A bootstrap algorithm for mixture models and interval data in inter-comparisons

P. Ciarlini and G. Regoliosi

Istituto per le Applicazioni del Calcolo "M.Picone", CNR, Roma, Italy

F. Pavese

Istituto di Metrologia "G.Colonnetti", Torino, Italy

Abstract

To combine the information from several laboratories to output a representative value x_r and its probability distribution function is the main aim of an inter-comparison in Metrology. Here, the proposed procedure identifies a simple model for this probability function, by taking into account only the probability interval estimates as a measure of the uncertainty in each laboratory. A mixture density model is chosen to characterize the stochastic variability of the inter-comparison population considered as a whole. The bootstrap method is applied to approximate the distribution function of the comparison output in an automatic way.

1 Introduction

The "mise en pratique" of the Mutual Recognition Arrangement (MRA), issued by national metrological Institutions in 1999, prompted new studies and projects in Metrology mainly concerning the inter-laboratory comparisons area.

Recently, considerable effort has been devoted to finalise the problem of the choice of a suitable statistical procedure to summarise inter-comparison data. The problem solution is influenced by both metrological and statistical considerations, but it can also depend on the physical quantity under comparison.

Some of the critical issues now emerging are related to several different reasons. For instance, the statistical information supplied by each laboratory is synthetic, since it comes from a data reduction process performed on several experimental datasets. In each laboratory, assumptions and statistical reduction procedures may be different and sometimes not fully documented or the *a priori* information on the original data may be insufficient to define a "credible" probability distribution function (pdf) for output quantities of the inter-comparison.

The use of the whole sets of original data from each laboratory might be an unfeasible approach in the inter-comparison case, due to the unavailability of all needed data or to practical reasons. At present, the practice is to supply synthetic information x_i by each participant to the inter-comparison and to use a location estimator to output the representative value.

Efforts should be given to improving the reliability of inter-comparison results by asking for the use of any *a priori* information and of its "credibility" to go ahead, towards the direct estimation of the output of the comparison, x_r.

This paper proposes the identification of a solution without resorting to the synthetic values and its point estimates of the standard uncertainty, but only to the probability interval estimates as the measure of the uncertainty. This approach consists of two parts: a modelling procedure to identify a simple mixture model able to approximate the stochastic variability of the inter-comparison population as a whole; a parametric Monte Carlo algorithm to automatically estimate the probability distribution of the output x_r and any accuracy measures at a prescribed precision.

The concept of a mixture of distribution functions occurs when a population made up of distinct subgroups is sampled, for example, in biostatistics, when it is required to measure certain characteristics in natural populations of a particular species. In an inter-comparison each participant constitutes a subgroup.

The Monte Carlo method, based on the principle of mimicking sampling behaviour, can always compute a numerical solution in an automatic way, also when the required analytic calculations may not be simple. If the Monte Carlo method is applied with the principle of substitution (of the unknown probability function with a probability model estimated from the given sample), the approach is known as *the bootstrap* approach [4] and is already used in Metrology [2]. In [1] the case of a multivariate normal mixture model is considered and the standard errors are estimated by means of the parametric bootstrap. The present algorithm will be applied to a thermometric inter-comparison, where data cannot be assumed to be normally distributed.

2 Data structure of an inter-comparison with interval data

The number, N, of laboratories involved in an inter-comparison is typically small. In the i-th laboratory, the $(\xi_1^{(i)}, \ldots, \xi_k^{(i)})$ measurements are supposed to pertain to a single probability distribution function, say $F_i(\Lambda)$, where Λ is the parameter vector, that may be partially unknown. The measurements are statistically analysed and reduced to provide to the comparison the synthetic value x_i and its uncertainty u_i at 95% confidence level, or a 95% uncertainty interval (95%CI): $((x_1, u_1) \ldots, (x_N, u_N))$.

In this work the uncertainty is considered as "a 95%CI rather than as a multiple of the standard deviation" (see 4.3.4 in [6]). Then an aim of an inter-comparison is to combine the input data in the labs to characterise a representative value of the inter-comparison, i.e., the random variable θ and its pdf F. Hence a good estimate of the 95%CI for θ can be obtained if the output pdf F is a simple known function, describing the stochastic variability of the inter-comparison data. In other cases a suitable approximation of the expected value $E_F[X] = \int x dF(x)$ could be accepted to output the reference value x_r. The inter-comparison data structure is summarised here in terms of interval estimates:

INPUT Sample — Each one of the N participants originates a 95%CI that is one element of the inter-comparison sample:

$$\{[u_{il}, u_{iu}], i = 1, \ldots, N\}. \tag{2.1}$$

Here no value x_i in the interval $[u_{il}, u_{iu}]$ is chosen as representative; possible information on F_i (such as limited or unlimited support, symmetric or not) should be added. If a laboratory does not supply any information on the pdf, the uniform distribution is assumed.

Comparison OUTPUT — It includes the representative value and its 95%CI

$$(\hat{\theta}, [\epsilon_l, \epsilon_u]). \tag{2.2}$$

In many inter-comparisons, the differences to θ are also defined: $(y_i, [w_{il}, w_{iu}])$, where $y_i = x_i - \hat{\theta}, i = 1, \ldots, N$.

3 A classical approach to inter-comparisons

Let us recall the solution to the inter-comparison problem through the traditional estimator, the weighted mean. It is a location statistic that combines several measures and their standard uncertainties $(x_i, u_i)_{i=1}^N$. It provides the following estimate for θ,

$$\theta_w = u_w^2 \sum_{j=1}^{N} \frac{x_j}{u_j^2}, \quad u_w^2 = \left(\sum_{j=1}^{N} \frac{1}{u_j^2} \right)^{-1}, \tag{3.1}$$

and the following symmetric 95%CI,

$$\theta_w \pm k u_w, \tag{3.2}$$

where the coverage factor k is taken as the value $t_{N-1,0.95}$ of the Student distribution, N being small. In this approach, each x_i is viewed as an unbiased estimate of the laboratory mean value and the random variable θ_w is defined to be a linear combination of N independent random variables $X_1, \ldots, X_N$, where $\{x_1, \ldots, x_N\}$ is an observed sample. θ_w is supposed to be asymptotically normally distributed [6]. This estimator can be correctly adopted to solve an inter-comparison problem if the assumption of the homogeneity of the data is valid. This is equivalent to saying that, after considering the extent of the real effect and bias in each laboratory, the laboratories yield on the average the same value, so that the differences between the estimates are entirely due to random error. In this case, the selected estimator θ_w appropriately estimates θ and (3.2) accurately estimates its 95%CI.

Obstacles to applying this approach to a key-comparison have been discussed in [3]. The "credibility" of the representative values x_i, and of their uncertainty can critically affect the accuracy of the estimate of the representative value x_r. Moreover, the peculiar characteristics of a typical inter-comparison sample ((1) its very limited size, from a statistical point of view, (2) different experimental methods, used in each laboratory) often imply that the statistical assumptions are not satisfied, as for example in several thermometric cases. Indeed, the first characteristic implies that the Central Limit Theorem and the asymptotic theory do not hold. Then the normal distribution cannot be properly used to infer the estimates in (3.2).

Another example of the inadequacy of the weighted mean approach is when some laboratories provide data affected by bias, resulting from skewed distributions underlying their measurements. The symmetric confidence interval of (3.2) cannot be considered an

accurate approximation[1] of the true one, since it does not adjust for the skewness. Finally, it is necessary to point out that the homogeneity condition among the laboratories must be assured in some sense, otherwise it would be impossible to attempt to the computation of any summary estimate and its associated uncertainty.

4 The approach based on interval data

4.1 The mixture density function

This paper proposes to construct a simple model for the output pdf, and to estimate its expected value θ without requiring strong assumptions such as N large or each F_i normal. This approach enables us to compute the probability interval of the output value in terms of the identified density in each laboratory. The stochastic variability of the population of inter-comparison data is directly considered in the modelling approach as a whole, by means of a so-called mixture distribution model [5]. This model, being a linear superposition of several (say N) component densities, appears to be suitable from a computational point of view and can be embedded in a bootstrap algorithm to simulate several data needed to predict the output quantities.

In an inter-comparison, let us suppose that a density function $f_i(x;\Lambda^{(i)})$ is assumed for the i-th laboratory, then the following density mixture is identified to model the output pdf, where the parameter vector is $\Lambda = (\Lambda^{(1)},\ldots,\Lambda^{(N)})$ and given weights $\pi_i \geq 0, i = 1,\ldots,N$, have summation normalised to one:

$$g(x;\Lambda) = \sum_{i=1}^{N} \pi_i f_i(x;\Lambda^{(i)}). \tag{4.1}$$

To compute the output as estimate of the expected value of the mixture, $\theta = E_{G(\Lambda)}[X]$, the probability function $G(\Lambda)$, corresponding to the density in (4.1), must be known. When some laboratory provides only partial information on a pdf, we propose to identify its experimental variability by one of the following simple probabilistic models: uniform, normal or triangular pdf (right or left or symmetric triangular). Indeed, in thermometric experiments these three probabilistic models can represent several common stochastic variabilities for measurements, such as a limited or unlimited support, symmetric or not.

We want the mixture parameters to be estimated by means of the *INPUT Sample*, (2.1), as required in a bootstrap approach. Let us call I_i the *probability interval* to which the 100% measurements of the laboratory are supposed to pertain. For the uniform and the triangular types, $\Lambda^{(i)}$ parameters are defined to be the extremes of $I_i = [\lambda_{il}, \lambda_{iu}]$. For the normal model the parameters are the mean x_i and the variance u_i, while I_i becomes $(-\infty, +\infty)$.

A right triangular pdf (RT), a left triangular pdf (LT) or symmetric triangular pdf (ST) is chosen according to the position where the maximum of the probability density occurs, i.e., one extreme or the middle point of I.

[1] A 95% CI $[\epsilon_l, \epsilon_u]$ for θ is defined to be accurate if the following holds for every possible value for θ: $\text{Prob}_G\{\theta \geq \epsilon_u\} = 0.025$ and $\text{Prob}_G\{\theta \leq \epsilon_l\} = 0.025$

To compute the two components of the vector $\Lambda^{(i)} = (\lambda_{il}, \lambda_{iu})^T$ given the i-th input interval, a 0.025% portion of probability mass is added outside of each extreme, according to the supplied density shape. For example, if the ST density is chosen, the parameters are computed by:

$$\lambda_{il} = (0.89u_{il} - 0.11u_{iu})/0.78 \quad \lambda_{iu} = (0.89u_{iu} - 0.11u_{il})/0.78.$$

The mixture weights could be used to associate a degree of "credibility" to each laboratory. Then the choice $\pi_i = 1/N, i = 1, \ldots, N$, implies that every laboratory equally contributes to the inter-comparison.

When the mixture $G(\hat{\Lambda})$ is completely identified, it can be used to simulate data and to approximate the output value in the Monte Carlo algorithm.

4.2 The bootstrap algorithm

To avoid integral computations to estimate θ and its variance, the Monte Carlo method is commonly used to approximate them within a given precision. Since the *parametric* bootstrap approach does resampling from a parametric distribution model, in this case the mixture model $G(\hat{\Lambda})$, is adopted to approximate the following distribution,

$$H(x) = \text{Prob}_{\hat{G}}\{\theta^* \leq x\}. \tag{4.2}$$

The Monte Carlo method simulates a sufficiently high number B of data θ^* from $\hat{G} = G(\hat{\Lambda})$, to compute,

$$H(x)^{(B)} = \frac{1}{B}\sum_{b=1}^{B} \Pi\{\theta_b^* \leq x\}, \tag{4.3}$$

where the function $\Pi\{A\}$ is the indicator function of the set A. With probability one, it is known that the Monte Carlo approximation converges to the true value as $B \to \infty$. The Monte Carlo algorithm has been developed for a mixture density to estimate the comparison output. A hierarchical resampling strategy is used to reproduce the hierarchical variability in the inter-comparison population, throughout the following steps:

(1) (a) Choose at random an index, say k, of k-th laboratory by randomly resampling with replacement from the set $\{1, \ldots, N\}$

$$K \sim Prob\{K = k\} = \pi_i.$$

(b) Given k, generate, at random from the selected F_k of the distribution, a bootstrap value θ^* in $[\lambda_{kl}, \lambda_{ku}]$.

Repeat Step 1 B times to simulate the full bootstrap sample $\theta_1^*, \ldots, \theta_B^*$.

(2) Approximate the bootstrap mixture distribution as in (4.3) to compute:

— the bootstrap estimate of the expected mean

$$\hat{\theta_B^*} = \frac{1}{B}\sum_{b=1}^{B} \theta_b^*, \tag{4.4}$$

Lab1 (-0.05; 0.15) [-0.347, 0.247]	Lab2 (0.03; 0.30) [-0.564, 0.624]
Lab3 (0.18; 0.15) [-0.117, 0.477]	Lab4 (0.04; 0.15) [-0.257, 0.337]
Lab5 (0.71; 0.15) [0.413, 1.007]	Lab6 (-0.01; 0.15) [-0.307, 0.287]
Lab7 (-0.03; 0.15) [-0.327, 0.267]	

TAB. 1. Inter-comparison of 7 laboratories [7]: point estimates and simulated interval data.

— the bootstrap standard deviation: $Sd_B^* = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\theta_b^* - \hat{\theta_B^*})^2 \right)^{1/2}$,

— the 95%CI $[\epsilon_l^*, \epsilon_u^*]$, where the two extremes are computed as the α-th quantile [2] ($\alpha = 0.025$) of the bootstrap distribution $H_{Boot}^B(\alpha))^{-1} = q_B^{*\alpha}$, hence $\varepsilon_l^* = q_B^{*\alpha}$ and $\varepsilon_u^* = q_B^{*(1-\alpha)}$.

In Step 1b) the inverse transformation method has been used for simulating a random variable X having a continuous distribution F_k. For example, $X = F_k^{-1}(U)$, for a $U(\lambda_{kl}, \lambda_{ku})$ random variable. In Step 2 the bootstrap CI has been computed by means of the percentile method (see footnote). However, when the normal distribution is involved in the mixture, the *t-bootstrap* method gives more appropriate results [4]. To determine B in approximating the bootstrap confidence interval the coefficient of variation [4] can be used. The value of B is increased until the coefficient of variation cv of the sample quantile approaches the given precision δ_0. Indeed, from a metrological point of view, it appears easier to choose δ_0 instead of B as stopping rule in Step 1.

We would like to have also an automatic tool to investigate how well every laboratory contributes to the comparison, or to detect the possible presence of heterogeneous data. Here the concept of jackknife-after-bootstrap has been adopted to compute the mean and the bootstrap 95%CI. It is simply obtained by the following algorithm:

— for $i = 1, ..., N$, leave out the i-th lab and compute $\hat{\theta}_B^*(-i)$ and $q_B^*(-i)$,

— compare the N jackknife estimates to detect outlier values.

5 An application in thermometry

The proposed method is shown applied to an inter-comparison of Temperature Fixed Points, involving N =7 laboratories [7]. Each lab provided data x_i with the 95% standard uncertainty (Table 1: first item).

The second item (square brackets in the same table) represent the interval data generated with (3.2), that used to perform this simulated example. Since no specific pdf was supplied, the mixture distribution density has been constructed assuming the uniform type for each participant and equal weights. The parameters of every uniform density was computed using interval data, and the obtained mixture density was used in the resampling step of the algorithm to compute the representative value and its

[2]The percentile method of a statistics θ, based on B bootstrap samples, simply gives for a α-percentile $q_B^{*\alpha}$ = {(αB)th largest for θ_b^*}

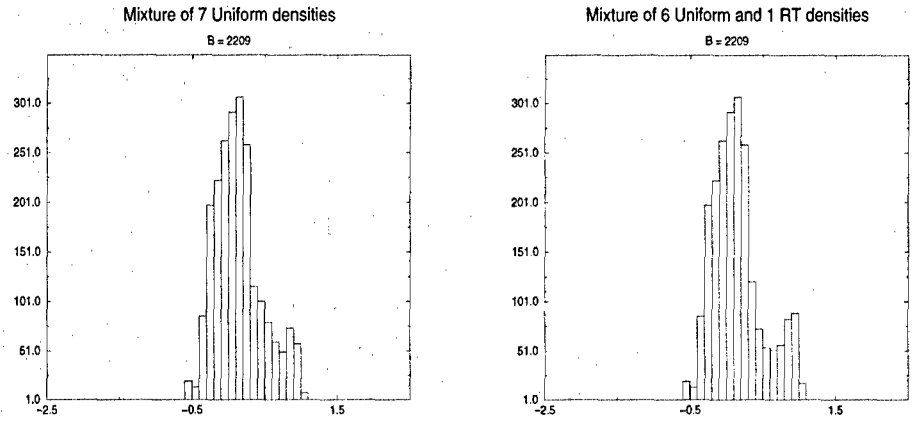

FIG. 1. Bootstrap histograms B =2209: left–mixture of 7 uniform distributions; right–mixture of 6 ST plus one RT density for Lab$i$.

probability interval with $\delta_0 = 0.05$. In Figure 1 (left) the bootstrap histogram, that approximates the mixture density, shows a bimodal behaviour. The computations are obtained for $\delta_0 = 0.05$ or $B = 2209$: $\hat{\theta}^* = 0.14$, bootstrap standard deviation Sd^*=0.33, 95%CI [-0.35, 0.92].

The proposed algorithm was also applied with a mixture of seven normal densities, and the results are $\hat{\theta}^* = 0.13$, $Sd^* = 0.43$, bootstrap 95%CI [-0.61, 1.1] for B =4752. The effect of assuming unlimited symmetric distributions to model the output pdf results in a wider 95%CI for a mixture of normal densities.

By comparing the jackknife results in Table 2, Lab5 appears to supply unusual values. To directly consider this behaviour in the inter-comparison, a mixture of six uniform densities plus a RT density, identifying Lab5, has been constructed. The approximated bootstrap distribution is displayed in Fig.1 (left), with bootstrap estimates, $\hat{\theta}^* = 0.15$, standard deviation $Sd^* = 0.35$ and [-0.35, 0.96] for the Bootstrap 95%CI, obtained for $B = 2209$.

6 Conclusions

The problem of the inter-comparison data has been described, and a new approach has been proposed. It is based on the uncertainty estimates, that should be provided by each Laboratory as interval estimate at 95% confidence level together with information, also partial, on the probability function. The constructive procedure directly characterises the stochastic variability of the reference value of the inter-comparison, by means of a mixture density model. The result of an inter-comparison is then viewed as a random variable, not directly measured, being the output of a complex process, that involves measures, statistical information and metrological considerations. These considerations suggest us constructing a mixture, with weights π_i to take into account each participating laboratory according to its credibility.

Lab1	0.34	[-0.45, 0.92]	Lab2	0.32	[-0.31, 0.94]
Lab3	0.34	[-0.40, 0.91]	Lab4	0.34	[-0.35, 0.92]
Lab5	0.23	[-0.42, 0.48]	Lab6	0.34	[-0.36, 0.95]
Lab7	0.34	[-0.42, 0.92]			

TAB. 2. Jackknife-after-bootstrap estimates. Standard deviation and 95%CI for mixture of 6 uniform densities ($B = 1000$): in the ith item, Lab$i$ is left out.

The parametric bootstrap approach has been adopted to estimate in a simple and automatic way the inter-comparison output, where information, even partial, on the probability hierarchical data of the participating laboratories, have been taken into account.

Also with a limited number of laboratories, the method can be applied, as it is shown in the thermal example, where ($N = 7$) and the experimental conditions implied to adopt skewed distributions. The automatic jackknife method of detecting the heterogeneous data succeeded in revealing an unusual value. To take into account this condition, a mixture of six uniform densities plus an RT density to identify Lab5 could be better used. The choice of equal weights emphasises that all the standards have equally contributed to the inter-comparison.

The bootstrap procedure, completely developed for a class of five simple distribution functions often used in thermal metrology, could be adapted to consider other distributions, when the synthetic data information provided by the laboratories, as summarised in Section 2, allow to compute the mixture parameters.

Bibliography

1. K. E. Basford, D. R. Greenway, G. J. McLachlan and D. Peel, *Standard errors of fitted component means of normal mixtures.* Computational Statistics **12**, 1–17, 1997.
2. P. Ciarlini et al., *Non-parametric bootstrap with application to metrological data.* In: Advanced Mathematical Tools in Metrology, Series on Advances in mathematics for applied sciences, **16**, Singapore, Ciarlini, Cox, Monaco, Pavese eds., World Scientific, 219–230, 1994.
3. M. Cox, *A discussion of approaches for determining a reference value in the analysis of key-comparison data.* In Advanced Mathematical and Computational Tools in Metrology IV, Series on Advances in mathematics for applied sciences, **53**, Singapore, Ciarlini, Cox, Pavese, Richter Eds , World Scientific, 45–65, 2000.
4. B. Efron and R. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, London, 1993.
5. B. S. Everitt, Finite Mixture Distributions, Chapman and Hall, London, 1981.
6. ISO, Guide to the Expression of Uncertainty in Measurement, Geneva, Switzerland, 1995.
7. F. Pavese, Monograph 84/4 of Bureau International des Poids et Mesures, BIPM Sevres, 1984.