

An Approach to Inter-rater Agreement through Stochastic Orderings: the Discrete Case

Un Approccio All'Accordo Tra Osservatori Mediante Gli Ordinamenti Stocastici: Il Caso Discreto

Alessandra Giovagnoli, Johnny Marzioletti
University of Bologna, Bologna, Italy

Henry Wynn
London School of Economics and Political Science, London, UK

Riassunto: Si intende studiare l'accordo tra d osservatori che classificano un campione di unità statistiche su una scala discreta. Il punto di partenza è una relazione di maggiore-minore "agreement" definita assiomaticamente sui vettori che esprimono i punteggi degli osservatori relativamente a una stessa unità statistica. Successivamente, si definisce un modo per confrontare due diverse distribuzioni di probabilità (o di frequenza) d -dimensionali che descrivono i giudizi di due gruppi di d osservatori su uno stesso campione. Alcune misure di agreement verranno discusse nel caso $d = 2$.

Keywords: Raters' agreement, Stochastic orderings, Discrete random variables.

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 1968). 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).

The approach of this paper is the following: given the sample space (the subjects) to each observer there corresponds the frequency distribution of his/her ratings. Thus an observer gives rise to a random variable with values in the set S of all possible classes or ratings and the agreement among a group of observers is a type of association among the jointly distributed random variables representing the 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.

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.

2. Agreement among observers rating just one subject

Assume the vector $\mathbf{x} = (x_1, x_2, \dots, x_d)$ in a set S expresses the ratings of d observers about

the same statistical unit. If the scores are on a continuous scale, let $\mathcal{S} = \mathbf{R}^d$, 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 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.

We now show that Axioms **A1-A4** uniquely define the agreement ordering $\geq_A$. It is well-known (Marshall and Olkin 1979) that axioms **A1, A2, A3** characterize the opposite of the so-called *majorization* ordering $\cup$:

Definition 1 Given two vectors $\mathbf{x} = (x_1, x_2, \dots, x_d)$ and $\mathbf{y} = (y_1, y_2, \dots, y_d)$, $\mathbf{x}$ is majorized by $\mathbf{y}$ ($\mathbf{x} \prec \mathbf{y}$) if:

$$\begin{cases} \sum_{i=1}^k x_{[i]} \leq \sum_{i=1}^k y_{[i]} & k = 1, \dots, d-1 \\ \sum_{i=1}^d x_{[i]} = \sum_{i=1}^d y_{[i]} \end{cases} \quad \text{where } x_{[1]} \geq x_{[2]} \geq \dots \geq x_{[d]}.$$

An equivalent condition is that $\mathbf{x}$ is a convex combination of the vectors obtained by permuting the coordinates of $\mathbf{y}$. To account for axiom **A4**, instead of just permutations we consider the group $\mathcal{G}$ of permutations AND simultaneous shifts of all the vector coordinates. This is called *G-majorization* (Giovagnoli and Wynn 1985) relative to group $\mathcal{G}$. It can be shown that if we set

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

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

Proposition 1 For an order relation $\geq_A$ satisfying **A1-A4**,

$$\mathbf{x} \geq_A \mathbf{y} \Leftrightarrow \mathbf{x}^* \prec \mathbf{y}^* \quad \square.$$

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

A large class of indicators of the agreement $\geq_A$ is given by all concave, permutation- and shift-invariant functions Φ . In particular $\mathbf{x}^* \prec \mathbf{y}^* \Leftrightarrow \sum \varphi(x_i^*) \leq \sum \varphi(y_i^*)$ for all continuous convex functions φ , and examples of measures of agreement are any decreasing function γ of the variance $(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_d - \bar{x})^2$.

3. Agreement ordering of multidimensional r.v's.

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

Definition 2: Given the pre-ordering $\geq_A$ on the set $\mathcal{S}$ and the d -dimensional random vectors X and Y with values in $\mathcal{S}$, let

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

This definition is in analogy with the definition of upper-orthant stochastic ordering for d -dimensional random variables (see for instance Shaked and Shantikumar 1994), where $\geq_A$ replaces the entry-wise ordering of vectors in the d -dimensional real space. Observe that Definition 2 implies

$$X \geq_{\text{Agr}} Y \Leftrightarrow \Pr \{X^* \prec z^*\} \geq \Pr \{Y^* \prec z^*\} \quad \forall z^* = z - \bar{z}1 \text{ such that } z \in S$$

We now show how to check that the condition holds in the special (but important) case of only two raters.

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

For simplicity consider the case of just two raters: then the ordering $\geq_A$ is total (i.e. every pair of vectors is comparable). This implies

Proposition 2 When $d = 2$

$$X \geq_{\text{Agr}} Y \Leftrightarrow E\{\Phi(X)\} \geq E\{\Phi(Y)\} \quad \forall \geq_A \text{-increasing } \Phi \text{ on } S \quad \square.$$

Applying Definition 2 to $m \times m$ tables of joint probabilities referring to the ratings of two observers on a scale 1 to m , we can prove:

Proposition 3 Given two tables $\mathbf{P} = [p_{ij}]$ and $\mathbf{Q} = [q_{ij}]$ $i, j = 1, \dots, m$, the agreement of $\mathbf{P}$ is higher - in the sense of $\geq_{\text{Agr}}$ - than that of $\mathbf{Q}$ (and we write $\mathbf{P} \geq_{\text{Agr}} \mathbf{Q}$) if and only if

$$\sum_{h=0}^k \text{tr}_h \mathbf{P} \geq \sum_{h=0}^k \text{tr}_h \mathbf{Q} \quad k = 1, \dots, m$$

where

$$\text{tr}_h \mathbf{P} = \sum_{|i-j|=h} p_{ij} = \sum_{i=1}^{m-h} p_{i, i+h} + \sum_{i=1}^{m-h} p_{i+h, i} \quad h = 1, \dots, m \quad \square.$$

We now want to check how this definition fits in with existing measures of agreement. The *weighted total proportion of agreement* is the index

$$\text{TPA}_w = \sum_{i=1}^m \sum_{j=1}^m w_{ij} p_{ij}$$

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 $w_{ii} = 1$. Typically $w_{ij} = 0$ for $i \neq j$; otherwise the weights are usually decreasing with respect to the distance from the main diagonal $|i-j|$. 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):

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

The ordinary (unweighted) kappa is when $w_{ij} = 0$ for $i \neq j$.

Proposition 4: The index TPA_w is order-preserving with respect to $\geq_{\text{Agr}}$

$$\mathbf{P} \geq_{\text{Agr}} \mathbf{Q} \Rightarrow \text{TPA}_w(\mathbf{P}) \geq \text{TPA}_w(\mathbf{Q}) \quad \square.$$

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

$$\mathbf{P} \geq_{\text{Agr}} \mathbf{Q} \Rightarrow \kappa_w(\mathbf{P}) \geq \kappa_w(\mathbf{Q}) \quad \square.$$

The case of $d=3$ raters is similar. However $\geq_A$ is now a partial ordering and the calculations are more complex.

6. Conclusions

Apart from definitions and properties of the stochastic agreement ordering $\geq_{\text{Agr}}$, we want to be able to test whether or not $\geq_{\text{Agr}}$ holds when we observe real data, namely frequency tables. More precisely, we may test the null hypothesis of independence against the existence of the ordering or the ordering against the unrestricted hypothesis. These problems belong to the field of statistical inference under order restrictions. The test is the likelihood ratio one and under the null the test statistics have a chi-bar-squared distribution. The theoretical developments related to this are described in Marzioletti (2006).

Acknowledgments

This research was carried out while the second author was a Socrates PG student in the Department of Statistics of L.S.E., London. Thanks are due to Prof A. 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*.

References

- Banerjee M., Capozzoli M., McSweeney L., Sinha D. (1999) Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, 27, 1, 3-23.
- Bishop Y.M.M., Fienberg S.E., Holland P.W. (1975) *Discrete Multivariate Analyses: Theory and practice*, MIT Press, Cambridge, MA.
- Cohen J. (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Giovagnoli A., (2002) Stochastic orderings and their use in statistics: the case of association between two variables. *Proceedings of the XLI Scientific Meeting of the Italian Statistical Society*, 5-7 June 2002, 95-104.
- Giovagnoli A., H.P. Wynn (1985) "G-majorization with applications to matrix orderings" *Linear Algebra Appl.*, 67, 111-135.
- Marshall A.W., Olkin I. (1979) *Inequalities: Theory of Majorization and Its Applications*. Academic Press.
- Marzioletti J. (2006) *Lo studio dell'agreement mediante gli ordinamenti*. PhD Thesis, Department of Statistics, University of Bologna, Italy, March 2006.
- Shaked M., Shanthikumar J. G. (1994) *Stochastic orders and their applications*. San Diego: Academic Press.
- Schouten H.J.A. (1982) Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica*, 36, 45-61.