number of rows of  $K$ ,  $\mathbb{R}_+^k = \{\mathbf{x} \in \mathbb{R}^k : \mathbf{x} \geq \mathbf{0}\}$  is the non-negative  $k$ -dimensional orthant,  $I(\pi)$  is the asymptotic Fisher Information matrix, and  $\chi_j^2$  is a Chi-squared variable with  $j$  degrees of freedom.

Expression (22) depends on the true value of the parameter vector. In order to compute the  $p$ -value of the asymptotic distribution we choose to find the least favourable value, i.e. that value of the parameter in  $\Theta_0$  which gives rise to the smallest rejection region (this procedure generates conservative tests). However, when this computation is too difficult and the sample size is large, we can apply a *bounds test* computing a lower and an upper limit for the asymptotic  $p$ -value, as described in [Silvapulle and Sen \(2005\)](#): if the upper bound is inferior to the significance level  $\alpha$  then we reject the null hypothesis; if the lower bound is greater than  $\alpha$  the null hypothesis is not rejected; in the other cases we cannot decide. Other approaches suggest performing a *local test* computing a point estimate of the asymptotic  $p$ -value by means of the M.L.E. of the parameters under the null hypothesis. [Dardanoni and Forcina \(1998\)](#) conducted a simulation study to assess the performance of the local test under  $H_0$ , concluding that even when the sample size is not too large the parameter estimation introduces no appreciable distortion, but, on the other hand, the rate of convergence to the asymptotic distribution is not as good as expected. In Appendix B we discuss the methods generally used to calculate the weights  $w_j$  with suitable approximation. After all this, the null distribution of the test statistics can be used to carry out the test.

The second testing problem considered here is the following: we want to test the hypothesis that the probability distribution associated to table  $N_1$  is stochastically greater in terms of agreement than the distribution related to table  $N_2$ , against the alternative hypothesis that this ordering does not hold. Therefore:

$$H_S : P \geq_{Agr} Q \quad vs \quad H_2 : \text{No restrictions} \tag{23}$$

or equivalently

$$H_S : \boldsymbol{\pi} \in \Theta_S \quad vs \quad H_2 : \boldsymbol{\pi} \in \Theta, \boldsymbol{\pi} \notin \Theta_S$$

We use the Likelihood Ratio Test statistic  $T_{S2}$

$$T_{S2} = -2 \left[ \sup_{\boldsymbol{\pi} \in \Theta_S} L(\boldsymbol{\pi}) - \sup_{\boldsymbol{\pi} \in \Theta_2} L(\boldsymbol{\pi}) \right] \tag{24}$$

having the following null asymptotic distribution:

$$\lim_{n \rightarrow \infty} Pr(T_{S2} \geq t \mid \boldsymbol{\pi} \in A) = \sum_{j=0}^s w_{s-j} \left( s, S I(\boldsymbol{\pi})^{-1} S^T, \mathbb{R}_+^s \right) \cdot Pr(\chi_j^2 \geq t) \tag{25}$$

where  $A = \{\boldsymbol{\pi} \in \Theta : K\boldsymbol{\pi} \geq \mathbf{0} \text{ and } S\boldsymbol{\pi} = \mathbf{0}, \text{ rank}(S) = s \geq 2\}$  and  $S$  is the matrix obtained from  $K$  considering just the rows corresponding to the active constraints, i.e. those constraints satisfied as equalities (see [Silvapulle and Sen 2005](#) for a detailed discussion). In this case too we need to compute the least favourable value of the parameters. When the computation of this supremum does not appear feasible, a bounds test can be performed, as in the previous case ([Silvapulle and Sen 2005](#)). Another approach proposed by [Wolak \(1991\)](#) consists in computing the weights of the asymptotic distribution using the unrestricted estimate of the parameter vector. However, he shows that if the least favourable value of the parameter vector does not satisfy



The problem becomes easier with the following reparameterization

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{bmatrix} = \begin{bmatrix} p_{111} + p_{222} + p_{333} \\ p_{322} + p_{232} + p_{223} + p_{211} + p_{121} + p_{112} \\ p_{332} + p_{323} + p_{233} + p_{221} + p_{212} + p_{122} \\ p_{311} + p_{131} + p_{113} \\ p_{321} + p_{312} + p_{231} + p_{213} + p_{132} + p_{123} \\ p_{331} + p_{313} + p_{133} \end{bmatrix}$$

and defining  $\mathbf{q}$  in the same way of  $\mathbf{p}$  we have that the parameter vector is

$$\tilde{\boldsymbol{\pi}} = \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}.$$

In this case the constraint matrix can be expressed as

$$\tilde{K} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & -1 & -1 & -1 & 0 & -1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & -1 & -1 & -1 & 0 & 0 & -1 \end{pmatrix}.$$

### 6.3 The case of $d = 2$ raters

The hypotheses testing problem for the case of just 2 raters (see Sect. 4) deserves special attention. Given two  $m \times m$  tables of joint probabilities  $P$  and  $Q$ , and operating the following reparameterization

$$\mathbf{p}^* = \begin{bmatrix} p_0^* \\ \vdots \\ p_h^* \\ \vdots \\ p_{m-1}^* \end{bmatrix} = \begin{bmatrix} tr_0 P \\ \vdots \\ tr_h P \\ \vdots \\ tr_{m-1} P \end{bmatrix} \qquad \mathbf{q}^* = \begin{bmatrix} q_0^* \\ \vdots \\ q_h^* \\ \vdots \\ q_{m-1}^* \end{bmatrix} = \begin{bmatrix} tr_0 Q \\ \vdots \\ tr_h Q \\ \vdots \\ tr_{m-1} Q \end{bmatrix}$$

where, we recall,

$$tr_h P = \sum_{i=1}^m \sum_{j=1}^m \sum_{|i-j|=h} p_{ij} = \begin{cases} \sum_{i=1}^m p_{ii} & \text{for } h = 0 \\ \sum_{i=1}^{m-h} p_{i,i+h} + \sum_{i=1}^{m-h} p_{i+h,i} & h = 1, \dots, m-1 \end{cases}$$

then the agreement ordering can be seen as the usual stochastic ordering between two vectors of multinomial probabilities:

$$P \geq_{Agr} Q \Leftrightarrow \sum_{h=0}^k p_h^* \geq \sum_{h=0}^k q_h^* \quad k = 0, 1, \dots, m - 2 \tag{26}$$

Therefore, using the following notation

$$\boldsymbol{\pi}^* = \begin{pmatrix} p^* \\ q^* \end{pmatrix}$$

we return to the testing problem described in the previous Section. The constraints defined by the agreement ordering can be expressed as

$$R\boldsymbol{\pi}^* \geq \mathbf{0} \quad \text{with} \quad R = [H, -H] \tag{27}$$

where  $H$  is a lower triangular matrix  $(m - 1) \times (m - 1)$  in which every element placed on the main diagonal and below it is equal to 1, and every element placed above it is equal to 0.

The peculiarity of this case is that we do know which is the least favourable value for the two testing problems we have described, and it is therefore possible to compute precise  $p$ -values: if we want to test the agreement ordering against the unrestricted alternative

$$H_S : R\boldsymbol{\pi}^* \geq \mathbf{0} \quad \text{vs} \quad H_2 : \text{No restrictions}$$

we have the following result (Silvapulle and Sen 2005):

$$\sup_{\boldsymbol{\pi}^* \in H_S} \lim_{n \rightarrow \infty} Pr (T_{S2} \geq t \mid \boldsymbol{\pi}^*) = \sum_{i=1}^{m-1} 2^{-(m-1)} \cdot \frac{(m - 1)!}{i!(m - 1 - i)!} \cdot Pr (\chi_i^2 \geq t) \tag{28}$$

whereas for testing the hypothesis of equivalence in terms of agreement against the alternative that the agreement ordering holds

$$H_0 : R\boldsymbol{\pi}^* = \mathbf{0} \quad \text{vs} \quad H_S : R\boldsymbol{\pi}^* \geq \mathbf{0}$$

we get

$$\sup_{\boldsymbol{\pi}^* \in H_0} \lim_{n \rightarrow \infty} Pr (T_{0S} \geq t \mid \boldsymbol{\pi}^*) = \frac{1}{2} \left\{ Pr (\chi_{m-2}^2 \geq t) + Pr (\chi_{m-1}^2 \geq t) \right\}. \tag{29}$$

We end this session with a general remark. In the applications sparse contingency tables, i.e. tables with a lot of cells having zero counts, will be frequently encountered. Empty cells and sparse tables can cause problems with existence of M.L.E. estimates and with the computational algorithms. One of the solutions generally adopted is to add a small constant to cell counts before performing the estimation

process. This is what has been done in the data analysis showed in the next paragraph. One important point, however, is the following: the parameters we actually need to estimate in order to perform the tests are suitable sums of the entries of  $P$  and  $Q$ , thus the number of essential parameters is much smaller than  $2(m^d - 1)$ . For instance in the above example of 3 observers and 3 categories only 10 independent probabilities  $p_1, p_2, \dots, p_5$  and  $q_1, q_2, \dots, q_5$  are needed instead of 52, namely  $p_{111}, p_{112}, \dots, p_{332}$  and  $q_{111}, q_{112}, \dots, q_{332}$ . When  $d = 2$  like in our case-study, just  $2(m - 1)$  unknown parameters are needed, instead of  $2(m^2 - 1)$ . For this reason, scarcity in contingency tables is less likely to affect our problem.

## 7 An example

The methodology illustrated in this article has been applied to data collected by the Hospital (Azienda Ospedaliera) of Perugia (Italy) on the evaluation of the quality of a certain number of clinical guidelines, referring to four Units of the hospital. This evaluation was carried out by two experienced doctors.

Clinical guidelines are a type of advice to doctors for their clinical practice. Every guideline of this study is formed by 23 items, each of them referring to one specific aspect of medical practice. Every item has been classified on a scale formed by 4 ordered categories, ranging from “Bottom quality” (the doctor was in complete disagreement with the guideline) to “Top quality” (the doctor was entirely in agreement). Therefore, each item of each guideline has been considered as a statistical unit to classify. The scale used to rate the items is clearly qualitative and ordinal. To apply the methodology described in this paper we have discretized the categories of the scale, assigning them the values 1, 2, 3 and 4; the comparison is among the agreement of the two doctors on their ratings of the guidelines relative to different hospital Units.

Table 1 shows the ratings of the two doctors on the items of the guidelines referring to four Units, denoted as A, B, C, D, of Perugia Hospital. At the upper left hand of each table the value of the Weighted Kappa index is shown, computed with Cicchetti’s weights (16). Thus the ratings of the two doctors can be represented as  $4 \times 4$  tables, one for each of the 4 Units (see Table 1).

Let  $P$  and  $Q$  be the joint probability distributions of the ratings of the two doctors, and tables  $N$  and  $M$  be the observed values. We want to test the following hypotheses:

- (1)  $H_0$  versus  $H_S - H_0$
- (2)  $H_0$  versus  $H_{\bar{S}} - H_0$
- (3)  $H_S$  versus  $H_2 - H_S$
- (4)  $H_{\bar{S}}$  versus  $H_2 - H_{\bar{S}}$

where

$$H_0 : P \cong_{Agr} Q$$

$$H_S : P \geq_{Agr} Q$$

$$H_{\bar{S}} : Q \geq_{Agr} P$$

$H_2 : P$  and  $Q$  cannot be compared in terms of agreement and the respective  $p$ -values have been labelled:  $p_{0S}$ ,  $p_{0\bar{S}}$ ,  $p_{S2}$  and  $p_{\bar{S}2}$ . A routine in Matlab created by Prof. Antonio Forcina (Department of Statistics of Perugia University) has been adopted to compute the constrained M.L.E. and the  $p$ -values of the tests.

**Table 1** Evaluation of the quality of the guidelines referring to Units A, B, C and D of Perugia Hospital

	First doctor		Second doctor		Total
	1	2	3	4	
UNIT 'A', $\kappa_w = 0,409$					
1	83	9	10	2	104
2	12	2	28	3	45
3	0	0	8	0	8
4	0	0	4	0	4
Total	95	11	50	5	161
UNIT 'B', $\kappa_w = 0,426$					
1	76	7	23	0	106
2	0	4	18	0	22
3	8	0	41	8	57
4	6	7	8	1	22
Total	90	18	90	9	207
UNIT 'C', $\kappa_w = 0,535$					
1	71	4	8	0	83
2	3	4	20	0	27
3	4	3	11	0	18
4	0	1	7	2	10
Total	78	12	46	2	138
UNIT 'D', $\kappa_w = 0,563$					
1	44	12	4	0	60
2	2	12	6	0	20
3	1	0	5	0	6
4	0	0	1	5	6
Total	47	24	20	1	92

After calculating the  $p$ -values of the tests described in Table 2, we reject the hypothesis of equal agreement of the doctors relative to the guidelines of tables A and B in favour of the hypothesis that more agreement is shown in A than in B ( $p_{0S} = 0.0015$ ); in addition, testing whether table A shows higher agreement than B against whether this is not true gives  $p_{S2} = 0,7979$  so the hypothesis is accepted, while the reverse relationship (table B shows higher agreement than A) is rejected ( $p_{\bar{S}2} = 0,0053$ ). Thus we can state that agreement is greater in table A than in B.

Comparing A against C, the hypothesis of equivalent agreement is not rejected for both alternatives ( $p_{0S} = 0,8089$  and  $p_{0\bar{S}} = 0,2011$ ), although the latter with less evidence, and both tests regarding the agreement ordering are non significant ( $p_{S2} = 0,1397$  and  $p_{\bar{S}0} = 0,8655$ ). Thus we can state that A is equivalent to C in terms of agreement. The same conclusions are drawn comparing A against D and C against D.

Comparing table B against table C, the hypothesis of equality in terms of agreement is rejected in favor of the hypothesis that C shows more agreement than B ( $p_{0\bar{S}} = 0.0018$ ); moreover, this hypothesis is accepted against the unrestricted alternative

**Table 2**  $p$ -values referring to the tests used: comparisons among units A, B, C and D

Comparisons	$P_{0S}$	$P_{S2}$	$P_{0\tilde{S}}$	$P_{\tilde{S}2}$
A and B	0,0015	0,7979	0,1572	0,0053
A and C	0,8089	0,1397	0,2011	0,8655
A and D	1,0000	0,1048	0,2061	1,0000
B and C	0,7873	0,0032	0,0018	1,0000
B and D	0,8882	0,0005	0,0009	1,0000
C and D	1,0000	0,3847	0,6287	0,8738

**Table 3** Decisions about the comparisons

	$\cong_{Agr}$	$\geq_{Agr}$	$\leq_{Agr}$	No order
A and B	No	Yes	No	No
A and C	Yes	–	–	–
A and D	Yes	–	–	–
B and C	No	No	Yes	No
B and D	No	No	Yes	No
C and D	Yes	–	–	–

( $p_{\tilde{S}2} = 1.0000$ ), while the reverse relationship (table B shows higher agreement than C) is rejected ( $p_{S2} = 0,0032$ ). Thus, we can conclude that C is greater than B in terms of agreement. Comparing table B against table D gives a similar result.

We summarize the conclusions in Table 3.

Thus the data indicate that  $B \leq_{Agr} A \cong_{Agr} C \cong_{Agr} D$ .

Comparing the values of weighted Kappa, table A shows less agreement than table B, since the Kappa values are respectively 0,409 and 0,426. Therefore, in this case the order relation expressed by the Kappa indices is not consistent with the results of the agreement test of Sect. 6.

Considering tables A and C, the tests performed lead to their equivalence in terms of agreement, while the Kappa index shows that C expresses more agreement than A. The same conclusions can be drawn comparing A against D and comparing C against D.

For the remaining two comparisons, i.e. B versus C and B versus D, the Kappa indices are consistent with the analyses performed by the test statistics.

### 8 Conclusions

In this paper the important issue of measuring how much two or more observers agree in their judgements is approached from a novel viewpoint, through the ordering of the random vectors of scores. We hope to have shown that our approach provides a useful insight into this problem and helps clarify the behaviour of well known agreement indices. It also suggests a path for defining some new measures of agreement, consistent with the ordering.

Although the paper concentrates on ratings on a discrete scale and focuses on just two observers, the methods employed are general. A similar approach has been used by the authors when the ratings are expressed on a categorical scale (Giovagnoli et al. 2007).

Apart from theoretical developments a crucial aspect concerns the possibility of testing the order relation on the presence of real data. In this paper we have shown how this can be achieved applying the principles of constrained statistical inference.

**Acknowledgments** This research was started while the second author was an Erasmus PG student in the Department of Statistics of the LSE, UK, under the joint supervision of the other two authors, and is part of his PhD thesis (Marzialetti 2006). The third author wishes to thank the Institute of Advanced Studies of Bologna University where he was senior fellow in October and November 2006. Thanks are also due to Prof Antonio Forcina, of Perugia University, and his collaborators for scientific and financial support of this research within the “PRIN 2002” Project: *Statistical methods for stochastic orderings with sociological, medical and environmental applications*.

## Appendix

### A Orderings in a set

Given a set  $S$ , a *pre-ordering* is a binary relation  $<$  defined on the elements of  $S$  which satisfies the following properties:

(1) Reflexive property:

$$x < x \quad \forall x \in S$$

(2) Transitive property:

$$\text{if } x < y \text{ and } y < z \Rightarrow x < z, \quad \forall x, y, z \in S$$

A *partial ordering* is a binary relation defined on the elements of  $S$  which satisfies the above properties and also the anti-symmetric property:

(3) Anti-symmetric property:

$$\text{if } x < y \text{ and } y < x \Rightarrow x = y, \quad \forall x, y \in S$$

An order relation is said to be *total* if the following property also holds:

(4)  $x < y$  or  $y < x \quad \forall x, y \in S$

### B Chi-bar-squared random variable

Let  $C$  be a closed convex cone in  $\mathbb{R}^k$ , let  $V$  be a  $k \times k$  symmetric and positive definite matrix, and let  $y_{V,C}$  be the projection of a vector  $y \in \mathbb{R}^k$  onto the cone  $C$  in the metric  $V^{-1}$ , i.e.  $y_{V,C}$  is the solution to the problem

$$\operatorname{argmin}_{y^* \in C} (y - y^*)^T V^{-1} (y - y^*).$$

Applying the standard properties of projections onto convex cones and their duals, it can be proven that

$$\|y\|^2 = \|y_{V,C}\|^2 + \|y_{V,C^o}\|^2$$

where  $\| \cdot \|$  is the norm defined by the metric  $V^{-1}$ ,  $C^o$  is the dual cone of  $C$  in the metric  $V^{-1}$ , and is defined as

$$C^o = \left\{ \mathbf{x} : \mathbf{y}^T V^{-1} \mathbf{x} \leq 0, \forall \mathbf{x} \in C \right\}.$$

Under the assumption that

$$\mathbf{Y} \sim N(\mathbf{0}, V)$$

the random variable

$$\chi^2(V, C) = \mathbf{Y}_{V,C}^T V^{-1} \mathbf{Y}_{V,C}$$

is called Chi-bar-squared, and has the following distribution

$$Pr \left( \chi^2(V, C) \geq t \right) = \sum_{i=0}^k w_i(k, V, C) \cdot Pr \left( \chi_i^2 \geq t \right)$$

where the  $w_i(c, V, C)$  are non-negative values which sum to one and depend on the matrix  $V$  and the cone  $C$  and  $\chi_i^2$  are Chi-squared variables with  $i$  degrees of freedom. This random variable has been intensively studied in the literature (Bartholomew 1959; Kudo 1963; Shapiro 1985; Shapiro 1988; Gourieroux et al. 1982; Dardanoni and Forcina 1998; Colombi and Forcina 2000), and used in contexts of hypothesis testing under linear inequality constraints.

Although the weights of Chi-bar-squared distributions can be computed exactly by integrating a normal density on a proper convex cone, in practice this operation is extremely laborious even when the space dimension is less than 4, but it is possible to obtain accurate estimates  $\hat{\mathbf{w}}$  of the weights using Monte Carlo simulation techniques, projecting onto the positive orthant a certain number (say  $t$ ) of pseudo-random vectors generated from  $N(\mathbf{0}, V)$ . Let  $X$  have a Chi-bar-squared distribution with weights  $\mathbf{w}$  and let  $Pr(X \geq r) = \alpha$  and  $Pr(\chi_i^2 \geq r) = r_i$ . Then by the central limit theorem (Dardanoni and Forcina 1998):

$$\hat{\mathbf{w}} \sim N \left( \mathbf{w}, \frac{1}{t} \left( \text{diag}(\mathbf{w}) - \mathbf{w} \mathbf{w}^T \right) \right).$$

Thus, starting from a preliminary estimate of  $\mathbf{w}$ , we can choose  $t$  in order to achieve the required level of precision by imposing that  $Pr(|r^T \hat{\mathbf{w}} - \alpha|/\alpha \leq \lambda)$  be close to 1, where  $\lambda$  is the error allowed. It follows that the estimated  $p$ -value referring to a certain observed value  $x$  of the Chi-bar-squared variable with estimated weights has a normal distribution with mean equal to the real  $p$ -value and variance equal to  $(\mathbf{1} - \mathbf{f})^T V (\mathbf{1} - \mathbf{f})$ , where the  $h^{th}$  element of the vector  $\mathbf{f}$  is

$$Pr \left( \chi_h^2 \leq x \right).$$

For the special case when  $V$  is the identity matrix and  $C$  is the non-negative orthant, the weights can be computed as (Gourieroux et al. 1982)

$$w_i \left( k, I_k, \mathbb{R}_+^k \right) = \binom{k}{i} \left( \frac{1}{2} \right)^k$$

while if the cone is the non-negative orthant and the matrix  $V$  is general we can adopt the solution proposed by Kudo (1963):

$$w_i \left( k, V, \mathbb{R}_+^k \right) = \sum_{\dim(M)=i} p \left\{ (V_{M'})^{-1} \right\} \cdot p \left\{ V_{M;M'} \right\}$$

where the summation is over all subsets  $M$  of  $\{1, \dots, k\}$  with  $i$  elements,  $M'$  is the complement of  $M$ ,  $V_M$  is the variance-covariance matrix of variables  $Y_i$  ( $i \in M$ ),  $V_{M;M'}$  is the variance-covariance matrix under the condition  $Y_j = 0$  ( $j \in M'$ ), and  $p \{A\}$  is the probability that  $\mathbf{Z} \geq \mathbf{0}$  for a normal variable  $\mathbf{Z} \sim N(\mathbf{0}, A)$ . However, when  $k > 4$  the computation of these probabilities might be very difficult.

## References

- Banerjee M, Capozzoli M, McSweeney L, Sinha D (1999) Beyond kappa: a review of interrater agreement measures. *Can J Stat* 27(1):3–23
- Barlow RE, Bartholomew DJ, Bremner DJ, Brunk HD (1972) *Statistical inference under order restrictions*. Wiley, New York
- Bartholomew DJ (1959) A test of homogeneity for ordered alternatives. *Biometrika* 46:36–48
- Bartolucci F, Forcina A (2002) Extended rc association models allowing for order restrictions and marginal modelling. *J Am Stat Assoc* 97:1192–1199
- Bishop YMM, Fienberg SE, Holland PW (1975) *Discrete multivariate analyses: theory and practice*. MIT Press, Cambridge
- Cicchetti DV (1972) A new measure of agreement between rank ordered variables. *Proc Am Psychol Assoc* 7:17–18
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46
- Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 70:213–220
- Colombi R, Forcina A (2000) Modellizzazione di dati discreti con vincoli di uguaglianza e disuguaglianza. *Statistica* 2:195–214
- Dardanoni V, Forcina A (1998) A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J Am Stat Assoc* 93:1112–1123
- Giovagnoli A (2002) Stochastic orderings and their use in statistics: the case of association between two variables. In: *Proceedings of the XLI scientific meeting of the Italian statistical society*, Milan, 5–7 June 2002, pp 95–104
- Giovagnoli A, Wynn HP (1985) G-majorization with applications to matrix orderings. *Linear Algebra Appl* 67:111–135
- Giovagnoli A, Marzioletti J, Wynn HP (2007) Bivariate stochastic orderings for unordered categorical variables. In: Pronzato L, Zhigljavsky A (eds) *W-Optimality in Design and Statistics*. Springer, Heidelberg, pp 81–96
- Gourieroux C, Holly A, Monfort A (1982) Likelihood ratio test, Wald test, and Kuhn–Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica* 50:63–80
- Kudo NM (1963) A multivariate analogue of the one-sided test. *Biometrika* 50:403–418
- Marshall AW, Olkin I (1979) *Inequalities: theory of majorization and its applications*. Academic, New York
- Marzioletti J (2006) *Lo Studio dell'Agreement mediante gli Ordinamenti*. Ph.D. thesis, Department of Statistical Sciences, University of Bologna

- Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. Wiley, New York
- Schouten HJA (1982) Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Stat Neerl* 36:45–61
- Shaked M, Shanthikumar JG (2007) Stochastic orders. Springer, Berlin
- Shapiro A (1985) Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* 72:133–144
- Shapiro A (1988) Towards a unified theory of inequality constrained testing in multivariate analysis. *Int Stat Rev* 56:49–62
- Silvapulle MJ, Sen PK (2005) Constrained statistical inference: inequality, order and shape restrictions. Wiley-Interscience, Hoboken
- Wolak FA (1991) The local and global nature of hypothesis tests involving inequality constraints in nonlinear models. *Econometrics* 59:981–995