Mechanical Systems and Signal Processing 37 (2013) 43–53

Contents lists available at SciVerse ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

Estimating correlation under interval uncertainty

Ali Jalal-Kamali, Vladik Kreinovich *

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

ARTICLE INFO

Article history:
Received 19 November 2011
Received in revised form
27 November 2012
Accepted 2 December 2012
Available online 20 January 2013

Keywords:
Imprecise probabilities
Correlation
Interval uncertainity

ABSTRACT

In many engineering situations, we are interested in finding the correlation ρ between different quantities x and y based on the values x_i and y_i of these quantities measured in different situations i. Measurements are never absolutely accurate; it is therefore necessary to take this inaccuracy into account when estimating the correlation ρ. Sometimes, we know the probabilities of different values of measurement errors, but in many cases, we only know the upper bounds Δ_{xi} and Δ_{yi} on the corresponding measurement errors. In such situations, after we get the measurement results $\tilde{x}_i$ and $\tilde{y}_i$, the only information that we have about the actual (unknown) values x_i and y_i is that they belong to the corresponding intervals $[\tilde{x}_i-\Delta_{xi},\tilde{x}_i+\Delta_{xi}]$ and $[\tilde{y}_i-\Delta_{yi},\tilde{y}_i+\Delta_{yi}]$. Different values from these intervals lead, in general, to different values of the correlation ρ. It is therefore desirable to find the range $[\underline{\rho},\overline{\rho}]$ of possible values of the correlation when x_i and y_i take values from the corresponding intervals. In general, the problem of computing this range is NP-hard. In this paper, we provide a feasible (=polynomial-time) algorithm for computing at least one of the endpoints of this interval: for computing $\overline{\rho}$ when $\overline{\rho}>0$ and for computing $\underline{\rho}$ when $\underline{\rho}<0$.

1. Introduction

Need for correlation. In engineering, we design systems for real-world applications. To make sure that the system functions correctly, we need to take into account all possible situations in which these systems will function. Each such situation can be characterized by the values of different quantities. To describe which combinations of these values are more probable and which are less probable, it is necessary to know which quantities are independent and which are correlated—positively or negatively.

To estimate the correlation between the quantities x and y, we repeatedly measure the values x_i and y_i of both quantities in different situations i. The correlation ρ is then estimated as the ratio

$$\rho=\frac{C}{\sigma_x\cdot\sigma_y}$$

of the covariance C to the product of standard deviations $\sigma_x=\sqrt{V_x}$ and $\sigma_y=\sqrt{V_y}$. Covariance and standard deviations, in their turn, are defined as follows:

$$C=\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-E_x)\cdot(y_i-E_y)=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i\cdot y_i-E_x\cdot E_y,$$

$$V_x=\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-E_x)^2,\quad V_y=\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i-E_y)^2,$$

* Corresponding author. Tel.: +1 915 747 6951; fax: +1 915 747 5030.
E-mail addresses: ajalalkamali@miners.utep.edu (A. Jalal-Kamali), vladik@utep.edu (V. Kreinovich).

0888-3270/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ymssp.2012.12.003

and the means E_x and E_y are estimates as follows:

$$E_x = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i, \quad E_y = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i.$$

Comment. In the above formulas, we use the estimates for C, V_x, and V_y which are known to be biased. Usually, correlation is defined by using unbiased definitions

$$C = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-E_x)\cdot(y_i-E_y) = \frac{1}{n-1}\cdot\sum_{i=1}^{n} x_i\cdot y_i - E_x\cdot E_y,$$
$$V_x = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-E_x)^2, \quad V_y = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-E_y)^2.$$

One can easily check that the resulting expression for ρ is the same whether we use biased or unbiased estimates; we use biased estimates because they make the computations slightly simpler.

Known facts about correlation: *brief reminder*. It is known that the value of this correlation coefficient ρ is always between -1 and 1. The correlation is equal to 1 if and only if the values are positively linearly dependent, i.e., when for some coefficient $k_x > 0$, we have $y_i = E_y + k_x\cdot(x_i - E_x)$ for every i. The correlation is equal to -1 if and only if the values are negatively linearly dependent, i.e., when for some coefficient $k_x < 0$, we have $y_i = E_y + k_x\cdot(x_i - E_x)$ for every i.

Need to take into account interval uncertainty. The values x_i and y_i used to estimate correlation come from measurements, and measurements are never absolutely accurate: the measurement results $\tilde{x}_i$ and $\tilde{y}_i$ are, in general, different from the actual (unknown) values x_i and y_i of the corresponding quantities. As a result, the value $\tilde{\rho}$ estimated based on these measurement results is, in general, different from the ideal value ρ which we would get if we could use the actual values x_i and y_i. It is therefore desirable to determine how accurate is the resulting estimate.

Sometimes, we know the probabilities of different values of measurement errors $\tilde{x}_i - x_i$ and $\tilde{y}_i - y_i$. However, in many cases, we do not know these probabilities, we only know the upper bounds Δ_{xi} and Δ_{yi} on the corresponding measurement errors: $|\tilde{x}_i - x_i| \le \Delta_{xi}$ and $|\tilde{y}_i - y_i| \le \Delta_{yi}$; see, e.g., [15]. In this case, the only information that we have about the actual values x_i and y_i is that they belong to the corresponding intervals $[\underline{x}_i,\overline{x}_i] = [\tilde{x}_i - \Delta_{xi}, \tilde{x}_i + \Delta_{xi}]$ and $[\underline{y}_i,\overline{y}_i] = [\tilde{y}_i - \Delta_{yi}, \tilde{y}_i + \Delta_{yi}]$. Different values $x_i \in [\underline{x}_i,\overline{x}_i]$ and $y_i \in [\underline{y}_i,\overline{y}_i]$ lead, in general, to different values of the covariance. It is therefore desirable to find the range of all possible values of the covariance ρ:

$$[\underline{\rho},\overline{\rho}] = \{\rho(x_1,\ldots,x_n,y_1,\ldots,y_n) : x_i \in [\underline{x}_i,\overline{x}_i], y_i \in [\underline{y}_i,\overline{y}_i]\}.$$

The problem of computing the range of correlation under interval uncertainty is a particular case of the general problem of *interval computations* (see, e.g., [8,12]): computing the range of a given function $f(x_1,\ldots,x_n)$ under the interval uncertainty $x_1 \in [\underline{x}_1,\overline{x}_1]$, …, $x_n \in [\underline{x}_n,\overline{x}_n]$. Interval computations – in particular, interval computations of statistical characteristics – have many applications, in particular, engineering applications; see, e.g., [2,7,8–14,16].

For example, if we perform a statistical analysis of the measurement results, then, for each statistical characteristic $S(x_1,\ldots,x_n)$, we need to find its range

$$\mathbf{S} = \{S(x_1,\ldots,x_n) : x_1 \in \mathbf{x}_1,\ldots,x_n \in \mathbf{x}_n\}.$$

For the mean E_x, the situation is simple: the mean is an increasing function of all its variables. So, its smallest value $\underline{E}_x$ is attained when each of the variables x_i attains its smallest value $\underline{x}_i$, and its largest value $\overline{E}_x$ is attained when each of the variables attains its largest value $\overline{x}_i$:

$$\underline{E}_x = \frac{1}{n}\cdot\sum_{i=1}^{n}\underline{x}_i, \quad \overline{E}_x = \frac{1}{n}\cdot\sum_{i=1}^{n}\overline{x}_i.$$

Estimating correlation under interval uncertainty is NP-hard. In contrast to the mean – which is always monotonic – variance, covariance, and correlation are sometimes non-monotonic. It turns out that, in general, computing the values of these characteristics under interval uncertainty is NP-hard [3,4,13,14]. This means, crudely speaking, that unless P=NP (which most computable scientists believe to be wrong), no feasible (i.e., no polynomial-time) algorithm is possible that would always compute the range of the corresponding characteristic under interval uncertainty.

The problem of estimating correlation under interval uncertainty is formulated and analyzed in [16]; in that paper, this problem is formulated and solved as an optimization problem. For reasonably small n, the corresponding optimization algorithms work well [16]. However, since the problem is NP-hard, the computation time becomes infeasible when n is large.

What we do in this paper. We show that while we cannot have an efficient algorithm for computing both bounds $\underline{\rho}$ and $\overline{\rho}$, we can effectively compute (at least) one of the bounds. Specifically, we show that we can compute $\overline{\rho}$ when $\overline{\rho} > 0$ and we can compute $\underline{\rho}$ when $\underline{\rho} < 0$. This means that, in the case of a non-degenerate interval $[\underline{\rho},\overline{\rho}]$ (i.e., $\underline{\rho} < \overline{\rho}$):

- when $\overline{\rho} \le 0$, we compute the lower endpoint $\underline{\rho}$;
- when $0 \le \underline{\rho}$, we compute the upper endpoint $\overline{\rho}$;
- in all remaining cases, when $\underline{\rho} < 0 < \overline{\rho}$, we compute both lower endpoint $\underline{\rho}$ and $\overline{\rho}$.