XXI IMEKO World Congress "Measurement in Research and Industry"
August 30 - September 4, 2015, Prague, Czech Republic

ANALYSIS OF INTERLABORATORY COMPARISON DATA USING T-DISTRIBUTIONS

Alistair B Forbes [1]

[1] National Physical Laboratory, Teddington, UK, alistair.forbes@npl.co.uk

Abstract - This paper discusses the analysis of interlaboratory comparison data where there may be some doubt about the validity of the uncertainty statements. The analysis of repeated measurements drawn from a Gaussian distribution with unknown variance provides a model for the posterior adjustment of the input distribution associated with each participant. Importantly, the shapes of the input distributions are adjusted, not just their standard deviations.

Keywords: Bayesian inference, interlaboratory comparisons, t-distributions, uncertainty evaluation

1. INTRODUCTION

One of the difficult problems in analysing interlaboratory comparison (ILC) data [4] is what to do when the data is judged to be inconsistent. There are a large number of schemes for adjusting the input uncertainties to bring about consistency, e.g., [2, 5, 6, 3, 7, 8, 10, 12]. They generally start from a premise that some or all of the stated uncertainties are unreliable to a degree and that the information available from all the participants in the ILC can be used to determine which input uncertainties need to be adjusted and by how much. This paper uses the analysis of repeated measurements drawn from a Gaussian distribution with unknown variance to provide a model for the posterior adjustment of the distribution associated with each participant, on the basis of the information provided by all the participants. Importantly, the shapes of the input distributions are adjusted, not just their standard deviations.

2. ANALYSIS OF ILC DATA DERIVED FROM REPEATED MEASUREMENTS

While the GUM framework [1] is based on propagating known uncertainties associated with influence quantities through to the measurand using an input-output model, the analysis of repeated data is one situation where the standard deviation of the sampling distribution is unknown, a priori, and is estimated on the basis of the observed data. The fact that the standard deviation is estimated leads to the assignment of a student t-distribution to the estimate of the quantity derived from the sample mean. In a Bayesian framework [9], the analysis is as follows.

Suppose observations $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$, arise in according to the model $y|\alpha \sim \mathrm{N}(\alpha, \phi^{-1})$ where $\phi = 1/\sigma^2$ is unknown. Assuming a noninformative prior distribution $p(\alpha, \phi) \propto 1/\phi$, the posterior distribution $p(\alpha, \phi|\mathbf{y})$ is such that

$$p(\alpha, \phi|\mathbf{y}) \propto \phi^{m/2-1} \exp\left\{-\frac{\phi}{2}\left[(m-1)s^2 + m(\bar{y}-\alpha)^2\right]\right\},$$

where

$$\bar{y} = \frac{1}{m}y_i, \quad s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \bar{y})^2.$$

The marginalised distributions $p(\alpha|\mathbf{y})$ and $p(\phi|\mathbf{y})$ can be determined analytically. The posterior distribution $\alpha|\mathbf{y}$ is the t-distribution $t_{m-1}(\bar{y}, s^2/m)$ with

$$\begin{aligned} p(\alpha|\mathbf{y}) &\propto \left[(m-1)s^2 + m(\bar{y}-\alpha)^2\right]^{-m/2}, \\ &\propto \left[1 + \frac{1}{m-1}\frac{m}{s^2}(\bar{y}-\alpha)^2\right]^{-m/2}, \end{aligned}$$

while the posterior distribution associated with $\phi|\mathbf{y}$ is the Gamma distribution $G((m-1)/2, (m-1)s^2/2)$ with

$$p(\phi|\mathbf{y}) \propto \phi^{(m-1)/2-1} \exp\left\{-\frac{\phi}{2}(m-1)s^2\right\}.$$

For $m > 3$, the standard deviation of $t_{m-1}(\bar{y}, s^2/m)$ is

$$\left(\frac{m-1}{m-3}\right)^{1/2} \frac{s}{m^{1/2}}.$$

If information about α is available from another independent source, e.g., through measurements $\mathbf{y}_0$ with (marginalised) likelihood $p(\mathbf{y}_0|\alpha)$ gathered using another instrument, then knowledge about α and ϕ can be updated according to

$$p(\alpha, \phi|\mathbf{y}, \mathbf{y}_0) \propto p(\mathbf{y}_0|\alpha, \phi)p(\alpha, \phi|\mathbf{y}) = p(\mathbf{y}_0|\alpha)p(\alpha, \phi|\mathbf{y}),$$

from which the marginalised distributions $p(\alpha|\mathbf{y}, \mathbf{y}_0)$ and $p(\phi|\mathbf{y}, \mathbf{y}_0)$ can be determined. In particular, if the additional information provides a very accurate estimate a_0 of α, then

$$\phi|\mathbf{y}, \mathbf{y}_0 \sim G(m/2, ms_0^2/2), \quad s_0^2 = \sum_{i=1}^{2}(y_i - a_0)^2.$$

If a_0 is significantly different from $\bar{y}$, the additional information about $\boldsymbol{\alpha}$ will have a similarly significant impact on the state of knowledge distribution for ϕ.

In the context on an ILC, the additional information about α comes from the results provided by other laboratories. We consider the case where all the participants measurement results are associated with the analysis of repeated measurements. Define $f(\alpha, \phi|m, y, s)$ to be the function

$$f(\alpha, \phi|m, y, s) = \phi^{m/2-1} \exp\left\{-\frac{\phi}{2}\left[(m-1)s^2 + m(\bar{y}-\alpha)^2\right]\right\}.$$

We assume that the jth participant, $j = 1, \ldots, n$, associates with the measurand, α, the distribution $t_{m_j-1}(\bar{y}_j, s_j^2/m_j)$, derived from a sample $\mathbf{y}_j$ generated according to $y_{j,i} \in \mathrm{N}(\alpha, \phi_j^{-1})$ using the prior $p(\alpha, \phi_j) \propto 1/\phi_j$. We note here that the distribution $p(\phi_j|\mathbf{y}_j)$ associated with ϕ_j is the Gamma distribution $G((m_j-1)/2, (m_j-1)s_j^2/2)$. The posterior distribution $p(\alpha, \boldsymbol{\phi}|\{\mathbf{y}_j\})$ is such that

$$p(\alpha, \boldsymbol{\phi}|\{\mathbf{y}_j\}) \propto \prod_{j=1}^{n} f(\alpha, \phi_j|m_j, \bar{y}_j, s_j).$$

The variable ϕ_j appears in only one factor of the product so that analytical marginalisation can be performed a factor at a time to determine $p(\alpha|\{\mathbf{y}_j\})$ with

$$p(\alpha|\{\mathbf{y}_j\}) \propto \prod_{j=1}^{n} t_{m_j-1}(\alpha|\bar{y}_j, s_j^2/m_j). \tag{1}$$

This distribution takes into account the fact that there is only partial information about σ_j derived from the m_j samples $\mathbf{y}_j$. (The posterior distribution depends on the data $\{\mathbf{y}_j\}$ only through $\{\bar{y}_j\}$, $\{s_j\}$ and $\{m_j\}$.)

If our primary concern in the analysis of the data is to derive a consensus state of knowledge distribution for α, then we need go no further than determining $p(\alpha|\{\mathbf{y}_j\})$ in (1). A single numerical quadrature calculation is required to determine the normalising constant for this distribution. If we also wish to update our knowledge about ϕ_k, we can choose to perform the marginalisation with respect to all but ϕ_k to determine $p(\alpha, \phi_k|\{\mathbf{y}_j\})$ with

$$p(\alpha, \phi_k|\{\mathbf{y}_j\}) \propto f(\alpha, \phi_k|m_k, \bar{y}_k, s_k) \times \prod_{j\neq k} t_{m_j-1}(\alpha|\bar{y}_j, s_j^2/m_j).$$

A two dimensional numerical quadrature scheme can be used to marginalise this distribution with respect to α to determine $p(\phi_k|\{\mathbf{y}_j\})$, and therefore the posterior distribution for σ_k, given the information available from all the ILC participants. Thus, the formulation involving t-distributions already provides a mechanism that is akin to adjusting the input uncertainties.

As noted above, the distribution for $\phi_k|\mathbf{y}_k$ derived from $\mathbf{y}_k$ alone is a Gamma distribution and the posterior distribution $\phi_k|\{\mathbf{y}_j\}$ determined from all the ILC data is likely to be similar to a Gamma distribution. We can use the method of moments to approximate the posterior distribution $p(\phi_k|\{\mathbf{y}_j\})$ by the Gamma distribution $G(\hat{\nu}_k/2, \hat{\nu}_k\hat{s}_k^2/2)$ where

$$\hat{\nu}_k = \frac{2E_k^2}{V_k}, \quad \hat{s}_k^2 = \frac{1}{E_k}, \tag{2}$$

and

$$E_k = \mathrm{E}(\phi_k), \quad V_k = \mathrm{Var}(\phi_k),$$

are the mean and variance of the posterior distribution $p(\phi_k|\{\mathbf{y}_j\})$, respectively.

We can regard $\hat{s}_k$ in (2) as an updated estimate of σ_k but the validity of this estimate will depend on the extent to which a Gamma distribution is a good approximant for the poster distribution $p(\phi_k|\{\mathbf{y}_j\})$. The new knowledge about ϕ_k can also be used to predict the measurement result $\tilde{y}$ if the kth instrument is used to measure a new quantity $\tilde{\alpha}$. The distribution for $\tilde{y}$ is given by the posterior predictive distribution $p_k(\tilde{y}|\tilde{\alpha}, \{\mathbf{y}_j\})$ which can be calculated by marginalising $p_k(\tilde{y}, \phi_k|\tilde{\alpha}, \{\mathbf{y}_j\})$ with respect to ϕ_k. We have

$$\begin{aligned} p_k(\tilde{y}, \phi_k|\tilde{\alpha}, \{\mathbf{y}_j\}) &\propto p(\tilde{y}, |\phi_k, \tilde{\alpha}, \{\mathbf{y}_j\})p(\phi_k|\tilde{\alpha}, \{\mathbf{y}_j\}), \\ &= p(\tilde{y}|\tilde{\alpha}, \phi_k)p(\phi_k|\{\mathbf{y}_j\}), \end{aligned} \tag{3}$$

with $\tilde{y}|\tilde{\alpha}, \phi_k \sim \mathrm{N}(\tilde{\alpha}, \phi_k^{-1})$. Thus, the two terms on the right hand side of (3) are known and the predictive distribution can be determined using a two dimensional numerical quadrature scheme. The predictive distribution is likely to be similar to a t-distribution.

If $m_j > 3$, $j = 1, \ldots, n$, a posterior estimate for α can also be determined using an estimation approach [4] to determine the weighted mean

$$\bar{a} = \left(\sum_{j=1}^{n} \frac{1}{u_j^2}\right)^{-1} \sum_{j=1}^{n} \frac{\bar{y}_j}{u_j^2}, \tag{4}$$

and associated uncertainty $u(\bar{a})$ with

$$u^2(\bar{a}) = \left(\sum_{j=1}^{n} \frac{1}{u_j^2}\right)^{-1}, \quad u_j = \left(\frac{m_j-1}{m_j-3}\right)^{1/2} \frac{s_j}{m_j^{1/2}}, \tag{5}$$

the standard deviation of the distribution $t_{m_j-1}(\bar{y}_j, s_j^2/m_j)$. The distribution associated with α is usually taken to be $\mathrm{N}(\bar{a}, u^2(\bar{a}))$ which we denote by $p_{\mathrm{LS}}(\alpha)$.

Figure 1 shows data simulating sets of repeat measurements $\mathbf{y}_j$, $j = 1, \ldots, n = 8$, drawn from $\mathrm{N}(0, \sigma_j^2)$. From $\mathbf{y}_j$, are calculated mean $\bar{y}_j$, standard deviation s_j. The collection of data has two slightly usual features. The data set $\mathbf{y}_5$ by chance has all but one of the 10 draws greater than zero, leading to a mean of 0.93 while data set $\mathbf{y}_8$ has all 9 draws less than zero, leading to a mean of -0.54. Furthermore, by chance the standard deviation of $\mathbf{y}_8$ is much smaller than σ_8. Figure 1 also shows shows the uncertainty bars $\bar{y}_j \pm 2u_j$ where u_j is calculated as in (5). Compared to the true mean of zero, $\bar{y}_5$ and $\bar{y}_8$ look outlying. The observed χ^2 value associated with the sum of squares of the residual errors for the weighted mean is 23.16 and $\Pr(\chi^2 > 23.15|\nu = 7)$ is approximately 0.002, indicating that it is very unlikely that the $\bar{y}_j/u_j$ are samples from $\mathrm{N}(\alpha, 1)$ for some α. (We know, of course, that they are not samples from such a distribution.) If we inflate all the uncertainties u_j by a factor of $\hat{\sigma} = 1.82$, then the resulting observed χ^2 value is equal to its expected value of $n-1 = 7$. This inflation corresponds to the Birge adjustment procedure for this data [2]. We denote by $p_{\mathrm{LS,B}}(\alpha)$ the distribution corresponding to $\mathrm{N}(\bar{a}, \hat{\sigma}^2 u^2(\bar{a}))$.

Figure 2 graphs the posterior distribution for $p(\alpha|\{\mathbf{y}_j\})$ calculated according to (1) and the distributions $p_{\mathrm{LS}}(\alpha)$ and $p_{\mathrm{LS,B}}(\alpha)$ associated with the weighted mean calculated in (4). The distributions are significantly different from each other. Figure 1 graphs the consensus value a for α given by the mean of the distribution $p(\alpha|\{\mathbf{y}_j\})$, solid horizontal bar, along with the uncertainty band $a \pm 2u(a)$, dashed horizontal bars, where $u(a)$ is the standard deviation of $p(\alpha|\{\mathbf{y}_j\})$.

Figure 3 shows the distribution $p(\phi_1|\{\mathbf{y}_j\})$, its Gamma approximant and $p(\phi_1|\mathbf{y}_1)$. All three distributions are very similar. For $j = 2$, 3, 4, 6 and 7, the three distributions are also similar. Figure 4 gives the three distributions for $j = 5$. In this case the distributions $p(\phi_5|\{\mathbf{y}_j\})$ and $p(\phi_5|\mathbf{y}_5)$ are significantly different, reflecting the fact that the posterior estimate $\hat{s}_5$ is larger than the estimate s_5 derived from $\mathbf{y}_5$ alone. The fact that $\bar{y}_5$ is far from the consensus value provides evidence that s_5 is an underestimate. The Gamma approximant in this case also is close to $p(\phi_5|\{\mathbf{y}_j\})$. Similarly, figure 5 gives the three distributions for $j = 8$. In this case the distributions $p(\phi_8|\{\mathbf{y}_j\})$ and $p(\phi_8|\mathbf{y}_8)$ are different, reflecting the fact that the posterior estimate $\hat{s}_8$ is larger than the estimate s_8 derived from $\mathbf{y}_8$ alone. The fact that $\bar{y}_8$ is far from the consensus value provides evidence that s_8 is an underestimate. The Gamma approximant in this case is somewhat different to $p(\phi_8|\{\mathbf{y}_j\})$.

The quantities $\hat{\nu}_j$ and $\hat{s}_j$ can be used to adjust the uncertainties u_j in a least squares analysis (but we do not necessarily recommend this approach). For example, setting

$$\hat{u}_j = \left(\frac{\hat{\nu}_j}{\hat{\nu}_j - 2}\right)^{1/2} \frac{\hat{s}_j}{{m_j}^{1/2}}, \qquad (6)$$

we can perform a least squares analysis as in (4) with these posterior adjusted uncertainties. Figure 6 shows the uncertainty bars $\bar{y}_j \pm 2u_j$, $\bar{y}_j \pm 2\hat{u}_j$ and $\bar{y}_j \pm 2\hat{\sigma}u_j$, this last adjustment that following the Birge procedure. For all but $j = 5$ and $j = 8$, $\hat{u}_j$ is very similar to u_j while for $j = 5$ and particularly $j = 8$, the $\hat{u}_j$ is considerably larger. The observed χ^2 value for the weighted mean calculated using $\hat{u}_j$ is 10.6, much closer to the expected value of 7. The Birge procedure inflates all uncertainties by the same amount.

3. ANALSIS OF ILC DATA USING t-DISTRIBUTIONS

The previous section showed that the analysis of ILC data where each laboratory assigns a t-distribution on the basis of observed repeated measurements allows the posterior distributions for $\phi_j = 1/\sigma_j^2$ to be evaluated. If the mean $\bar{y}_j$ is far from the consensus value relative to u_j, the posterior distribution for ϕ_j will usually assign more probabil-

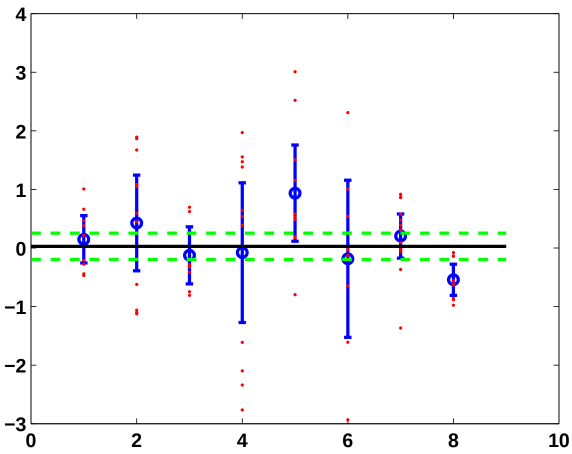

Fig. 1. Data $\mathbf{y}_j$, $j = 1, \ldots, n = 8$, dots, along with means $\bar{y}_j$, circles and uncertainty bars $\pm 2u_j$ associated with the means $\bar{y}_j$.

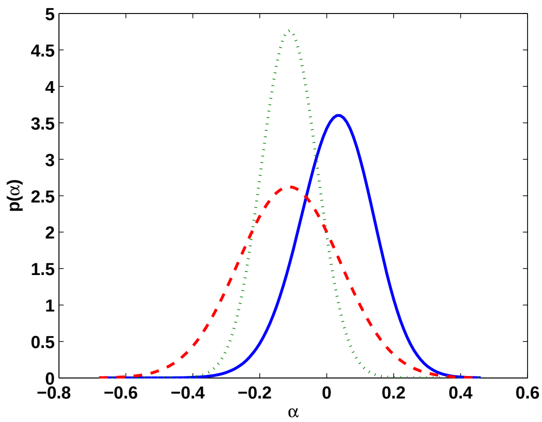

Fig. 2. Posterior distribution $p(\alpha|\{\mathbf{y}_j\})$, solid, calculated according to (1) and the distributions $p_{\mathrm{LS}}(\alpha)$, dotted, and $p_{\mathrm{LS,B}}(\alpha)$, dashed, associated with the weighted mean calculated in (4).

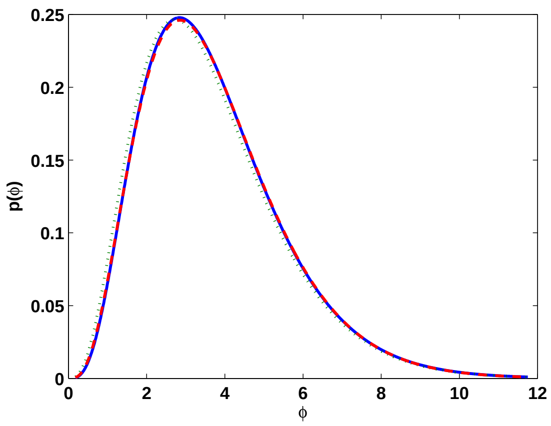

Fig. 3. Posterior distribution $p(\phi_1|\{\mathbf{y}_j\})$ calculated from all the ILC data, solid, along with its Gamma approximant, dashed, and the distribution $p(\phi_1|\mathbf{y}_1)$, dotted, derived from $\mathbf{y}_1$ alone.

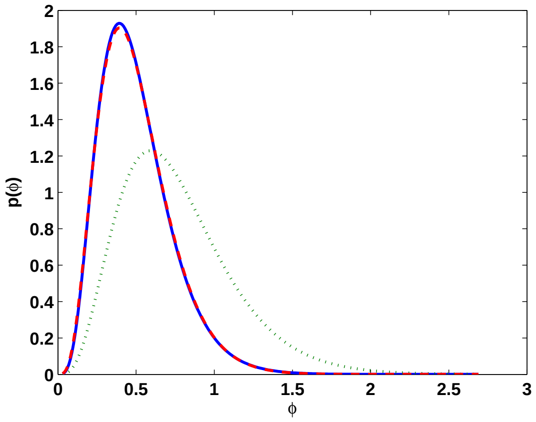

Fig. 4. Posterior distribution $p(\phi_5|\{\mathbf{y}_j\})$ calculated from all the ILC data, solid, along with its Gamma approximant, dashed, and the distribution $p(\phi_5|\mathbf{y}_5)$, dotted, derived from $\mathbf{y}_5$ alone.

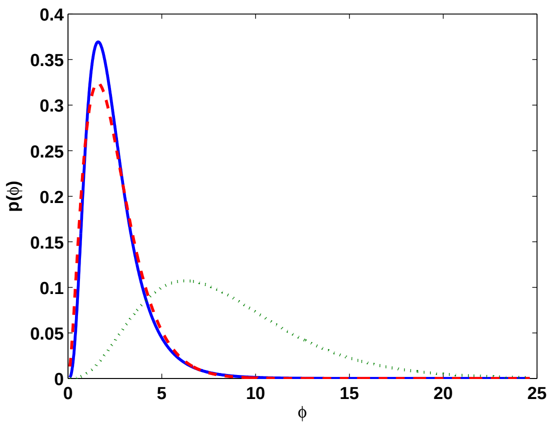

Fig. 5. Posterior distribution $p(\phi_8|\{\mathbf{y}_j\})$ calculated from all the ILC data, solid, along with its Gamma approximant, dashed, and the distribution $p(\phi_8|\mathbf{y}_8)$, dotted, derived from $\mathbf{y}_8$ alone.

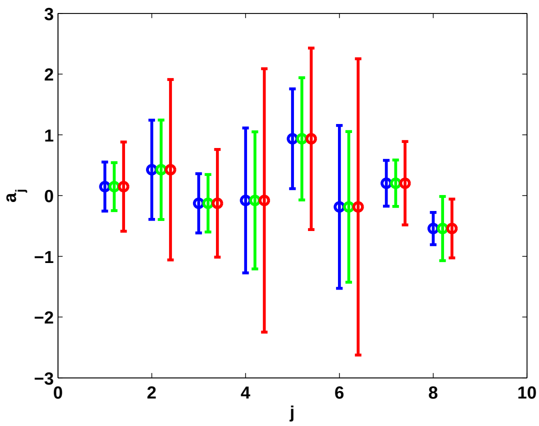

Fig. 6. Uncertainty bars $\bar{y}_j \pm 2u_j$, left, $\bar{y}_j \pm 2\hat{u}_j$, middle, and $\bar{y}_j \pm 2\hat{\sigma}u_j$, right.

ity mass to larger values of σ_j. This adjustment procedure uses no additional modelling assumptions. The above analysis suggests the following approach for analysing ILC data $(x_j, u_j = u(x_j))$. Instead of starting with the assumption $x_j|\alpha \in \mathrm{N}(\alpha, u_j^2)$, $j = 1, \ldots, n$, we assume instead

$$x_j|\alpha, \phi_j \in \mathrm{N}(\alpha, \phi_j^{-1}), \quad j = 1, \ldots, n$$

with (independent) priors

$$\phi_j|\nu_j \sim G(\nu_j/2, \nu_j u_j^2/2)$$

for ϕ_j, with the degrees of freedom parameter ν_j encoding our degree of belief in the estimate u_j of $\sigma_j = 1/\phi_j^{1/2}$. The simplest case to consider is where $\nu_j = \nu \geq 1$ is the same for all participants and is hyper-parameter that represents a single measure of the degree of belief associated with the set of quoted uncertainties $\{u_j\}$.

Given $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, the analysis above determines the posterior distribution

$$p(\alpha, \nu|\mathbf{x}) \propto p(\nu) \prod_{j=1}^{n} t_\nu(\alpha|x_j, u_j^2). \tag{7}$$

Using a two-dimensional quadrature scheme, the posterior distributions $p(\alpha|\mathbf{x})$ and $p(\nu|\mathbf{x})$ can be determined. The posterior distribution $p(\alpha|\mathbf{x})$ is the consensus distribution for α determined from all the participants while $p(\nu|\mathbf{x})$ reflects the mutual consistency of the results. The choice of a prior distribution for ν needs to be considered. In the calculations below, we set $\zeta = 1/\nu$, $0 \leq \zeta \leq 1$ and assign an exponential distribution $p(\zeta) \propto e^{-\lambda\zeta}$. As λ increases, the more probability mass is associated with higher values of ν, corresponding to greater belief in the input $\{u_j\}$.

As in the case of repeated measurements, posterior information about ϕ_k can be derived by marginalising with respect to ϕ_j, $j \neq k$, yielding

$$p(\alpha, \phi_k, \nu|\mathbf{x}) \quad \propto \quad p(\nu)\phi_k^{\left(\frac{\nu+1}{2}-1\right)} \prod_{j \neq k} t_\nu(\alpha|x_j, u_j^2) \times \exp\left\{-\frac{\phi_k}{2}\left[\nu u_k^2 + (x_k - \alpha)^2\right]\right\}.$$

The posterior distributions $p(\phi_j|\mathbf{x})$ can be determined using a three-dimensional quadrature scheme. (Such schemes can be implemented on a standard PC and take of the order of a few seconds to compute; quadrature for dimensions much higher than three are not practical.) A computationally cheaper alternative is to determine the posterior distribution $p(\nu|\mathbf{x})$ using a two-dimensional quadrature scheme for $p(\alpha, \nu|\mathbf{x})$ in (7), set ν to be the maximum likelihood estimate ν_0 that maximises $p(\nu|\mathbf{x})$, and perform two-dimensional quadratures to determine $p(\phi_j|\mathbf{x}, \nu = \nu_0)$.

Once the distribution $p(\phi_k|\mathbf{x})$ has been determined, it can be used to used to determine the posterior predictive measurement for the k laboratory as in (3). This distribution is symmetric and the 2.5 and 97.5 percentiles $\tilde{\alpha} \pm \tilde{U}_k$ for this distribution can be used to provide 95 % uncertainty bars for the measurement result x_k and provide a measure of the posterior adjustment to the input uncertainty u_k.

4. EXAMPLE CALCULATIONS: GAUGE BLOCK KEY COMPARISONS

We give calculations for data derived from the Length Consultative Committee key comparison CCL-K1, calibration of gauge blocks by interferometry. The measured results x_j and associated standard uncertainties u_j are from eleven participating laboratories measuring nine tungsten carbide gauge blocks of nominal lengths ranging from 0.5 mm to 100 mm [11, Table 5]. The data was analysed using the approach given in section 3. The prior for $\zeta = 1/\nu$ chosen was $p(\zeta) \propto e^{-\lambda\zeta}$ with $\lambda = 10$.

Figures 7–9 show the results of the analysis for gauge blocks of nominal length 0.5 mm, 1 mm and 6 mm. All three data sets show inconsistency. The graphs show uncertainty bars $x_j \pm 2u_j$ based on the input uncertainties u_j, $x_j \pm \tilde{U}_j$ derived from the posterior 95 % coverage interval for the posterior predictive distributions and $x_j \pm 2\hat{\sigma}u_j$ based on a Birge adjustment in which all input uncertainties are scaled to achieve consistency. Figures 10–12 graph the posterior distributions $p(\alpha|\mathbf{x})$ for the three data sets in figures 7–9 along with the Gaussian distributions $p_{\text{LS}}(\alpha)$ associated with the weighted mean calculations and the Gaussian distributions $p_{\text{LS,B}}(\alpha)$ associated with Birge adjustment procedure. The posterior distributions $p(\zeta|\mathbf{x})$, $\zeta = 1/\nu$, for the three data sets in figures 7–9 are given in figure 13. These distributions indicate the level of consistency in the data sets.

5. CONCLUSIONS

This paper has been concerned with analysis of interlaboratory comparison (ILC) data and approaches to resolving inconsistency in such data. We have discussed the situation in which the measurement result from each laboratory is derived from the analysis of repeated measurements from Gaussian distributions with unknown variances. For this case, it is appropriate to associate a t-distribution with each measurement result and the distribution for the consensus value of the measurand is then given by a product of t-distributions. However, using a Bayesian approach, it is also possible and computationally straightforward to derive posterior distributions associated with the variances of the sampling distributions for each laboratory. These updated distributions for the variance parameters are based on the analysis of all the measurements and providing new information about each laboratory, amounting to a posterior adjustment of their uncertainty information. We have used this idea to provide a general approach to analysing ILC data using a parameter ν, a degrees of freedom parameter, that encodes the degree of belief in the stated uncertainties. The approach depends only on the measurement results, their associated uncertainties and a prior distribution for ν and delivers a state of knowledge distribution for the measurand along with updated distributions for the variance parameters associated with each laboratory. We have illustrated the approach on data associated with a gauge block key comparison.

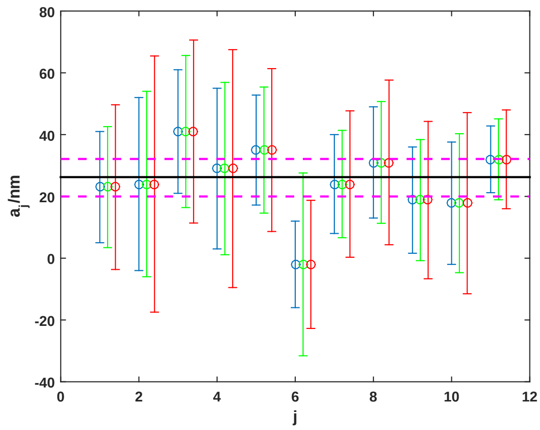

Fig. 7. Analysis of key comparison measurement data for a 0.5 mm tungsten carbide gauge block. Uncertainty bars $x_j \pm 2u_j$, left, $x_j \pm \tilde{U}_j$ derived from the posterior predictive distributions, middle, and $x_j \pm 2\hat{\sigma}u_j$ derived using a Birge adjustment, right. The solid horizonal line indicates the mean and the horizontal dotted lines the 2.5 and 97.5 percentiles of the posterior distribution $p(\alpha|\mathbf{x})$.

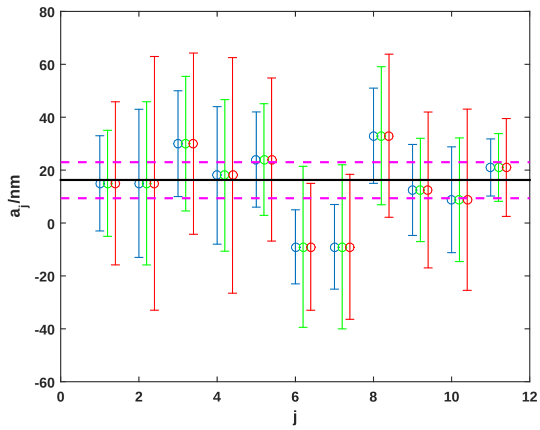

Fig. 8. As figure 7 but for measurement data for a 1 mm tungsten carbide gauge block.

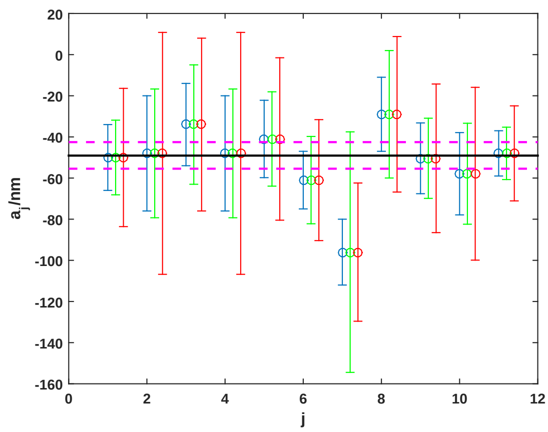

Fig. 9. As figure 7 but for measurement data for a 6 mm tungsten carbide gauge block.

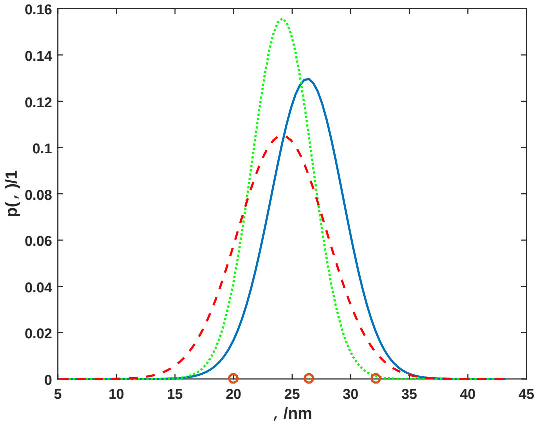

Fig. 10. Posterior distribution $p(\alpha|\mathbf{x})$, solid, for the data in figure 7 derived from marginalising $p(\alpha, \nu|\mathbf{x})$ in (7) and the distributions $p_{\mathrm{LS}}(\alpha)$, dotted, and $p_{\mathrm{LS,B}}(\alpha)$, dashed, associated with the weighted mean calculations and the Birge adjustment procedure.

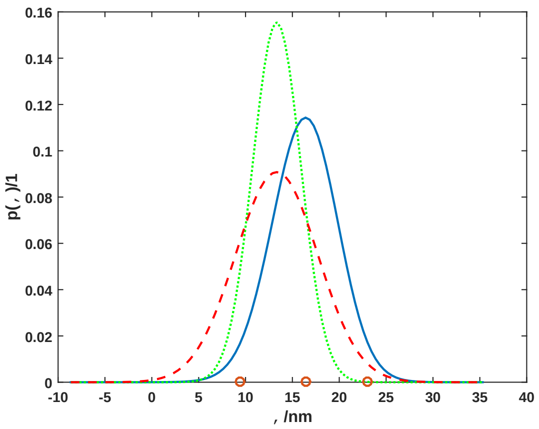

Fig. 11. As figure 10 but for measurement data for a 1 mm tungsten carbide gauge block.

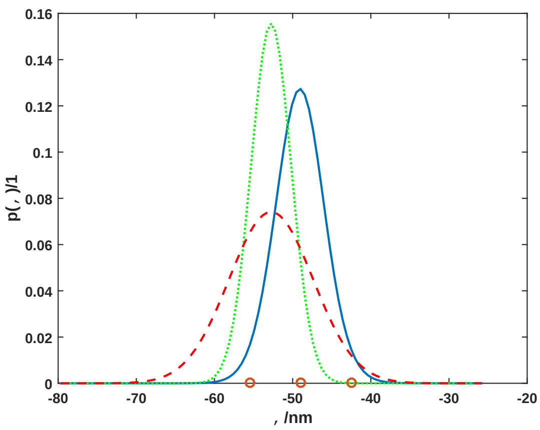

Fig. 12. As figure 10 but for measurement data for a 6 mm tungsten carbide gauge block.

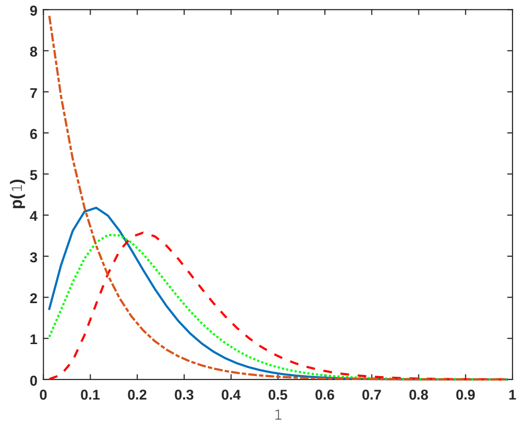

Fig. 13. Posterior distributions $p(\zeta|\mathbf{x})$, $\zeta = 1/\nu$, for data in figures 7–9, solid, dotted, dashed, respectively, along with the prior distribution $p(\zeta)$, dot-dashed.

REFERENCES

[1] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008.

[2] R. T. Birge. Probable values of the general physical constants. *Rev. Mod. Phys*, 1:1–73, 1929.

[3] A. G. Chunovkina, C. Elster, I. Lira, and W. Wöger. Analysis of key comparison data and laboratory biases. *Metrologia*, 45(2):211–216, 2008.

[4] M. G. Cox. The evaluation of key comparison data. *Metrologia*, 39:589–595, 2002.

[5] M. G. Cox, A. B. Forbes, J. Flowers, and P. M. Harris. Least squares adjustment in the presence of discrepant data. In P. Ciarlini, M. G. Cox, F. Pavese, and G. B. Rossi, editors, *Advanced Mathematical and Computational Tools in Metrology VI*, pages 37–51, Singapore, 2004. World Scientific.

[6] M.G. Cox. The evaluation of key comparison data: determining the largest consistent subset. *Metrologia*, 44(3):187–200, 2007.

[7] C. Elster and B. Toman. Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using Bayesian model averaging. *Metrologia*, 47(3):113–119, 2010.

[8] A. B. Forbes and C. Perruchet. Measurement systems analysis: concepts and computational approaches. In *IMEKO World Congress, September 18–22, 2006, Rio de Janeiro*, 2006.

[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Fl., second edition, 2004.

[10] R. C. Paule and J. Mandel. Consensus values and weighting factors. *J Res. Natl. Bur. Stand.*, 87:377–385, 1982.

[11] R. Thalmann. CCL key comparison: calibration of gauge blocks by interferometry. *Metrologia*, 39:165–177, 2002.

[12] K. Weise and W. Woeger. Removing model and data nonconformity in measurement evaluation. *Measurement Science and Technology*, 11:1649–1658, 2000.