Fusion learning for Inter-laboratory Comparisons

Jan Hannig[*], Qing Feng
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill

Hari Iyer, C. M. Wang
National Institute of Standards and Technology

Xuhua Liu[†]
China Agricultural University

July 25, 2017

Abstract

In this paper we propose a Generalized Fiducial Inference inspired method for finding a robust consensus of several independently derived collection of confidence distributions (CDs) for a quantity of interest. The resulting fused CD is robust to the existence of potentially discrepant CDs in the collection. The method uses computationally efficient fiducial model averaging to obtain a robust consensus distribution without the need to eliminate discrepant CDs from the analysis. This work is motivated by a commonly occurring problem in inter-laboratory trials, where different national laboratories all measure the same unknown true value of a quantity and report their CDs. These CDs need to be fused to obtain a consensus CD for the quantity of interest. When some of the CDs appear to be discrepant, simply eliminating them from the analysis is often not an acceptable approach, particularly so in view of the fact that the *true value* being measured is not known and a discrepant result from a lab may be closer to the true value than the rest of the results. Additionally, eliminating one or more labs from the analysis can lead to political complications since all labs are regarded as equally competent. These considerations make the proposed method well suited for the task since no laboratory is explicitly eliminated from consideration. We report results of three simulation experiments showing that the proposed fiducial approach has better small sample properties than the currently used naive approaches. Finally, we apply the proposed method to obtain consensus CDs for gauge block calibration inter-laboratory trials and measurements of Newton's constant of gravitation (G) by several laboratories.

Keywords: Confidence distributions, generalized fiducial inference, model averaging, inter-laboratory trials, key comparison experiments

[*]Jan Hannig's research was supported in part by the National Science Foundation under Grant No. 1016441 and 1512945.

[†]Xuhua Liu's research was partly supported by the National Natural Science Foundation of China (Grant No.11201478).

1 Introduction

Inter-laboratory trials are often conducted by leading metrology laboratories in the world to compare each others' capabilities for measuring various fundamental properties of substances. Such a trial typically involves two or more participants each of whom measures the (nominally) same unknown value (called *measurand*) and provides the result along with an assessment of the uncertainty in the result. The results are meant to be the best estimates of the measurand the participating laboratories are able to provide. Often the same or very similar protocols are used by the participating laboratories. In some cases different subsets of participants use different methods for measuring the same unknown quantity. This is particularly so when specific laboratories have special expertise in particular measurement methods. The results from such experiments are used to gauge how comparable the measurement capabilities are across the participating laboratories. In some cases such experiments lead to the creation of certified reference materials (CRMs) and a consensus value for the measurand is arrived at by combining the results from the participating laboratories. This consensus value is used as the certified value for the CRM. The uncertainty associated with this certified value is used to provide an interval estimate of the value for the CRM.

Key Comparisons

There is a particular class of inter-laboratory trials which takes on international significance. With the signing of the Mutual Recognition Arrangement (MRA) CIPM (1999) in 1999, National Metrology Institutes (NMI's) and Regional Metrology Organizations (RMO's) around the world have undertaken the task of examining the *degree of equivalence* of their measurement standards. The CIPM (*Comité international des poids et mesures* – The International Committee on Weights and Measures), an entity whose principal task is to promote world-wide uniformity in units of measurement, works with member countries on issues related to the creation of measurement standards and comparisons of measurement capabilities of various national metrological laboratories (such as the National Institute of Standards and Technology (NIST) in the U.S, the National Physical Laboratory (NPL) in Great Britain, and Physikalisch-Technische Bundesanstalt (PTB) in Germany), and oversees the conduct of inter-laboratory experiments by participating NMIs to evaluate the relative measurement capabilities of each other and also to establish standard reference values (called Key Comparison Reference Value(s) or KCRV) for many important fundamental measurements and standards. The results obtained by the different laboratories are combined to arrive at the consensus KCRV value. Such comparisons *provide for the mutual recognition of calibration and measurement certificates issued by NMIs and thereby to provide governments and other parties with a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs.*

During any inter-laboratory trial it is generally the case that the results from one or a few laboratories differ noticeably from the rest even though all participating laboratories are considered to be more or less equally competent. It is natural to think that these apparently nonconforming values should perhaps be excluded from the calculation of a consensus value. There are at least two problems with this thinking. First, since the true value of the measurand is not known, one cannot say, based on any objective evidence, that one result is more believable than another. Second,

there are political overtones associated with leaving out measured results of a laboratory since all participating laboratories are considered to be competent in their own right. Although discrepant results are subjected to further scrutiny to make sure such discrepancies are not the result of identifiable errors, when no errors are identified, each laboratory stands behind its result and the associated statements of uncertainty. Hence the problem of arriving at a consensus value takes on a greater level of significance when it comes to International Key Comparison Studies.

Gauge Blocks

A gauge block (Thalmann (2002)) is a length standard having flat and parallel opposing surfaces. The cross-sectional shape is not very important, although the standard does give suggested dimensions for rectangular, square and circular cross-sections. Gauge blocks have nominal lengths defined in either the metric system (millimeters) or in the English system (1 inch = 25.4 mm). The length of the gauge block is defined at standard reference conditions:

temperature = 20 °C (68 °F)
barometric pressure = 101,325 Pa (1 atmosphere)
water vapor pressure = 1,333 Pa (10 mm of mercury)
CO2 content of air = 0.03%.

The length of a gauge block is defined as the perpendicular distance from a gauging point on one end of the block to an auxiliary true plane wrung to the other end of the block, as shown in Figure 1.

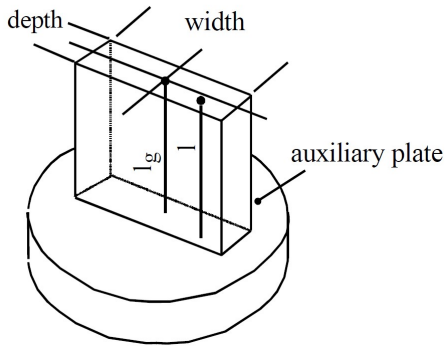

Figure 1: The length of a gauge block is the distance from the gauging point on the top surface to the plane of the platen adjacent to the wrung gauge block.

Figure 2 shows a portion of the results from an international key comparison study (CCL-K1) involving the measurement of the central length of steel gauge blocks (nominal length 8 mm) using interferometry according to ISO 3650. Detailed results are available from the website of the International Bureau of Weights & Measures (BIPM). The URL for the website is `http://kcdb.bipm.org/`. For instance, one can see, given the reported uncertainties, VNIIM (D.I. Mendeleev All-Russian Institute for Metrology) appears to deviate the most from the rest of the measurements.

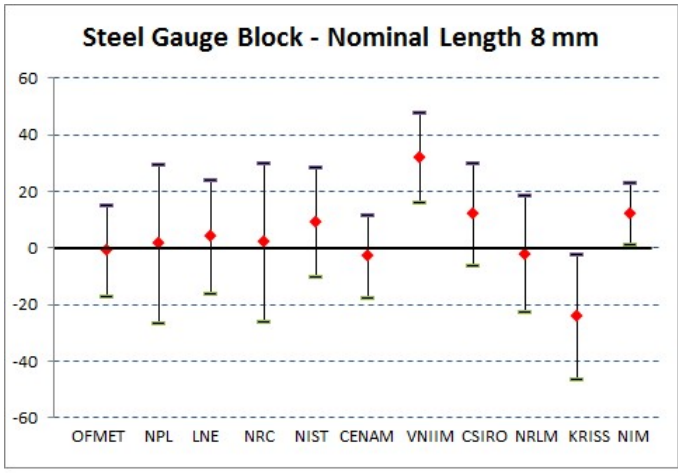

Figure 2: Gauge Block Measurements by 11 National Metrological Laboratories. Nominal length is 8 mm. The horizontal axis shows deviations (in nm) from the nominal value.

One of the issues that needs to be resolved is "how to treat this apparent outlier?" Alternatively, how much weight should be given to this particular measurement if one were using a weighted average approach to arrive at the KCRV?

The Key Comparison Study also considered gauge blocks of other nominal lengths besides the 8 mm gauge block. The estimates and uncertainties for the full set of steel gauge blocks for the 11 NMIs is given in Table 1. The entire array of issues related to this problem is more involved than what we are able to present here.

Although the potential applications of Fiducial methods in this area has been investigated in the literature (Iyer *et al.* (2004a,b)), a systematic and thorough treatment of Fiducial Methods robust to potential outliers has not been carried out. In this paper we propose the use of generalized fiducial model averaging approach applied to confidence distributions reported by the laboratories to finding a consensus value robust to potentially discrepant laboratories.

When combining together information from the participating labs we use *fusion learning* techniques (CD combination techniques) based on Generalized Fiducial Inference ideas of Hannig and Xie (2012). This is described in Section 2. In particular, a novel, highly computationally efficient algorithm for model averaging is presented in Section 2.3. We show good small sample properties of the proposed method in Section 3. We demonstrate the new technique on the steel gauge block data and measurements of Newton's constant of gravitation (G) in Section 4. Section 5 concludes with a summary.

Lab	Nominal Lengths (in mm)								
	0.5	1.01	6	7	8	15	80	90	100
OFMET	17 ± 9	34 ± 9	52 ± 8	31 ± 8	-1 ± 8	16 ± 8	22 ± 11	-21 ± 12	-96 ± 13
NPL	20 ± 14	25.5 ± 14	54.5 ± 14	33.5 ± 14	1.5 ± 14	22.5 ± 15	38.5 ± 28	-14 ± 31	-140 ± 33
LNE	15 ± 10	25 ± 10	54 ± 10	35 ± 10	4 ± 10	20 ± 10	28 ± 14	-24 ± 15	-110 ± 16
NRC	29 ± 13	28 ± 13	36 ± 14	30 ± 14	2 ± 14	14 ± 14	9 ± 21	-37 ± 22	-126 ± 24
NIST	26 ± 8.9	42 ± 9	57 ± 9.4	34 ± 9.5	9 ± 9.6	30 ± 10.3	33 ± 16.1	-23 ± 17	-117 ± 17.9
CENAM	15 ± 7	20 ± 7	47 ± 7.1	26 ± 7.1	-3 ± 7.2	13 ± 7.4	21 ± 15.6	-19 ± 17.3	-119 ± 18.7
VNIIM	*	60 ± 8	68 ± 8	25 ± 8	32 ± 8	36 ± 12	25 ± 14	-32 ± 15 -	104 ±
CSIRO	28 ± 9	46 ± 9	53 ± 9	37 ± 9	12 ± 9	51 ± 9	27 ± 14	-20 ± 15	-114 ± 16
NRLM	23.9 ± 8.6	17.7 ± 10.3	44.1 ± 10.3	27 ± 8.7	-2.2 ± 10.3	15.1 ± 10.9	47.3 ± 13.5	9.1 ± 14.3	-89.4 ± 16.3
KRISS	18.7 ± 13.1	20.3 ± 12.2	22.1 ± 13.6	12.8 ± 11	-24.2 ± 11	8.1 ± 13.2	30.4 ± 17	-18.4 ± 18.9	-104.3 ± 20.6
NIM	30 ± 5.4	48 ± 5.4	56 ± 5.5	42 ± 5.5	12 ± 5.5	28 ± 5.6	44 ± 8.9	18 ± 9.6	-90 ± 10.3

Table 1: CCL-K1 Measured results by 11 NMIs and combined standard uncertainties for steel gauge blocks for 9 different nominal lengths. The nominal lengths are in millimeters (mm). The values shown in the table are deviations from the nominal values (in nm) plus or minus the combined standard uncertainty (also in nm)

2 Method

2.1 Confidence Distributions

A Confidence Distribution (CD) is a way to summarize information about a parameter contained in the data. It is similar to Bayes posterior but is grounded in frequentist methodology. Heuristically speaking, the CD function is obtained by stacking up one-sided confidence intervals of all levels. Schweder and Hjort (2002); Singh *et al.* (2005) provide the following formal definition of a CD function.

Definition 1. A function $H(\theta|\mathbf{y})$ on $\Theta \times \mathcal{Y} \to [0,1]$ is called a confidence distribution (CD) for a parameter θ if it follows two requirements:

1. For each given $\mathbf{Y} \in \mathcal{Y}$, $H(\cdot|\mathbf{y})$ is a continuous cumulative distribution function on Θ;

2. At the true parameter value $H(\theta_0|\mathbf{Y})$, as a function of the random $\mathbf{Y}$ (generated from the distribution determined by the true parameter θ_0), follows a uniform distribution $U[0,1]$, i.e., $P_{\theta_0}\left(H(\theta_0|\mathbf{Y}) \leq u\right) = u$ for all $0 < u < 1$.

The function $H(\theta_0|\mathbf{Y})$ is a *conservative CD* if condition 2 is replaced by $P_{\theta_0}\left(H(\theta_0|\mathbf{Y}) \leq u\right) \leq u$, for all $0 < u < 1/2$ and $P_{\theta_0}\left(H(\theta_0|\mathbf{Y}) \leq u\right) \geq u$, for all $1/2 < u < 1$.

The function $H(\theta|\mathbf{y})$ is an *asymptotic CD (aCD)*, if condition 2 is true only asymptotically (as sample size goes to infinity) and the continuity requirement in condition 1 is dropped.

In general, fiducial distribution, objective Bayes posterior distribution, inversion of one sided confidence intervals are all examples of CDs or aCDs. To demonstrate the idea of a CD on a simple example, consider a sample of size n from a $N(\theta, \sigma^2)$ distribution with sample mean $\bar{x}$ and sample standard deviation s. The data is summarized as $\mathbf{y} = (\bar{x}, s^2)$ and the corresponding CD is the location-scale t distribution with distribution function

$$H(\theta|\mathbf{y}) = F^t_{n-1}\left(\frac{\theta - \bar{x}}{s/\sqrt{n}}\right),$$

where F^t_{n-1} is the distribution function of the Student's T distribution with $n-1$ degrees of freedom.

A useful graphical tool for visualizing a confidence distribution is a confidence curve (Birnbaum, 1961). For a given confidence distribution $H(\theta|\mathbf{y})$, a corresponding confidence curve is defined as $CV(\theta) = 2|H(\theta|\mathbf{y}) - 0.5|$. On a plot of $CV(\theta)$ versus θ, each sub-level set $\{\theta : CV(\theta) \leq \alpha\}$ is α-level confidence set, $0 < \alpha < 1$. Thus, a confidence curve is a graphical device that shows confidence intervals of all levels; see, e.g. Birnbaum (1961); Bender *et al.* (2005). The minimum of a confidence curve is the median of the confidence distribution. It provides a point estimator which is typically median unbiased (Birnbaum, 1961). Figure 3 shows an example of a confidence curve.

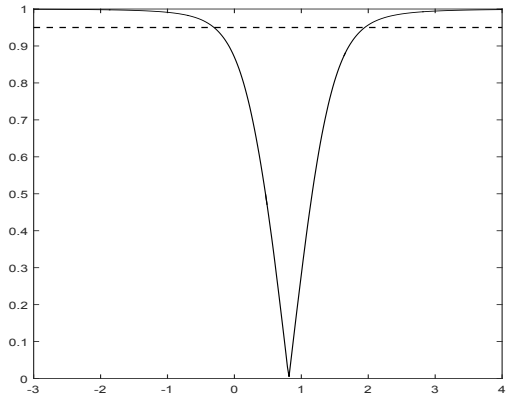

Figure 3: Confidence curve for the mean of the normal distribution based on a sample of size 8. The lowest point is the median of the CD that can be used as a point estimator and the interval between the two points where the dotted line intersects the CD is the 95% confidence interval.

2.2 Fiducial inspired Confidence Distribution Fusion

Suppose there are K labs and lab i measures the object n_i times, $i = 1, \ldots, K$, and reports the mean, X_i, of these n_i measurements. We assume that the data generating equation for these measurements is

$$X_i = \mu + B_i + \frac{\sigma_{A_i}}{\sqrt{n_i}} Z_i, \quad i = 1, \ldots, K \tag{1}$$

Here μ is the true value of the measurand, ${n_i}^{-1/2}\sigma_{A_i} Z_i$ are measurement errors (called type A error)assumed to have $N(0, \sigma^2_{A_i}/n_i)$ distribution and B_i are lab specific unknown systematic errors (called type B errors). The B_i cannot be measured directly. However it is assumed that each lab has some prior information available to asses its size. This prior information often differs significantly from lab to lab, so modeling it as a random effect with a common distribution across labs is not appropriate. For example, see Table 7 in Thalmann (2002).

Typically, B_i is modeled as a random variable with zero mean and known standard deviation

σ_{B_i} often referred to as type-B uncertainty. Hence, the variance of X_i, denoted as $\sigma^2_{C_i}$, is given by

$$\sigma^2_{C_i} = \sigma^2_{B_i} + \frac{\sigma^2_{A_i}}{n_i}.$$

The inferences for μ are performed separately by each lab and reported as the triple $\mathbf{y}_i = (x_i, u_i, d_i)$ where x_i is the realized value of X_i, u_i is an estimate of σ_{C_i} and d_i is an *effective degrees of freedom* associated with u_i. The quantity u_i is called the *combined standard uncertainty* (GUM, 1995). The value of d_i is generally determined using the Satterthwaite (1946) approximation for a linear combination of independent χ^2 random variables.

In particular, the estimate of the combined variance $\sigma^2_{C_i}$ is

$$u_i^2 = \sigma^2_{B_i} + \frac{s_i^2}{n_i}$$

where s_i is the sample standard deviation of the n_i observations from lab i whose mean is x_i. It is assumed that $d_i u_i^2/\sigma^2_{C_i}$ is distributed (approximately) as a χ^2 random variable with d_i degrees of freedom. Since we assume that the type B error has a known variance, the corresponding type B degrees of freedom $d_B = \infty$ and Satterthwaite (1946) approximation provides the following formula for effective degrees of freedom

$$d_i = (n_i - 1)\frac{u_i^4}{s_i^4/n_i^2}. \tag{2}$$

The labs therefore report what is essentially a conservative Confidence Distribution given by the location-scale t distribution with distribution function

$$H_i(\mu_i \mid \mathbf{y}_i) = F^t_{\lfloor d_i \rfloor}\left(\frac{\mu_i - x_i}{u_i}\right), \tag{3}$$

where $\lfloor d_i \rfloor$ is the floor of d_i; largest integer smaller than d_i. We take these lab reported CDs and their type A degrees of freedom $d_{A_i} = n_i - 1$ as a starting point for our model averaging. Trying to improve the lab reported CDs goes beyond the scope of this work and will be subject of future work.

Because the labs are measuring the same quantity, it is reasonable to assume that most, if not all, of the labs are actually providing unbiased estimates of μ. However, it is not uncommon for a handful of labs to provide discrepant results. This may be the consequence of incorrect adjustments by the labs to account for systematic errors or incorrect specification of σ_{B_i}. Our goal is to provide a combined confidence distribution for the common value μ that is robust to discrepant results. We first derive a combined CD assuming that all labs measure the same $\mu_i = \mu$. Then, in the next section, we will address the issue of discrepant labs using a model selection approach.

Hannig and Xie (2012) provide a simple formula based on Dempster's rule of recombination (Dempster, 2008; Shafer, 1976) and generalized fiducial distribution (Hannig *et al.*, 2016). The density of the combined CD for μ is

$$h(\mu \mid \mathbf{y}) = \frac{\sum_{i=1}^K \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{y}_i) \prod_{j \neq i} D_{\mathbf{y}_j} H_j(\mu \mid \mathbf{y}_j)}{\int_{-\infty}^{\infty} \sum_{i=1}^K \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{y}_i) \prod_{j \neq i} D_{\mathbf{y}_j} H_j(\bar{\mu} \mid \mathbf{y}_j)\, d\bar{\mu}} \tag{4}$$

where $D_{\mathbf{y}_j} H_j(\theta|\mathbf{y}_j) = \|\nabla_{\mathbf{y}_j} H_j(\theta|\mathbf{y}_j)\|_2$ is the Euclidean norm of the gradient of the $H_j(\theta|\mathbf{y}_j)$ computed with respect to the observed measurements $\mathbf{y}_j$.

From equation (3) we have $\frac{\partial}{\partial\mu} H_i(\mu \mid \mathbf{y}_i) = t_{\lfloor d_j \rfloor}\left(\frac{\mu - x_j}{u_j}\right)$, where $t_{\lfloor d_j \rfloor}(s)$ is the density of a Student's T distribution with $\lfloor d_j \rfloor$ degrees of freedom. When calculating $D_{\mathbf{y}_j} H_j(\mu|\mathbf{y}_j)$ we follow Hannig and Xie (2012) and interpret the gradient as a vector of partial derivatives with respect to each of the observations (measurements) rather than just the sufficient statistics. Thus

$$\nabla_{\mathbf{y}_j} H_j(\theta|\mathbf{y}_j) = t_{\lfloor d_j \rfloor}\left(\frac{\mu - x_j}{u_j}\right)\left(-\frac{1}{u_j}\nabla_{\mathbf{y}_j} x_j - \frac{\mu - x_j}{u_j^2}\nabla_{\mathbf{y}_j} u_j\right),$$

with

$$\nabla_{\mathbf{y}_j} x_j = \frac{1}{n_j}\mathbf{1}_{\mathbf{n_j}}, \quad \nabla_{\mathbf{y}_j} u_j = \frac{1}{2n_j u_j}\nabla_{\mathbf{y}_j} s_j^2 \quad \text{and} \quad \nabla_{\mathbf{y}_j} s_j^2 = \frac{2}{n_j - 1}(\mathbf{x}_j - x_j \mathbf{1}_{n_j}).$$

Here $\mathbf{x}_j$ denotes the vector of observations used to calculate the lab mean x_j and $\mathbf{1}_{n_j}$ is a vector of 1s. The Euclidean norm of the gradient is then easily computed and using (2) simplified to

$$D_{\mathbf{y}_j} H_j(\mu|\mathbf{y}_j) = t_{\lfloor d_j \rfloor}\left(\frac{\mu - x_j}{u_j}\right)\frac{1}{n_j^{1/2} u_j}\left(1 + \frac{(\mu - x_j)^2}{((n_j - 1)d_j)^{1/2} u_j^2}\right)^{1/2}. \tag{5}$$

To numerically compute the confidence interval based on the combined generalized fiducial distribution (4) we can use the following importance sampling algorithm (Robert and Casella, 2004, Section 3.3).

1. Generate $R_{i,l}$, a sample of size m from each of the generalized fiducial distribution $H_i(\mu|\mathbf{y}_i)$, using $R_{i,l} = x_i - u_i T_{i,l}$ where $T_{i,l}, l = 1, \ldots, m$ are independent Student's T random variables with d_i degrees of freedom.

2. For each $R_{i,l}$, $i = 1, \ldots K$, $l = 1 \ldots, m$, compute unnormalized weights $W_{i,l} = \prod_{j \neq i} D_{\mathbf{x}} H_j(R_{i,l}|\mathbf{y}_j)$ using (5).

3. Compute the importance sampling estimate of the distribution function of (4) by

$$\hat{H}(\mu \mid \mathbf{y}) = \frac{\sum_{i=1}^{K} m^{-1} \sum_{l=1}^{m} W_{i,l} I_{[R_{i,l},\infty)}(\mu)}{\sum_{i=1}^{K} m^{-1} \sum_{l=1}^{m} W_{i,l}}, \tag{6}$$

   where the indicator $I_{\lfloor R_{i,j},\infty)}(\mu) = 1$ if $R_{i,j} \leq \mu$ and $I_{[R_{i,j},\infty)}(\mu) = 0$ otherwise. To form approximate confidence intervals use the appropriate quantiles of $\hat{H}(\mu)$

Finally notice that the normalizing constant in (6) is an estimate of the normalizing constant in (4).

2.3 Model Selection

Let us now consider the situation where *most* of the labs are measuring the same correct value μ while each remaining lab is measuring some incorrect value. We are interested in making inferences about the true value μ without making any a priori assumptions about which labs are correct. There are $2^K - 1$ possible such models ranging from only a single lab measuring the true value to all the labs measuring the correct value.

Hannig and Lee (2009) have introduced model selection into the generalized fiducial paradigm. Their results have been used for a multivariate normal model by Wandler and Hannig (2011, 2012). The idea is to include the various models as a parameter in the setup of the problem and has been formalized in Theorem 3.1 of Hannig *et al.* (2016) where a formula for fiducial probability of each model is described.

Evaluating the fiducial probability for all models might be prohibitive in terms of computational cost even for moderate values of K. One could build an MCMC chain that could estimate these probabilities as has been done in Hannig and Lee (2009). In this paper we take a different route. Instead of estimating the fiducial probability of each model, we propose an extremely efficient algorithm that directly calculates the fiducial model averaged distribution for the common μ without explicitly estimating any of the model probabilities. Our idea is based on a simple mathematical observation. However, to our knowledge, this is the first time this computational trick has been used for model averaging.

In our situation we consider as a model $\mathbf{i} \subset \{1, \ldots, K\}$, with $i \in \mathbf{i}$ if lab i was measuring μ and $s \notin \mathbf{i}$ if lab s measured some value other than μ. For a fixed model $\mathbf{i}$, the joint fiducial density of the common mean μ and the discrepant means $\mu_s,\ s \notin \mathbf{i}$ is equal to

$$h_{\mathbf{i}}(\mu, \mu_s, s \notin \mathbf{i}|\mathbf{y}) = h_{\mathbf{i}}(\mathbf{y})^{-1} \left(\sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{y}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{y}_j} H_j(\mu \mid \mathbf{y}_j) \right) \prod_{s \notin \mathbf{i}} \frac{\partial}{\partial \mu} H_s(\mu_s \mid \mathbf{y}_s).$$

Notice that $\int_{-\infty}^{\infty} \frac{\partial}{\partial \mu} H_s(\mu_s \mid \mathbf{y}_s)\, d\mu_s = 1$ and therefore marginal density for the common parameter μ and model $\mathbf{i}$ is

$$h_{\mathbf{i}}(\mu|\mathbf{y}) = h_{\mathbf{i}}(\mathbf{y})^{-1} \left(\sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{y}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{y}_j} H_j(\mu \mid \mathbf{y}_j) \right) \tag{7}$$

where the normalizing constant is

$$h_{\mathbf{i}}(\mathbf{y}) = \int_{-\infty}^{\infty} \sum_{i \in \mathbf{i}} \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{y}_i) \prod_{j \in \mathbf{i},\ j \neq i} D_{\mathbf{y}_j} H_j(\bar{\mu} \mid \mathbf{y}_j)\, d\bar{\mu}. \tag{8}$$

Theorem 3.1 of Hannig *et al.* (2016) gives a generalized fiducial probability of each model as

$$h(\mathbf{i}|\mathbf{y}) = \frac{q^{K-|\mathbf{i}|+1} h_{\mathbf{i}}(\mathbf{y})}{\sum_{\mathbf{j} \in 2^{\{1,\ldots,K\}}} q^{K-|\mathbf{j}|+1} h_{\mathbf{j}}(\mathbf{y})},$$

where q is a penalty term to be specified below. The model averaged combined CD density for μ obtained by weighing models by their fiducial probabilities is given by

$$h(\mu \mid \mathbf{y}) = \sum_{\mathbf{j}\in 2^{\{1,\ldots,K\}}} h_{\mathbf{j}}(\mu|\mathbf{y}) h(\mathbf{j}|\mathbf{y}) = \frac{\sum_{\mathbf{j}\in 2^{\{1,\ldots,K\}}} q^{K-|\mathbf{j}|+1} h_{\mathbf{j}}(\mu|\mathbf{y}) h_{\mathbf{j}}(\mathbf{y})}{\sum_{\mathbf{j}\in 2^{\{1,\ldots,K\}}} q^{K-|\mathbf{j}|+1} h_{\mathbf{j}}(\mathbf{y})}. \tag{9}$$

The sum above is over $2^K - 1$ summands which could be prohibitively large even for medium values of K. Instead of implementing (9) using an MCMC we instead insert the formulas from (7) and (8) into (9), rearrange the terms and combine them into a product. After some algebra we get the following computationally friendly version of the model averaged combined CD density

$$h(\mu \mid \mathbf{y}) = \frac{\sum_{i=1}^{K} \frac{\partial}{\partial \mu} H_i(\mu \mid \mathbf{y}_i) \prod_{j\neq i} \left(1 + q^{-1} D_{\mathbf{y}_j} H_j(\mu \mid \mathbf{y}_j)\right)}{\int_{-\infty}^{\infty} \sum_{i=1}^{K} \frac{\partial}{\partial \mu} H_i(\bar{\mu} \mid \mathbf{y}_i) \prod_{j\neq i} \left(1 + q^{-1} D_{\mathbf{y}_j} H_j(\bar{\mu} \mid \mathbf{y}_j)\right) \, d\bar{\mu}}. \tag{10}$$

Notice that the numerator of (10) can be computed in K^2 operations; it is a sum of K terms that each are a product of K numbers.

Based on (10) we propose the following importance sampling algorithm that is usable for practical computations:

1. Generate $R_{i,l}$, a sample of size m from each of the generalized fiducial distribution $H_i(\mu|\mathbf{y}_i)$, using $R_{i,l} = x_i - u_i T_{i,l}$ where $T_{i,l}, l = 1, \ldots, m$ are independent Student's T random variables with d_i degrees of freedom.

2. For each $R_{i,l}$, $i = 1, \ldots K$, $l = 1 \ldots, m$, compute unnormalized weights

$$\tilde{W}_{i,l} = \prod_{j\neq i} \left[1 + D_{\mathbf{y}_j} H_j(R_{i,l}|\mathbf{y}_j)\, q^{-1}\right],$$

   where $D_{\mathbf{y}_j} H_j$ is given in (5) and q is in (11).

3. Compute the importance sampling estimate of the distribution function of (10) by

$$\hat{H}(\mu \mid \mathbf{y}) = \frac{\sum_{i=1}^{K} \sum_{l=1}^{M} \tilde{W}_{i,l} I_{[R_{i,l},\infty)}(\mu)}{\sum_{i=1}^{K} \sum_{j=1}^{M} \tilde{W}_{i,l}}.$$

   To form approximate confidence intervals use the appropriate quantiles of $\hat{H}(\mu|\mathbf{y})$.

The penalty term q is required to offset the propensity of the generalized fiducial distribution to select models with larger number of parameters. Selection of the right penalty is somewhat of an art very similar to selection a prior probability to each model. It is our experience that using Minimum Description Length principle (Lee, 2001) and adjusting the MDL penalty by a multiplicative term to make the procedure scale invariant often leads to good repeating sampling performance, see Hannig *et al.* (2016); Wandler and Hannig (2011). Therefore, we propose to use the following penalty

$$q = \text{MSE} \, \left(\sum_{i=1}^{K} u_i^{-2}\right)^{-1/2} \left(\sum_{i=1}^{K} n_i\right)^{-1/2} \tag{11}$$

where the type A mean square error MSE $= K^{-1}\sum_{i=1}^{K} n_i u_i^2 \sqrt{(n_i-1)/d_i}$.

Remark 1. The combined confidence distribution in (10) treats all the labs equally. However in some situations we want to combine results that are similar to a preferred reference lab. This is achieved by making sure that this lab is included in all the models considered. If the preferred lab is lab r, this exhibits itself in (10) and the corresponding part of the importance sampling algorithm by replacing "1+" in the formula with "$I_{\{j\neq r\}}$+".

3 Simulation Study

To demonstrate the small sample performance of our algorithm proposed in Section 2.3, we conducted a simulation study consisting of measurements from 7 labs generated from each of three different scenarios listed below.

- *Scenario 0*: All 7 labs provide unbiased estimates of the true value. We take $\mu_i = 45,\ i = 1,\ldots,7$ for concreteness.
- *Scenario 1*: Six labs provide unbiased estimates of the true value $\mu_i = 45,\ i = 1,\ldots,6$ and while one lab provides a biased estimate whose expectation is $\mu_7 = 48$. This mimics the situation where one lab may incorrectly estimate the lab bias B_k and/or the standard deviation of lab bias σ_{B_k}.
- *Scenario 2*: Two clusters of labs. Labs in one cluster of size 4 make measurements with expected value equal to 45 (that is $\mu_i = 45,\ i = 1,\ldots,4$) and labs in the other cluster of size 3 make measurements with expected value equal to 48 (that is, $\mu_i = 48,\ i = 5,\ldots,7$). This setting simulates the situation where labs use fundamentally different methods for measurement and it is impossible to know which of the labs, if any, are providing unbiased estimates of the true value μ. Thus, there is no answer to which value is the truth.

For each scenario, we assume each lab makes the same number of measurements n_i and thus same type A degrees of freedom $d_{A_i} = n_i - 1$. Two values $n_i = 5, 15$ are used in the simulation study. To model the heterogeneity among the labs, different standard deviations of type A error and type B error, σ_{A_i} and σ_{B_i} respectively, are generated from a Gamma distribution for each lab, i.e.

$$\sigma_{A_i} \sim \Gamma\left(n_i, \frac{1}{n_i}\right), \quad \sigma_{B_i} \sim \Gamma\left(n_i, \frac{R}{n_i}\right)$$

in which R is the ratio of the mean of σ_{B_i}'s over the mean of σ_{A_i}'s. Four different ratios are considered ($R = 0,\ 1/3,\ 1,\ 2$) for generating the data sets. Note that $R = 0$ implies type B error is not present. For each collection of σ_{B_i}'s, the type B errors B_i are independently simulated, one per lab, from normal distributions $N(0, \sigma_{B_i}^2)$.

One hundred parameter sets of $\{\sigma_{A_i}, \sigma_{B_i}, B_i, i = 1, \cdots, 7\}$ are simulated for each combination of n_i and R. For each fixed parameter set, 1000 repetitions of the laboratory measurements, type A

and combined standard errors are generated using

$$X_i = \mu_i + B_i + \frac{\sigma_{A_i}}{\sqrt{n_i}} Z_i, \quad s_i = \sigma_{A_i} \sqrt{\frac{W_i}{(n_i - 1)}}, \quad u_i = \sqrt{\frac{s_i^2}{n_i} + \sigma_{B_i}^2},$$

where $Z_i \sim N(0, 1)$ and $W_i \sim \chi^2_{n_i - 1}$ are independent. The true parameter values $\mu_i,\ i = 1, \cdots, 7$ are set based on different scenarios.

Notice that the fact that the value of type B error B_i is fixed within each of the 1000 repetitions leads to a behavior of the empirical coverage that is different from our usual experience. When R is close to 0 (type B error is negligible) the empirical coverage computed based on the 1000 repetitions should be close to the stated value for each of the 100 parameter sets. On the other hand, when R is large (type B error dominates) the empirical coverage will be close to either 100% or 0% depending on the value of B_i sampled. This is because the differences between the 1000 datasets are negligible compared to the fixed value of B_i. Thus for R large, one should also evaluate whether the average of empirical coverages computed across the 100 parameter sets is at or above the nominal coverage value.

We compared the proposed method to two classical methods that are most commonly used for calculating a consensus value in metrology (GUM, 1995; Rukhin, 2009): the arithmetic mean and the variance weighted mean. The arithmetic mean $\bar{x}_A$ and its estimated standard error are

$$\bar{x}_A = \frac{\sum_{i=1}^K x_i}{K}, \quad \hat{\sigma}_{\bar{x}_A} = \frac{\sqrt{\sum_{i=1}^K u_i^2}}{K}.$$

The weighted mean $\bar{x}_W$ and its estimated standard error are

$$\bar{x}_W = \frac{\sum_{i=1}^K w_i x_i}{\sum_{i=1}^K w_i}, \quad \hat{\sigma}_{\bar{x}_W} = \left(\sum_{i=1}^K w_i \right)^{-1/2}, \quad \text{where} \quad w_i = u_i^{-2}.$$

Details for each scenario are discussed in Section 3.1, Section 3.2 and Section 3.3, respectively.

3.1 Scenario 0

The expected values for the measurements by the 7 labs are all equal to $\mu_i = 45,\ i = 1, \ldots, 7$. For illustration, Figure 4 provides an example of the fiducial distribution of the consensus value for one of the datasets generated for type A degrees of freedom $d_{A_i} = 4$ and $R = 0$. The blue curves in the left panel are kernel density estimates, i.e. smoothed histograms, of the fiducial samples for each lab. It can be seen that the expected result for each lab deviates slightly from the true value of 45 with different amounts of dispersion. The top black kernel density estimator shows that the center of the consensus value distribution is around the true value 45. The top black confidence curve in the right panel depicts the median estimate as 44.9 and 95% fiducial confidence interval as [44.4, 45.4] which successfully covers the truth.

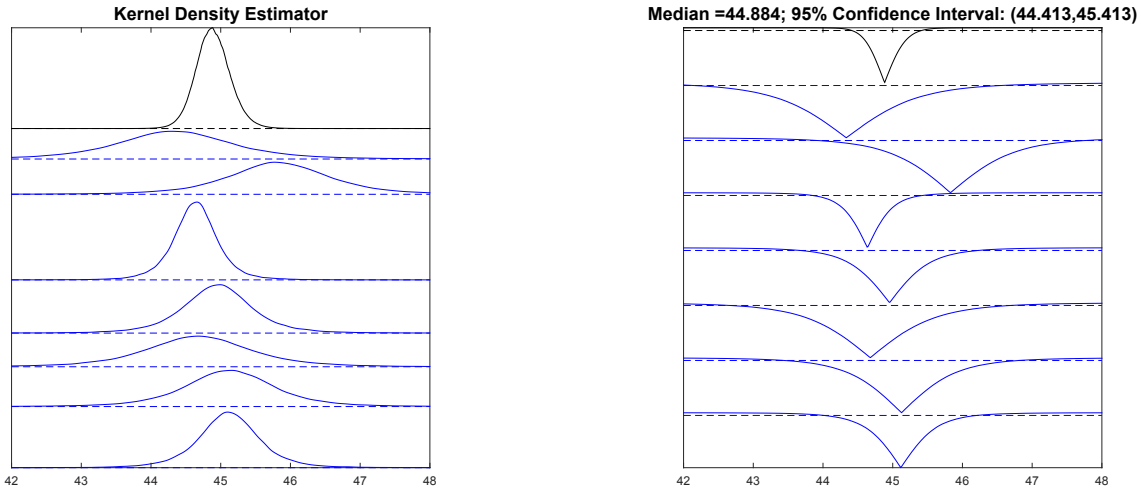

Figure 4: Fiducial estimate of one simulated data with σ_{A_i}, σ_{B_i}, B_i generated under Scenario 0 with $d_{A_i} = 4$ and $R = 0$. The top left black curve shows the kernel density estimate for the consensus value with mode around 45. The top right black curve shows that the confidence curve covers the true value 45 at 95% confidence level.

For each of the 100 parameter sets, we compute the coverage and lengths of the 95% confidence intervals based on the 1000 simulated data sets. Box-plots shown in Figure 5 summarize the results for the 100 parameter sets under different ratios R and type A degrees of freedom $d_{A_i} = n_i - 1$. The blue boxes display the coverages for fiducial method, while the green and yellow boxes, respectively, show the coverages of arithmetic mean and weighted mean. Results are grouped by ratios for $d_{A_i} = 4$ (left) and $d_{A_i} = 14$ (right). The average coverages are given underneath each box. When only type A error is present ($R = 0$), the coverage of fiducial estimates and arithmetic mean are around 95%, while the weighted mean has a much lower coverage, especially when $d_{A_i} = 4$ (with median coverage being around 80%). When type B error exists and increases (larger R), all three methods tend to get 100% coverage. However, the arithmetic mean and the weighted mean are less robust in the sense that they might get 0 coverage for certain parameter sets.

Additionally, we compute the average length of 95% confidence intervals of 1000 simulated data sets for comparing the three methods. As before, Figure 6 displays the box-plots of fiducial method (blue), arithmetic mean (green) and weighted mean (yellow) for different choices of type A degrees of freedom and R. The confidence intervals gets wider with an increase in the ratio R and gets shorter when the degree of freedom increases for all three methods. In general, the fiducial intervals are wider than the others which is consistent with the coverage comparison in Figure 5.

3.2 Scenario 1

In this scenario, we try to mimic the consensus value estimation with a single apparently discrepant lab, i.e. $\mu_i = 45,\ i = 1, \ldots, 6$ and $\mu_7 = 48$. Again, Figure 7 provides an illustration of the GFD of

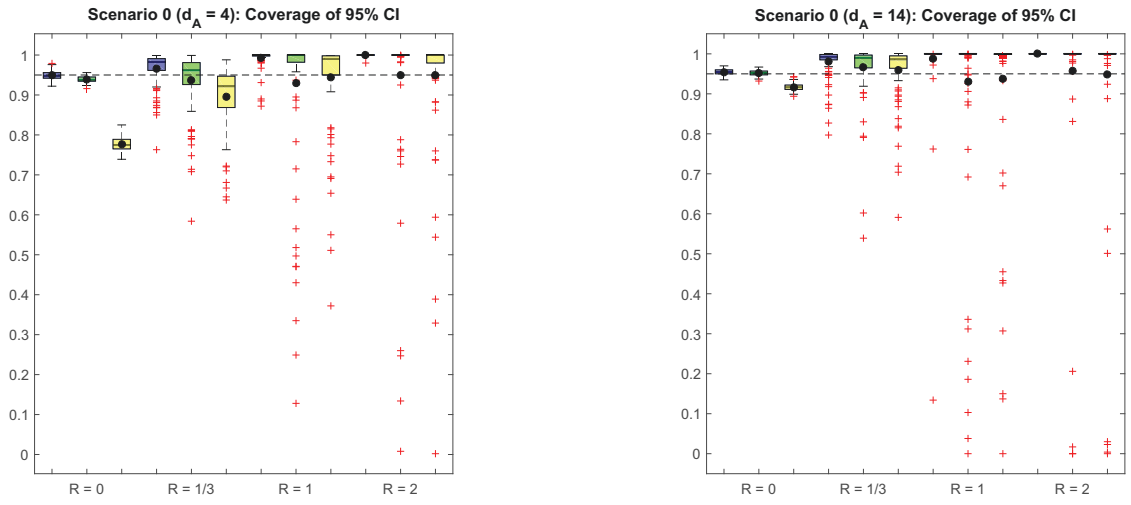

Figure 5: Coverage Comparison for Scenario 0 grouped by ratios R for $d_{A_i} = 4$ (left) and $d_{A_i} = 14$ (right). For each of the ratios are the boxplots are ordered left to right: fiducial method (the blue boxes), arithmetic (green boxes) and weighted mean (yellow boxes). The average of the coverages for each boxplot is indicated by a bold dot.

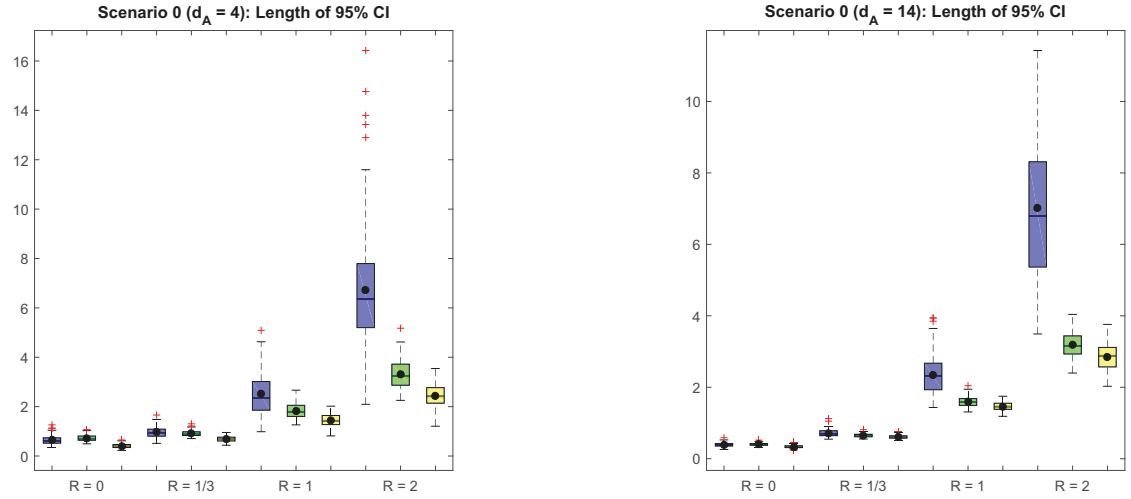

Figure 6: 95% CI length comparison, under Scenario 0, for $d_{A_i} = 4$ (left) and $d_{A_i} = 14$ (right) with different ratios R. For each R the boxplots are in the following order: fiducial method (blue), arithmetic mean (green), weighted mean (yellow). The average of lengths in each boxplot is indicated by a bold dot. The CI gets wider with the increase in the ratio and gets narrower with the increase in degree of freedom for all three methods.

the consensus value for one simulated data set with $d_{A_i} = 4$and $R = 0$. The bottom blue curves in both panels indicates the presence of a discrepant lab. The black curves on the top show that the consensus value estimate from the fiducial approach is around 45 and appears to be uninfluenced by the apparently discrepant lab. The 95% confidence interval is [44.40, 45.18] which covers the true value of the six *non-discrepant* labs. It can be seen that the fiducial consensus estimate is robust against an outlier measurement from the apparently discrepant lab.

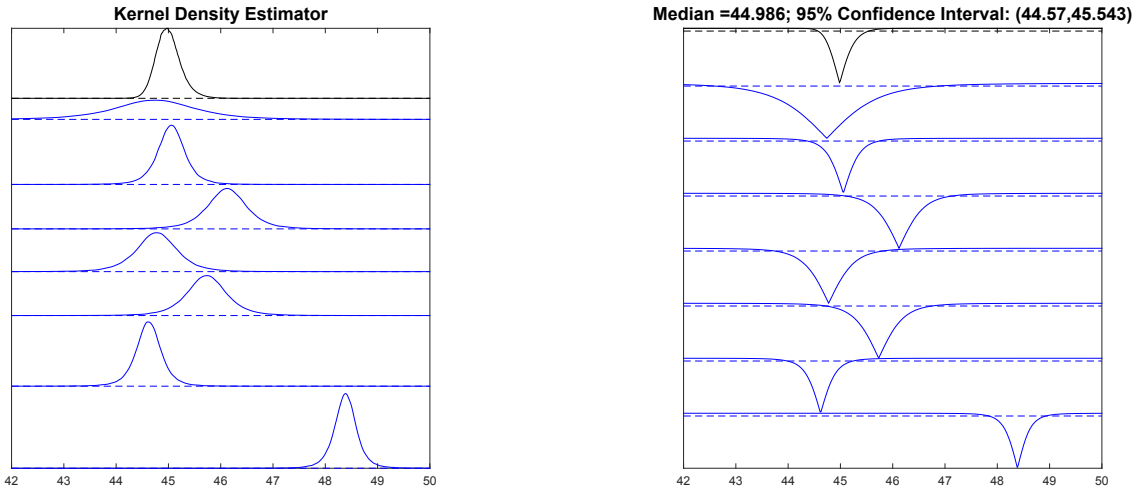

Figure 7: One simulated data with σ_{A_i} and σ_{B_i} generated under Scenario 1 with $d_{A_i} = 4$ and $R = 0$. The kernel density estimates (left) indicate an apparently discrepant measurement from the last lab. The black kernel density estimate and confidence curves for the consensus value demonstrate the robustness of our proposed method against discrepant measurements.

We similarly compute the coverage of the 95% confidence interval for the true value $\mu = 45$. Box-plots for different choices of type A degrees of freedom and R are shown in Figure 8. The fiducial method stays robust against the discrepant lab measurement and obtains similar coverages as Scenario 0. Both arithmetic mean and weighted mean are adversly influenced by the discrepant lab. When $d_{A_i} = 4$, the coverages are only around 40% with no type B error or a small ratio of type B error. Differing from Scenario 0, the coverages get worse with an increase in degree of freedom since the evidence of the outlier lab gets stronger. The median coverages even drop to near 0 when $R = 0$ and $d_{A_i} = 14$. When type B error dominates, both arithmetic and weighted mean are unstable with the average coverage still well bellow the target 95%.

3.3 Scenario 2

We consider two clusters of labs in this scenario for simulating the situation where labs might use different measuring methods. Recall that the true value of 4 labs is equal to 45 and the true value of the other 3 labs is equal to 48. In this situation, it is not clear which of the two, 45 or 48, should be the consensus value. This is illustrated in Figure 9 where GFD of the consensus value for two

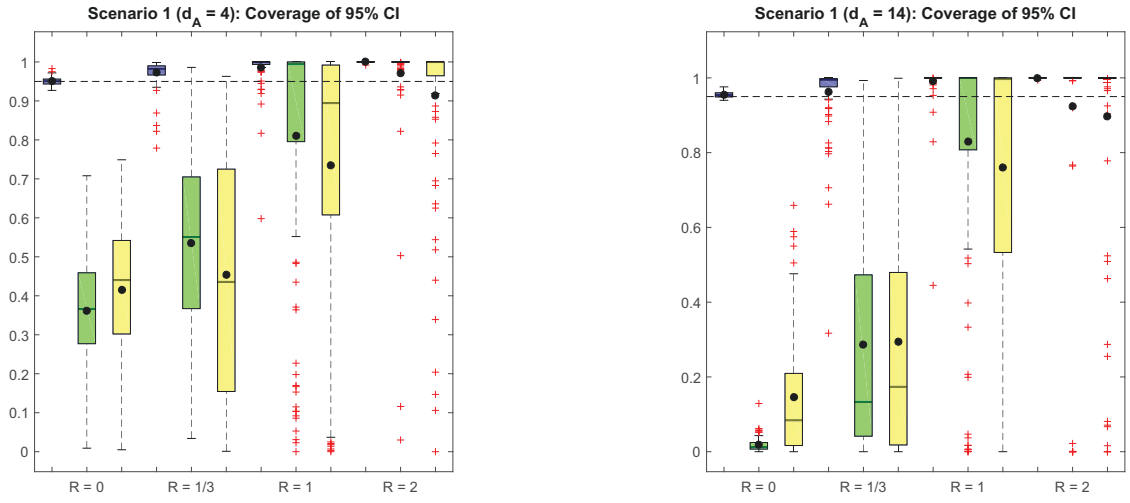

Figure 8: Coverage Comparison for Scenario 1: For each ratio R the methods are in the following order: fiducial estimate (blue), arithmetic mean (green), weighted mean (yellow). The average of the coverages for each boxplot are indicated by a bold dot. Fiducial estimate is robust against the apparent discrepancy of one of the labs, while the other methods are strongly influenced, especially in the case without type B error ($R = 0$).

simulated data sets generated with $d_{A_i} = 4$ and $R = 0$ are shown. The first four blue curves are centered around 45 representing the first cluster and the last three blue curves are centered around 48 from the second cluster. In the first row, the top black density estimate curve suggests that the estimate of the consensus value is about 45 which is dominated by the cluster whose true value is 45. A small peak can be seen around 48 indicating the impact from the other cluster. Besides, the black confidence curve shows the 95% confidence interval is [44.6, 47.9] which stretches towards the true value of the other cluster of labs. The second row shows an example of a situation where the value of 48 dominates. One of the labs in the second cluster has much smaller uncertainty compared with the other labs. This is enough to move the mode of the fiducial distribution of the consensus value to 48. The confidence curve suggests the 95% confidence interval is [44.6, 48.4], successfully covering both true values.

The assessment of coverage is tricky in the current situation since there is no single correct value. We evaluate two different coverage probabilities – (a) probability that at least one of the two values (45 or 48) will be covered, and (b) the probability of covering both 45 and 48. Results are shown in Figure 10 and Figure 11. The fiducial confidence intervals (blue boxes) cover at least one of the two values nearly 100% of the time under the different parameter settings. However, both arithmetic mean and weighted mean fail to capture any of the true values when ratio $R = 0$ and $R = 1/3$.

When it comes to simultaneous coverage of both values, 75% of the fiducial confidence intervals have a coverage around or above 60% for $R = 0$ and $R = 1/3$. This should not be surprising because our method was designed to capture the most dominant value, not both values. The other two methods are unable to simultaneously cover both of the true values for any of the cases.

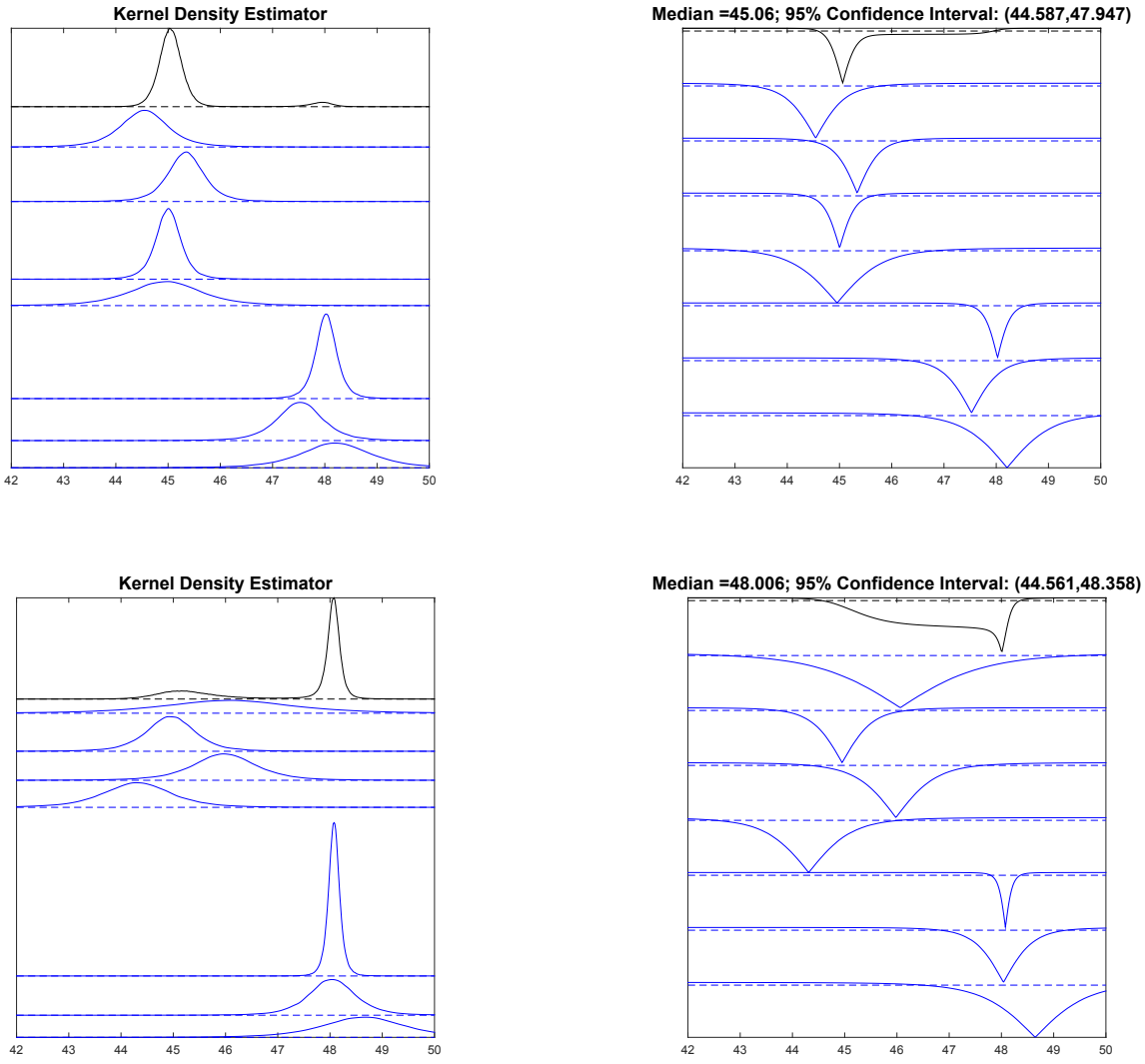

Figure 9: Two simulated data sets with σ_{A_i} and σ_{B_i} generated under Scenario 2 with $d_{A_i} = 4$ and $R = 0$. The first example shows that the first cluster with 4 labs dominate the consensus value estimation. The second example contains one lab with low estimated uncertainty in the cluster with true value being 48. Therefore, the consensus estimate is shifted towards 48.

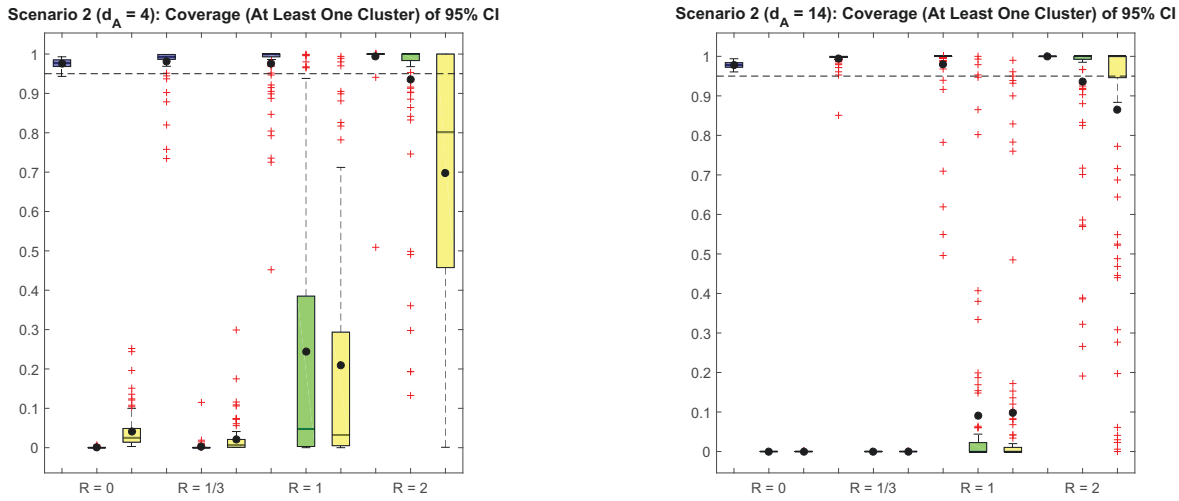

Figure 10: Coverage Comparison (At Least One Cluster) for Scenario 2 for different ratios and degrees of freedom. For each ratio R the methods are in the following order: fiducial estimate (blue), arithmetic mean (green), weighted mean (yellow). The average of the coverages for each boxplot is indicated by a bold dot. Notice that fiducial estimate is reliable at capturing at least one cluster while the other two methods are not.

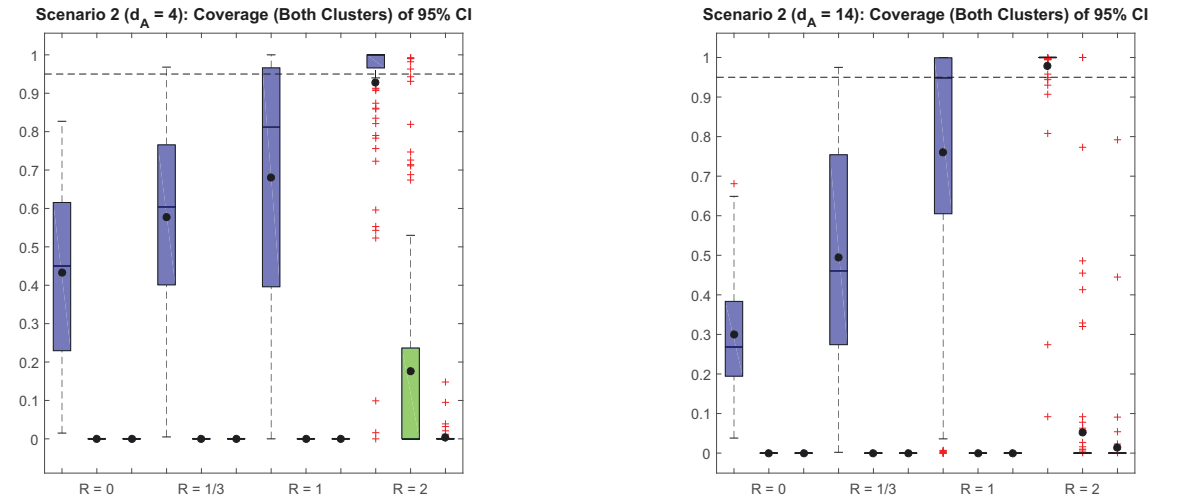

Figure 11: Coverage Comparison (Both Clusters) for Scenario 2 for different ratios and degrees of freedom. For each ratio R the methods are in the following order: fiducial estimate (blue), arithmetic mean (green), weighted mean (yellow). The average of the coverages for each boxplot is indicated by a bold dot. We see that even though the fiducial estimate is only designed to capture the dominant cluster it often captures both. The two methods almost never capture both clusters.

3.4 Discussion of Simulation Results

Below is a brief summary of some main observations from the simulation study.

- In Scenario 0 and Scenario 1 the fiducial intervals cover the true μ with confidence level greater than or equal to the nominal value of 95%. For Scenario 2 the coverage probability for covering at least one of the two cluster means is greater than or equal to the nominal value.
- For arithmetic mean and weighted mean approaches, the coverage is nowhere near nominal in most situations examined. For scenarios 1 and 2 the coverage is particularly bad. These methods are unsuitable for situations when there may be discrepant labs.
- Although the arithmetic mean and the weighted mean provide 95% confidence intervals of shorter expected length this is meaningless since their coverages are highly inadequate.

4 Data examples

4.1 Steel Gauge Blocks

In order to establish the metrological equivalence of national measurement standards and of calibration certificates issued by national metrology institutes a set of key comparisons are chosen and organized by the Consultative Committees of the CIPM or by the regional metrology organizations in collaboration with the Consultative Committees (Thalmann, 2002). In September 1997, the Consultative Committee for Length, CCL, decided upon a key comparison on gauge block measurements by interferometry, named CCL-K1, starting in spring 1998, with the Swiss Federal Office of Metrology (OFMET) as the pilot laboratory. The results of this international comparison contribute to the mutual recognition arrangement (MRA) between the national metrology institutes of the Metre Convention.

Ten gauge blocks of steel and 10 gauge blocks of tungsten carbide, of varying nominal lengths, were circulated to 11 different NMIs. For the purpose of illustration we considered one particular set of gauge block measurements corresponding to the nominal value of 8 mm. The results along with their associated uncertainties are shown in Figure 2 and Table 1. What is actually reported by each participating lab is the *deviation (in nm)* of the measured length from the nominal value.

The published reports Thalmann (2002) did not clearly spell out the degrees of freedom. In order to apply the proposed method we selected the effective degrees of freedom $d_i = 60$, corresponding to the degrees of freedom needed to get a critical value of 2 which is the typical multiplier used in the metrology literature for constructing confidence intervals. The type B to type A standard error ratio is typically around 1.5 in these problems and (2) gives the corresponding type A degrees of freedom as $d_{A_i} = n_i - 1 = 5$. Figure 12 presents the estimates of kernel density curve (left) and confidence curves (right). The 95% confidence interval is $[-31.6,\ 37.3]$ nm with the median estimate being 4.08 nm. The consensus value estimate mainly picks up the measurements of the labs with mode around 0. The confidence interval takes the uncertainty caused by two discrepant labs into consideration.

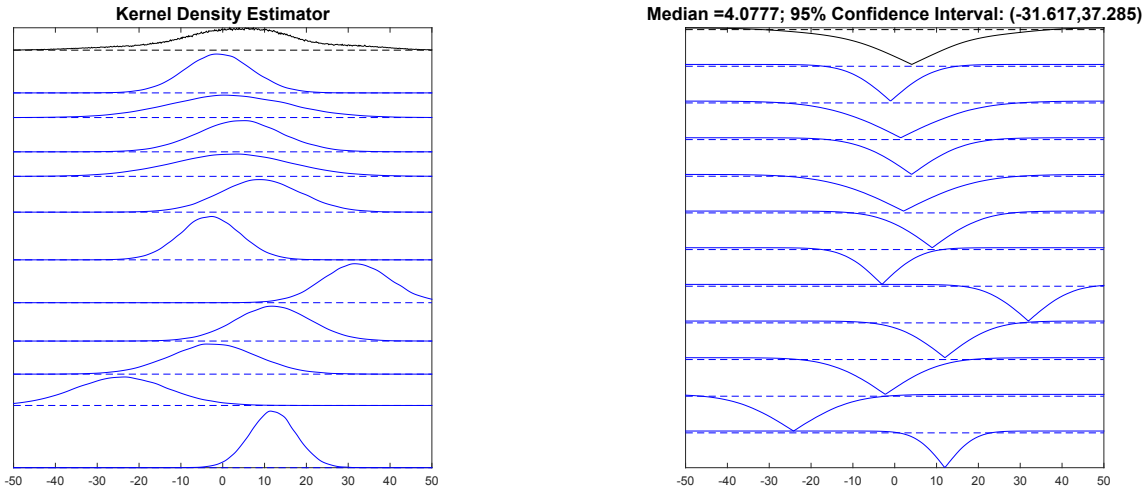

Figure 12: Results of CCL data sets with effective degrees of freedom equal to 60 and type A degrees of freedom equal to 5.

The arithmetic mean -0.2 ± 3.5 nm and the weighted mean 0.1 ± 3.2 nm are given in Thalmann (2002) as the reference values. These results exclude the values of VNIIM and NIM based on the decision of the CCL Working Group Dimensional Metrology (WGDM). Hence their confidence intervals are narrower as VNIIM is the one most different from the others.

4.2 Newton's Constant of Gravitation, G

Newton's constant of gravitation G is a key constant that is needed for much fundamental research in physics. Many advanced scientific labs measure G and report a value and an uncertainty. The data set contains the values from 11 labs shown in Table 2. See Mohr *et al.* (2012) for details.

It turns out that the confidence interval for G from some labs exclude values from other labs, so there is some inconsistency. This is perhaps due to severe underestimation by some or all the labs of uncertainties in their results. The community seeks a consensus value that uses all available information. We applied the proposed method and obtained an estimate depicted in Figure 13 computed using the default degrees of freedom values of $d_i = 60$ and $d_{A_i} = 5$.

The blue curves show that two labs, with small uncertainties, perhaps coincidently, have nearly the same mode around $6.674 \times 10^{-11}\text{m}^3\text{kg}^{-1}\text{s}^{-2}$. Besides, there are several labs whose results are near this value with varying levels of uncertainties. The consensus estimate is therefore pulled towards this number with 95% confidence interval being $[6.6740,\ 6.6743] \times 10^{-11}\text{m}^3\text{kg}^{-1}\text{s}^{-2}$.

The value of G given by Mohr *et al.* (2012) is $6.67384 \times 10^{-11}\text{m}^3\text{kg}^{-1}\text{s}^{-2}$. The uncertainty is $0.00080 \times 10^{-11}\text{m}^3\text{kg}^{-1}\text{s}^{-2}$ which is the standard error of the weighted mean of the 11 values in Table 2 multiplied by the factor of 14. This multiplication factor 14 is chosen ad hoc in order to cover all the 11 values of G as none of them has an apparent issue besides the disagreement.

Organization	X_i: Result	u_i: Combined Standard Uncertainty
NIST-82	6.67248	0.00043
TR&D-96	6.6729	0.00050
LANL-97	6.67398	0.00070
UWash-00	6.674255	0.000092
BIPM-01	6.67559	0.00027
UWup-02	6.67422	0.00098
MSL-03	6.67387	0.00027
HUST-05	6.67228	0.00087
UZur-06	6.67425	0.00012
HUST-09	6.67349	0.00018
JILA-10	6.67234	0.00014

Table 2: Summary of the results of measurements of the Newton's constant of gravitation G. The units are $10^{-11}\mathrm{m}^3\mathrm{kg}^{-1}\mathrm{s}^{-2}$.

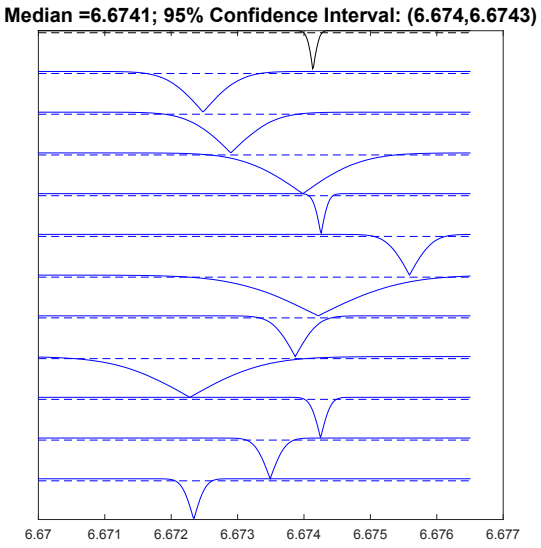

Figure 13: Results of Big-G data sets with effective degrees of freedom equal to 60 and type A degrees of freedom equal to 5.

5 Summary

Many methods have been proposed for obtaining a consensus from inter-laboratory trials. See Rukhin (2009) for a discussion. However, these methods do not adequately address robustness issues when addressing the discrepancy among inter-laboratory results. In this paper we introduce a new method based on generalized fiducial inference to obtain a robust consensus. The proposed new automatic and computationally efficient fiducial model averaging algorithm achieves a self-weighting capability, without the need to eliminate discrepant CDs. The three simulation experiments show that

our proposed method has better small sample properties than the commonly used methods, e.g., the arithmetic mean and weighted mean. Applications to gauge block calibration inter-laboratory trials and measurements of Newton's constant of gravitation (G) also display the robustness of our method since the reference value given in the reports either comes from the elimination of discrepant laboratory values or from the multiplication of the uncertainty by an arbitrary factor.

6 Acknowledgements

The authors are also most grateful to all the reviewers and the guest editors for their constructive and insightful comments and suggestions.

References

Bender, R., Berg, G. and Zeeb, H. (2005) Tutorial: Using Confidence Curves in Medical Research. *Biometrical Journal*, **47**, 237–247.

Birnbaum, A. (1961) Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of American Statistical Association*, **56**, 246–249.

CIPM (1999) Text of the CIPM MRA. *http://www.bipm.org/en/cipm-mra/cipm-mra-text/*.

Dempster, A. P. (2008) The Dempster-Shafer Calculus for Statisticians. *International Journal of Approximate Reasoning*, **48**, 365–377.

GUM (1995) *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization (ISO), Geneva, Switzerland.

Hannig, J., Iyer, H. K., Lai, R. C. S. and Lee, T. C. M. (2016) Generalized Fiducial Inference: A Review and New Results. *Journal of the American Statistical Association*, **111**, 1346–1361. Doi:10.1080/01621459.2016.1165102.

Hannig, J. and Lee, T. C. M. (2009) Generalized Fiducial Inference for Wavelet Regression. *Biometrika*, **96**, 847–860.

Hannig, J. and Xie, M. (2012) A note on Dempster-Shafer Recombinations of Confidence Distributions. *Electronic Journal of Statistics*, **6**.

Iyer, H. K., Wang, C. and Vecchia, D. (2004a) Consistency tests for key comparison data. *Metrologia*, **41**, 223.

Iyer, H. K., Wang, C. M. J. and Mathew, T. (2004b) Models and Confidence Intervals for True Values in Interlaboratory Trials. *Journal of the American Statistical Association*, **99**, 1060–1071.

Lee, T. C. M. (2001) An introduction to coding theory and the two–part minimum description length principle. *International Statistical Review*, **69**, 169–183.

Mohr, P., Taylor, B. and Newell, D. (2012) Codata recommended values of the fundamental physical constants: 2010. *Reviews of Modern Physics*, **84**, 1527–1605.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York: Springer-Verlag, second edn.

Rukhin, A. L. (2009) Weighted means statistics in interlaboratory studies. *Metrologia*, **46**, 323.

Satterthwaite, F. E. (1946) An approximate distribution of estimates of variance components. *Biometrics bulletin*, 110–114.

Schweder, T. and Hjort, N. L. (2002) Confidence and Likelihood. *Scandinavian Journal of Statistics. Theory and Applications*, **29**, 309–332.

Shafer, G. (1976) *A mathematical theory of evidence*. Princeton, New Jersey: Princeton University Press.

Singh, K., Xie, M. and Strawderman, W. E. (2005) Combining Information from Independent Sources Through Confidence Distributions. *The Annals of Statistics*, **33**, 159–183.

Thalmann, R. (2002) Ccl key comparison: calibration of gauge blocks by interferometry. *Metrologia*, **39**, 165.

Wandler, D. V. and Hannig, J. (2011) Fiducial Inference on the Maximum Mean of a Multivariate Normal Distribution. *Journal of Multivariate Analysis*, **102**, 87–104.

— (2012) A Fiducial Approach to Multiple Comparisons. *Journal of Statistical Planning and Inference*, **142**, 878–895.