CCT/01-
(Accepted at "Algorithms for Approximation IV", Huddersfield, UK, July 2001)

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

P. Ciarlini[1], F. Pavese[2], and G. Regoliosi[1]

[1] *Istituto per le Applicazioni del Calcolo "M.Picone", CNR, Viale del Policlinico, 137, 00161 Roma, Italia*
[2] *Istituto di Metrologia "G.Colonnetti", CNR, Strada delle Cacce, 73, 10135, Torino, Italia*

Abstract

The aim of an inter-comparison in Metrology is to combine the input information from several Laboratories to output a representative value x_r and its probability distribution function.

The procedure proposed in this paper identifies a simple model for this probability function, by taking into account the probability interval estimates as a measure of the uncertainty in each Lab. 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. The algorithm, developed for a thermometric inter-comparison, is described and the results are shown.

1 Introduction

The "mise en pratique" of the Mutual Recognition Arrangement (MRA) [1] prompted new studies and projects in Metrology mainly concerning the inter-Laboratory comparisons area, in order to define the degree of equivalence according to the results of the "key comparisons".

Recently, a considerable effort has been devoted to finalise the problem of the choice of a suitable statistical procedure to summarise inter-comparison results. It is influenced by both metrological considerations and by statistical assumptions, but it can also depend on the physical quantity under comparison as discussed in a recent (1999) Workshop at NPL, UK.

Some of the critical issues now emerging are related to several different conditions. For instance: the statistical information supplied by each Laboratory for the comparison 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. Moreover, the largely unknown intensive pre-processing on the dataset may introduce additional masking effects or bias, that cannot be modelled, as shown in [10] in processing tabulated literature data, where it was impossible to trace back to the original dataset.

It is known that the use of the whole sets of original data from each Laboratory, as mentioned in [8], might be an approach to avoid some of the inconveniences arising from the indirect process of performing several separate reductions of the dataset and subsequent data fusion. But this approach may be unfeasible 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.

Because of the above mentioned reasons, 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 identification of a solution without resorting to the synthetic values and the point estimates of the related 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 characterise the stochastic variability of the inter-comparison population as a whole; a parametric Monte Carlo algorithm to approximate the probability distribution of the output x_r, to estimate its variance or other accuracy measures at a prescribed level of confidence.

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. Here each participant in an inter-comparison constitutes a subgroup.

The choice of the Monte Carlo method arises from its ability to automatically compute a numerical solution, also when the required analytic calculations may not be simple. It is based on the principle of mimicking sampling behaviour by generating a large amount of simulated data. If the Monte Carlo approach is applied with the principle of substitution (of the unknown probability function with a probability model estimated from the given sample) [12], the approach is known as *bootstrap* approach [6]. In [2] 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 has been developed and applied to a thermometric inter-comparison, where data cannot necessarily assumed to be Normally distributed.

2 Data structure of an inter-comparison with interval data

The number N of the 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 [9]). In this general frame, let us say that one aim of an inter-comparison is to combine the input data and the information on the probability functions $F_i(\Lambda)$ in the Labs to characterise a representative value of the inter-comparison, in other words the random variable θ and its pdf F. Hence an accurate approximation of the $95\%CI$ for θ can be obtained if the output pdf F is known, where F depends on the stochastic variability of the data in each Lab. A suitable estimate $\hat{\theta}$ 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, in terms of interval estimates, are here considered and summarised as follows:

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

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

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 pdf, the Uniform distribution is assumed.

Comparison OUTPUT - It includes the representative value and its $95\%CI$

$$(\hat{\theta}, [\epsilon_l, \epsilon_u]). \tag{2}$$

In many inter-comparisons, the differences to θ are also defined: $(y_i, [w_{il}, w_{iu}])$, where $y_i = x_i - \hat{\theta}, i = 1, \ldots, N$. In the case of the key-comparisons in [1], $\hat{\theta}$ is called *Key Comparison Reference Value*.

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 :

$$\theta_w = u_w^2 \sum_{j=1}^{N} \frac{x_j}{u_j^2}, \quad u_w^2 = \left(\sum_{j=1}^{N} \frac{1}{u_j^2} \right)^{-1} \tag{3}$$

and the following symmetric $95\%CI$:

$$\theta_w \pm k u_w \tag{4}$$

where the coverage factor k is taken as the value $t_{N-1,0.95}$ of the Student distribution, being N 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. The variable θ_w is supposed to be asymptotically normally distributed [9]. 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 say 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 eq.4 accurately estimates its $95\%CI$.

Inconveniences to apply this approach to a key-comparison have been discussed in [5]. 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 [10]. Indeed, the first characteristic implies that the Central limit Theorem and the asymptotic theory do not hold. Then the Normality distribution cannot be properly used to infer the estimates in eq.4.

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 eq.4 cannot be considered an accurate approximation [1] of

[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: $Prob_G\{\theta \geq \epsilon_u\} = 0,025$ and $Prob_G\{\theta \leq \epsilon_l\} = 0,025$

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 to compute the probability interval of the output value in term of the identified density in each Lab. The stochastic variability of the population of inter-comparison data is directly considered in this modelling approach as a whole, by means of a so-called mixture distribution model [7]. This model 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.

A finite mixture density function is a linear superposition of several (say N) component densities.

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, where the parameter vector is $\Lambda = (\Lambda^{(\mathbf{1})}, \ldots, \Lambda^{(\mathbf{N})})$ and weights $\pi_i \geq 0, i = 1, \ldots, N$ have summation normalised to 1, can be identified to model the output pdf:

$$g(x;\Lambda) = \sum_{i=1}^{N} \pi_i f_i(x;\Lambda^{(\mathbf{i})}) \tag{5}$$

To compute the output value as estimate of the expected value of this mixture distribution, $\theta = E[G(\Lambda)]$, the probability function $G(\Lambda)$ should be known. Since some Labs may provide only partial information on pdf, we propose to model its stochastic variability by using 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, such as a limited or unlimited support, symmetric or not, for the measurements.

We want the mixture parameters to be estimated by means of the *INPUT Sample* of eq.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 $\Lambda^{(\mathbf{i})}$ 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 (LT) or symmetric triangular (ST) is chosen according to the position where the maximum of the probabilty density occurs, i.e. one extreme or the middle point of I.

To compute a bi-dimensional vector $\Lambda^{(\mathbf{i})} = [\lambda_{il}, \lambda_{iu}]^{-1}$ 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 Lab. 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 bootstrap algorithm.

4.2 The bootstrap algorithm

To avoid integral computations for the expected mean θ and the variance, the Monte Carlo method is used to approximate them within a given precision ([6]). The *parametric* bootstrap does resampling from a parametric distribution model (in this case a mixture model $G(\hat{\Lambda})$). Then the following distribution:

$$H(x) = Prob_{\hat{G}}\{\theta^* \leq \ x\} \tag{6}$$

is approximated by applying the Monte Carlo method, i.e. a sufficiently high number B of data θ^* are simulated from $\hat{G} = G(\hat{\Lambda})$, to compute:

$$H(x)^{(B)} = \frac{1}{B}\sum_{b=1}^{B} \Pi\{\theta_b^* \leq x\} \tag{7}$$

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 $\{1, \ldots, N\}$:

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

- 1.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 eq.7 to compute:
  - the bootstrap estimate of the expected mean

$$\hat{\theta_B^*} = \frac{1}{B}\sum_{b=1}^{B} \theta_b^* \tag{8}$$

  - the bootstrap standard deviation

$$Sd_B^* = (\frac{1}{B-1}\sum_{b=1}^{B} (\theta_b^* - \hat{\theta_B^*})^2)^{1/2}$$

  - 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 $\varepsilon_u^* = q_B^{*(1-\alpha)}$.

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

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]	

Table 1: Inter-comparison of 7 Laboratories [11]: point estimates and simulated interval data

In step 1.b) 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 [6]. To determine B in approximating the bootstrap confidence interval the coefficient of variation [12] 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) as in [4].

We would like to have also an automatic tool to investigate how well every Lab 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 now shown applied to an inter-comparison of Temperature Fixed Points, involving N =7 Laboratories [11]. In each Lab the provided data are: x_i with the 95% standard uncertainty (Table 1).

In the same Table, the second items (in square brackets) represent the interval data generated with eq.4, that have been used in this simulated example. Since no specific pdf was supplied, the Mixture distribution density has been constructed assuming the Uniform model for each participant with equal weights. The parameters of every Uniform density was computed using the input interval data, and the obtained Mixture density was used in the bootstrap algorithm to compute the representative value and its probability interval with $\delta_0 = 0.05$. The results 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]. Figure 1-$left$ the bootstrap histogramme, that approximates the Mixture density, shows a bimodal behaviour

The proposed procedure 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 the 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 Boostrap 95%CI, obtained for $B = 2209$.

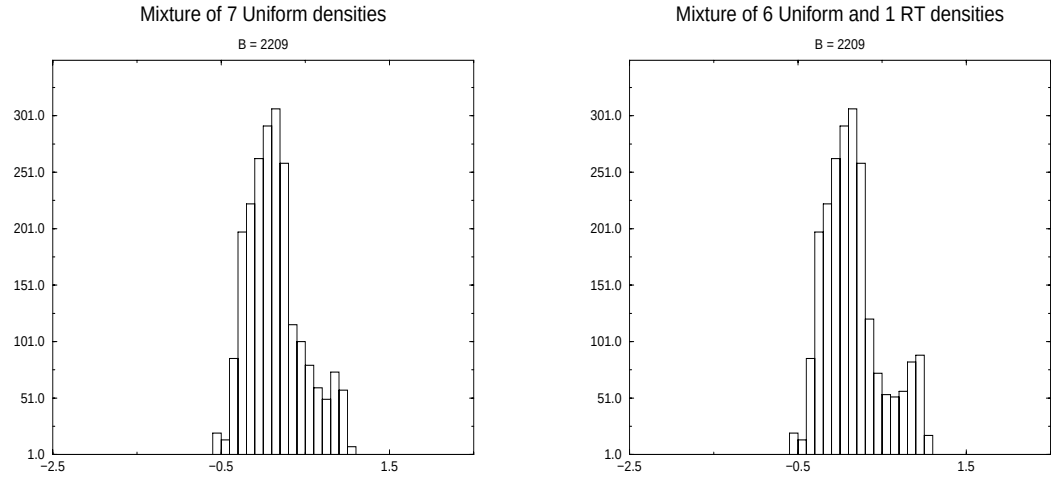

Figure 1: Bootstrap histograms B =2209: $left$- Mixture of 7 Uniform distributions, $right$ Mixture of 6 ST plus one RT density for Lab$i$.

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]			

Table 2: Jackknife-after-boostrap estimates. Standard deviation and 95%CI for Mixture of 6 Uniform densities ($B = 1000$): in the $i-th$ item, Lab$i$ is left out

6 Conclusions

The problem of the inter-comparison data has been briefly illustrated, and an approach different from the classical one has been proposed. It is based on the uncertainty estimates that should be provided by each Laboratory as the confidence interval at 95% level together with information, also partial, on the probability function. The proposed 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 that cannot be directly measured, being the output of a complex process, that involves measures, summary statistical information and metrological considerations. These considerations suggest constructing a mixture, with weights π_i that must be attributed to each participating Laboratory according to its credibility. When the same credibility is associated to each Laboratory the choice $\pi_i = 1/N$ implies that each Laboratory equally contributes to the inter-comparison.

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

Also with a limited number of involved Laboratories ($N = 7$) the method can be applied, as it happens in the thermal metrology example, when the experimental conditions imply to adopt skewed distributions. The automatic method of detecting the heterogeneous data, based on a

jackknife strategy, has revealed an unusual value. To take into account this condition, a mixture of six Uniform densities and a RT density to identify Lab5 could be better used. The choice of equal weights emphasises that all the standards have contributed.

The parametric bootstrap procedure has been completely developed for a class of five simple distribution functions, that can simply model several practical experimental conditions in thermal metrology. However, the algorithm could be adapted to other distribution functions, where the synthetic information provided by the Laboratories, as summarised in section 2, unables the preliminary estimate of the distribution parameters of the mixture model.

References

[1] BIPM, Mutual Recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, BIPM publication, Sevres (1999)

[2] K.E. Basford, D.R. Greenway, G.J. McLachlan and D.Peel, Computational Statistics, 12 (1997)

[3] P. Ciarlini et al., in: Advanced Mathematical Tools in Metrology, Series on Advances in mathematics for applied sciences, 16, Singapore, Ciarlini, Cox, Monaco, Pavese Eds , World Scientific (1994), 219

[4] P. Ciarlini et al, Computation in Statistics, Simulation and computation, 28 (1999), 821

[5] M. Cox,. 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 (2000), 45

[6] B. Efron and R. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall (1993)

[7] B. S. Everitt, Finite Mixture Distributions, Chapman and Hall, London (1981)

[8] E. Filipe, in Advanced Mathematical and Computational Tools in Metrology V, Series on Advances in mathematics for applied sciences, 57, Singapore, Ciarlini, Cox, Filipe, Pavese, Richter Eds , World Scientific (2001), 142

[9] ISO, Guide to the Expression of Uncertainty in measurement, Geneva, Switzerland (1995)

[10] F.Pavese, P.Ciarlini, Metrologia 27 (1990) 145

[11] F.Pavese "Final Report of the International Intercomparison of fixed points by means of sealed cells: 13.81 K to 90.686 K", Monograph 84/4 of Bureau International des Poids et Mesures, BIPM Sevres (1984)

[12] J. Shao, D.Tu, The Jackknife and Bootstrap, Springer, NY (1995)