International Journal of Approximate Reasoning 73 (2016) 56–75

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Computation of general correlation coefficients for interval data

Karol R. Opara *, Olgierd Hryniewicz

Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

ARTICLE INFO

Article history:
Received 27 November 2015
Received in revised form 18 February 2016
Accepted 29 February 2016
Available online 7 March 2016

Keywords:
Measures of dependence
Interval data
Kendall's tau
Spearman's rho
Partial orders

ABSTRACT

This paper provides a comprehensive analysis of computational problems concerning calculation of general correlation coefficients for interval data. Exact algorithms solving this task have unacceptable computational complexity for larger samples, therefore we concentrate on computational problems arising in approximate algorithms. General correlation coefficients for interval data are also given by intervals. We derive bounds on their lower and upper endpoints. Moreover, we propose a set of heuristic solutions and optimization methods for approximate computation. Extensive simulation experiments show that the heuristics yield very good solutions for strong dependencies. In other cases, global optimization using evolutionary algorithm performs best. A real data example of autocorrelation of cloud cover data confirms the applicability of the approach.

1. Introduction

The analysis of statistical dependence is one of the most important parts of statistics. First statistical procedures for the analysis of dependent data were introduced more than one hundred years ago. Since that time multitude of particular methods have been devised. The introduction of data mining techniques significantly extended the area of applications where the analysis of dependencies in data plays a crucial role. Despite the fact that the real statistical data are often imprecise, as the data are gathered from intrinsically imprecise measurements, the interest in the statistical analysis of imprecise data is relatively new. First statistical methods applicable in the analysis of imprecise data were proposed in papers published in the 1980s. The most important publication from that time which established a new statistical methodology for coping with imprecise (fuzzy) data is the book by Kruse and Meyer [25]. Since that time many books and papers on fuzzy statistics have been published. Interesting overview of these methods can be found e.g. in the paper by Gil and Hryniewicz [12] or in the book by Viertl [35]. The general methodology for the statistical analysis of imprecise (fuzzy) data is still under development, and some new techniques, see e.g. Couso and Sanchez [3], have been proposed recently.

First publications devoted to the problem of the interval correlation coefficient – framed as testing statistical hypotheses for dependent data – were published in the early 2000s. Testing statistical hypotheses for categorical data displayed in the form of contingency tables was considered in [15,16]. The statistical analysis of dependence using the well-known Kendall's τ statistics for imprecise data was considered for the first time in the paper by Hébert et al. [14]. The most important paper related to the problem of statistical testing of independence with fuzzy data is written by Denœux et al. [4], where this problem has been presented in a more general framework of using rank tests for fuzzy data. In all these papers

* Corresponding author.
E-mail address: karol.opara@ibspan.waw.pl (K.R. Opara).

http://dx.doi.org/10.1016/j.ijar.2016.02.007
0888-613X/

the authors have noticed important difficulties with the calculation of the values of fuzzy statistics. Some variants of the problem are known to be NP-hard. Hébert et al. [14] proposed an algorithm for the calculation of the exact interval value of Kendall's τ which, unfortunately, was computationally effective only for very small samples (less than 10 elements). Denœux et al. [4] considered an algorithm for the calculation of the approximate interval value of Kendall's τ which was effective also for relatively small samples (max. 30 elements). These findings prompted Hryniewicz and Szediw [22] to look for the approximate interval value of Kendall's τ that could be used as the starting point in the procedure for the calculation of more precise value of this statistic. In that paper the approximate interval value of Kendall's τ was calculated using a heuristic algorithm for autocorrelated imprecise data coming from statistical quality control. The results appeared promising, especially for highly correlated data. In the paper by Hryniewicz and Opara [19] an extended set of heuristic algorithms has been proposed to calculate the approximate interval value of Kendall's τ for usual bivariate interval data. Preliminary results presented in this paper have shown that further investigations are needed. One result of these investigations is described in the paper by Hryniewicz and Opara [20], who used a general purpose genetic optimization algorithm for finding better approximations.

In this paper we present a comprehensive analysis of computing general correlation coefficients for interval data. They are also given by intervals. We derive inequalities bounding their lower endpoint from below and upper endpoint from above, which we will further refer to as "outer bounds". We also report on several experiments, which compare the efficiency of the algorithms for constructive calculation of correlation coefficients, which we call "inner bounds".

The analyzed problems of computing interval statistics can be reformulated as finding linear extensions to partial orders or as bilinear quadratic programming problems. The proposed results are hence applicable in such areas as testing hypotheses for fuzzy data, comparing partial rankings in recommender systems or finding equilibria in bimatrix games.

Investigation of these issues requires a review of results obtained for crisp correlation coefficients, which is given in section 2. Next, we show different formulations of the problem of computing interval correlation coefficients and discuss their computational complexity. Section 4 provides derivation of outer bounds for Kendall's τ and Spearman's ρ. Inner bounds can be obtained using specialized heuristics or performing constrained optimization, which is discussed in section 5. Finally, we evaluate the proposed approach on an extensive simulation study and show an example of its application for analyzing real data.

2. Correlation coefficients

2.1. General correlation coefficient

Suppose we have a set of n objects characterized by two properties x and y. To any pair of individuals, say i-th and j-th, one can assign x-score $a_{ij} = -a_{ji}$ and y-score $b_{ij} = -b_{ji}$. Kendall [23] describes a general correlation coefficient Γ as

$$\Gamma = \frac{\sum_{i,j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i,j=1}^{n} a_{ij}^2 \sum_{i,j=1}^{n} b_{ij}^2}} \tag{1}$$

Scores a_{ij} and b_{ij} are regarded zero if $i = j$. In formula (1) each pair is counted twice – operationally one can exploit the symmetry of scores to half the necessary amount of computation.

The three most common correlation coefficients, Pearson's r, Spearman's ρ and Kendall's τ are special cases of (1). Pearson's product–moment correlation is obtained by basing the scores on the actual variate values

$$a_{ij} = x_j - x_i \tag{2}$$

$$b_{ij} = y_j - y_i \tag{3}$$

Spearman's ρ can be stated by a similar formula, in which variate values are replaced with ranks

$$a_{ij} = p_j - p_i \tag{4}$$

$$b_{ij} = q_j - q_i \tag{5}$$

where p_i denotes the rank of the i-th object according to the x-quality. Similarly q_i is a rank according to the y-quality, $p_i, q_i \in \{1, \ldots, n\}$. Kendall's τ is obtained from the general correlation coefficient by allotting scores ± 1 depending on the ranks of observations

$$a_{ij} = \begin{cases} +1 & \text{if } p_i < p_j \\ -1 & \text{if } p_i > p_j \end{cases} \tag{6}$$

$$b_{ij} = \begin{cases} +1 & \text{if } q_i < q_j \\ -1 & \text{if } q_i > q_j \end{cases} \tag{7}$$