Contents lists available at ScienceDirect

Measurement

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

Interval data fusion with preference aggregation

Sergey V. Muravyov[*], Liudmila I. Khudonogova, Ekaterina Y. Emelyanova

Department of Control Systems and Mechatronics, National Research Tomsk Polytechnic University, Pr Lenina 30, 634050 Tomsk, Russia

ARTICLE INFO

Keywords:
Data fusion
Rank aggregation
Interval data processing
Weak orders
Kemeny ranking problem
Triangle numbers
Preferences
Ties
Robustness
Monte Carlo simulation
Comparisons

ABSTRACT

It is proposed in the paper the interval data fusion procedure intended for determination of an interval to be consistent with maximal number of given initial intervals (not necessary consistent among each other) and to be with maximal likelihood including a value x^* that can serve as representative of all the given intervals. An algorithm of the interval fusion with preference aggregation (IF & PA) is proposed and discussed that can be carried out with help of representation of intervals on the real line by weak order relations (or rankings) over a set of discrete values belonging to these intervals. It is possible to determine a consensus ranking for collection of discrete values rankings, corresponding to initial intervals. The highest ranked value, accepted as a result of the fusion, guarantees improved accuracy and robustness of the interval data fusion procedure outputs. It is considered a space of weak orders induced by the intervals, its properties and dimension. A reasonable number choice problem of discrete values, representing the interval data, is investigated. Related to the problem, computing experiment results and recommendations are given. The interval data fusion procedures can be widely applied in interlaboratory comparisons, prediction of fundamental constant values on the base of different measured values, conformity testing, enhancement of multisensor readings accuracy in sensor networks, etc.

1. Introduction

Let x be a measurand, x_i be a result of i-th measurement, ε_i be a uncertainty of i-th measurement, and m be a number of measurements. Description of the measurement results in form of *intervals*

$$x_i = x \pm \varepsilon_i = x + |\varepsilon_i|, i = 1,...,m, \tag{1}$$

(the simplified form $x_i = x + \varepsilon_i$ is often used) whose bounds defined by experimentally obtained or given beforehand uncertainty values is rather common both in theory and practice of measurement [1–3]. Usually, sets of such data are input of measurement results processing procedure, aim of which, as a rule, is determination of a *unique summarized value* $\bar{x}$, by some justified way representing the input intervals.

Notice, carrying out of *multiple* measurements instead of a single one provides an opportunity to avoid blunders and enhance reliability of the determination of $\bar{x}$.

The classic procedure of *direct repeated* measurements processing described in many textbooks and guides in the field of metrology [4–7] allows to find the summarized value $\bar{x}$ as the *arithmetic mean* and its expanded uncertainty on the base of data of the same sample belonging to the same population where observations x_i were obtained by the same observer by means of the same methods and instruments under the same ambient conditions. Additionally, the biases are supposed to be absent in the measurement results, i.e. expectation $E(\varepsilon_i) = 0$; values ε_i are assumed to be independent and to have the same variances, i.e. $D(\varepsilon_i) = \sigma^2$ for all i. Such the observations sometimes are deemed to be *equidispersed*, that is equally distributed random variables [5,8].

Legitimacy of the arithmetic mean use can be easily justified if, first of all, to suggest that the intervals are equal to zero, i.e. $\varepsilon_i = 0$, and to try to find x from the m equations

$$x = x_i, i = 1,...,m. \tag{2}$$

For each separate equation, x is uniquely defined, however, these solutions are inconsistent among each other, and the whole system (2) is *inconsistent*. As the system solution, one can use an approximate value (estimate)

$$\bar{x} = f(x_1, x_2, ..., x_m) \tag{3}$$

that satisfies all m Eq. (2) with minimal error $\varphi = \sum_{i=1}^{m} (\bar{x} - x_i)^2$. The function φ has a minimum in a point where its derivative equals to 0, i.e. $\frac{d\varphi}{d\bar{x}} = 2\sum_{i=1}^{m} (\bar{x} - x_i) = 0$, whence it appears the estimate $\bar{x}$:

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i. \tag{4}$$

One can see from Eq. (4) that $\bar{x}$ is the arithmetic mean of measurement results x_i.

[*] Corresponding author.
E-mail address: muravyov@camsam.tpu.ru (S.V. Muravyov).

http://dx.doi.org/10.1016/j.measurement.2017.08.045
Received 5 April 2017; Received in revised form 5 July 2017; Accepted 28 August 2017
0263-2241/ © 2017 Elsevier Ltd. All rights reserved.

In order to check how the existence of uncertainty intervals affects the estimate (4) let us cancel the above supposition $\varepsilon_i = 0$ and substitute Eq. (1) for x_i into Eq. (4); we obtain $\bar{x} = \frac{1}{m}\sum_{i=1}^{m}(x + \varepsilon_i)$. Accounting that $E(\varepsilon_i) = 0$, we have the expectation of estimate (4)

$$\mu = E(\bar{x}) = E\left[\frac{1}{m}\sum_{i=1}^{m}(x + \varepsilon_i)\right] = E\left(\frac{mx}{m}\right) + E\left(\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\right) = x. \tag{5}$$

It follows from Eq. (5) that the estimate (4) coincides with the measurand x, i.e. it is *unbiased*. Taking into account $D(\varepsilon) = E[\varepsilon - E(\varepsilon)]^2$ and $D(a\varepsilon) = a^2 D(\varepsilon)$, if $a = \text{const}$, we obtain the variance of estimate (4)

$$D(\bar{x}) = D\left[\frac{1}{m}\sum_{i=1}^{m}(x + \varepsilon_i)\right] = D\left(\frac{mx}{m}\right) + D\left(\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\right) = \frac{1}{m^2}\sum_{i=1}^{m}D(\varepsilon_i) = \frac{m\sigma^2}{m^2} = \frac{\sigma^2}{m}, \tag{6}$$

whence it follows that the standard deviation of mean (4) is $\sqrt{m}$ times less than the sample standard deviation σ, that is, the estimate (4) is *consistent*. It is well known that the properties of unbiasedness, consistency, and also efficiency [1] of the estimate (4) are valid under *normal distribution* $N(\mu, \sigma^2)$ of measurement results.

There are many situations where it is necessary to find a most trustworthy quantity value and its uncertainty on the base of measurements carried out by different observers by means of various methods and instruments in different laboratories and/or ambient conditions. Series of obtained in this way observations are considered to be *not equidispersed*, if estimates of their variances are considerably differ from each other, i.e. $D(\varepsilon_i) = \sigma_i^2 \neq D(\varepsilon_{i+1}) = \sigma_{i+1}^2$, $i = 1, \ldots, m - 1$, and arithmetic means are estimates of the same value μ [4,5,8].

For example, it may be required to process the non-equidispersed measurement data in the following situations:

- evaluation of key comparison data provided by national metrology institutes (NMIs) for a single stable travelling standard [9,10];
- estimation of uncertainty of a fundamental physical constant that can be merely carried out by means of several principally different and independent methods [11];
- revealing uncorrected systematic measurement errors, which requires use of multiple investigators to measure a quantity in question [4,12];
- interlaboratory comparisons (ILC) to verify the technical competence of calibration or measurement laboratories where similar measurements are fulfilled by different laboratories by distinct means, and outcomes are compatible or not [13–15]; and
- analysis of some measuring instrument regular calibrations data accumulated for a long-term period where accuracy of observation series is different due to change of instrument's metrological performance in the course of time [16,17].

The list of the cases can be continued.

For produced in some of similar situations set of input intervals, the summarized value is defined as the *weighted arithmetic mean* [4,18]

$$y = \sum_{i=1}^{m} x_i\varepsilon_i^{-2} / \sum_{i=1}^{m}\varepsilon_i^{-2}, \tag{7}$$

whose uncertainty is

$$\varepsilon_y^2 = 1/\sum_{i=1}^{m}\varepsilon_i^{-2}, \tag{8}$$

where the weight of the measurement result x_i is typically the reciprocal square of the corresponding standard uncertainty $1/\varepsilon_i^2 = 1/\sigma_i^2$ [4]. It is well known that the weighted mean is the maximum likelihood estimate of the mean of independent *normal* distributions $N(\mu, \sigma_i^2)$ with the same mean μ [18].

Consider, for example, how the weighted mean (7) is employed in the popular Procedure A [19] used for processing ICL data where the main task is to establish a reference value x_{ref} that characterizes a largest consistent subset (LCS) of (reliable) measurement results provided by participating laboratories. When estimating the reference value x_{ref} the Procedure A uses a weighted mean value y and corresponding uncertainty ε_y calculated by Eqs. (7) and (8) where m is the number of participating laboratories; x_i is the nominal value estimate provided by i-th laboratory; and ε_i is corresponding standard uncertainty. The calculated weighted average value y is accepted as the reference value x_{ref} if it is consistent with the data provided by the participating laboratories in accordance with the criterion χ^2. If the consistency test is not passed, it is proposed in [19] to use a scheme of successive exclusion of outliers, i.e. measurement results which are not consistent with the others. A result is considered as inconsistent if the following condition is valid

$$|x_i - y| / \sqrt{\varepsilon_i^2 \pm \varepsilon_y^2} > 2, i = 1,\ldots,m. \tag{9}$$

The process of exclusion of one inconsistent result is repeated until the consistency of results by the criterion χ^2 is confirmed. The reference value for the obtained largest consistent subset is determined by Eq. (7), where the number of reliable laboratories m' is used instead of m. Clear that the Procedure A can be reasonably applied only if the measurement results provided by each participating laboratories are characterized by a normal probability distribution.

One can see from the above brief overview that purely statistical methods for interval processing have a lot of limitations imposed on permissible properties of the input intervals, such as normality of (series of) the data probability distributions, independence of observations, requirement of equidispersion, absence of outliers, etc. The typical recommendation to overcome the difficulties would be to use the non-parametric methods that, due to the reliance on fewer assumptions, are more *robust* than the parametric ones [20,21]. However, the non-parametric methods have their own shortcomings, in particular, even if they may work well on abnormal data, they have considerably less efficiency in cases when the normal distribution well enough approximates the measurement results. For example, in standard [13], in order to check the consistency of different laboratories measurement results, there are used estimates of laboratory bias, percentage differences, ranks and percentage ranks based on calculation of robust arithmetic mean and standard deviation. Authors of the document [13] point out that many of these methods are unlikely to be applicable when the number of participating laboratories is small (e.g. $m < 10$). Besides, the recommended procedures for outliers' detection can lead to unnecessary data removal [14].

In this paper we propose the approach to interval data processing that fundamentally does not use any their statistical properties and does not employ any parametric (like t- and F-tests) and non-parametric statistical significance tests. We call this approach *interval data fusion* where the latter two words "data fusion" designate the popular field of investigations and the former word refers to its interval-oriented specificity.

Data fusion is a process of joint processing of data on some object obtained from multiple sources aiming to acquire fuller, more objective and accurate knowledge of a characteristic under investigation than knowledge derived from a single source. List of data fusion methods usually includes mathematical statistics and probability theory (just in the context discussed above), fuzzy sets theory [22], possibility theory [3], Dempster-Shafer evidence theory [23], Bayesian inference [18], different artificial intelligence methods [24], methods of voting and preference (or rank) aggregation [2,25].

Under *interval data fusion* we will understand a procedure of shaping an interval to be consistent with maximal number of given initial intervals (not necessary consistent among each other) and to be with maximal likelihood including a value x^* that can serve as representative of all the given intervals. Evidently, both the arithmetic mean $\bar{x}$ and the