Int. J. Metrol. Qual. Eng. **10**, 3 (2019)

https://doi.org/10.1051/ijmqe/2019003

International Journal of Metrology and Quality Engineering

Available online at:
www.metrology-journal.org

RESEARCH ARTICLE OPEN ACCESS

Assessing interlaboratory comparison data adjustment procedures

Kavya Jagan* and Alistair B. Forbes

Data Science Group, National Physical Laboratory, Teddington, Middlesex, UK

Received: 4 December 2018 / Accepted: 16 March 2019

Abstract. Interlaboratory comparisons (ILCs) are one of the key activities in metrology. Estimates $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ of a measurand α along with their associated standard uncertainties $\boldsymbol{u}_0 = (u_{0,1}, \ldots, u_{0,n})^{\mathrm{T}}$, $u_{0,j} = u_0(x_j)$ are provided by each of n laboratories. Employing a model of the form

$$x_j \in N(\alpha, v_{0,j}), \qquad j = 1, \ldots, n,$$

$v_{0,j} = u_{0,j}^2$, we may wish to find a consensus value for α. A χ^2 test can be used to assess the degree to which the spread of the estimates $\boldsymbol{x}$ are consistent with the stated uncertainties $\boldsymbol{u}_0$. If they are judged to be inconsistent, then an adjustment procedure can be applied to determine $v_j \geq v_{0,j}$, so that $\boldsymbol{x}$ and $\boldsymbol{v}$ represent consistency. The underlying assumption behind this approach is that some or all of the laboratories have underestimated or neglected some uncertainty contributions, sometimes referred to as 'dark uncertainty', and the adjusted $\boldsymbol{v}$ provides an estimate of this dark uncertainty derived from the complete set of laboratory results. There are many such adjustment procedures, including the Birge and Mandel–Paule (M-P) procedures.

In implementing an adjustment procedure, a desirable objective is to make as minimal an adjustment as necessary in order to bring about the required degree of consistency. In this paper, we discuss the use of relative entropy, also known as the Kullback–Leibler divergence, as a measure of the degree of adjustment. We consider parameterising $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{b})$ as a function of parameters $\boldsymbol{b}$ with the input $\boldsymbol{v}_0 = \boldsymbol{v}(\boldsymbol{b}_0)$ for some $\boldsymbol{b}_0$. We look to perturb $\boldsymbol{b}$ from $\boldsymbol{b}_0$ to bring about consistency in a way that minimises how far $\boldsymbol{b}$ is from $\boldsymbol{b}_0$ in terms of the relative entropy or Kullback–Leibler divergence.

Keywords: Inter-laboratory comparisons / consistency / Birge adjustment / Kullback–Liebler divergence

1 Introduction

Interlaboratory comparisons (ILCs) are one of the key activities in metrology, and are undertaken for a number of reasons, e.g., to determine a consensus value for a quantity by combining experimental results from a number of laboratories. In this paper, we are primarily interested in finding a consensus estimate for the quantity being measured. The starting point for the analysis of ILC data is usually a model [1] of the form

$$x_j \in \mathrm{N}(\alpha, v_j), \qquad j = 1, \ldots, n, \tag{1}$$

where α is the value of the measurand, x_j is the estimate of α produced by the jth laboratory and $v_j = u_j^2$ its associated variance, where u_j is its standard uncertainty. Given the model (1), the distribution for α is derived from the weighted least squares (WLS) estimate

$$a = a(\boldsymbol{v}) = \left(C^{\mathrm{T}}C\right)^{-1}C^{\mathrm{T}}\boldsymbol{y}, \qquad y_j = \frac{x_j}{u_j}, \qquad C_j = \frac{1}{u_j},$$

involving the $n \times 1$ observation matrix C. If the data is regarded as a sample from Gaussian distributions as in (1), then a is a sample from the Gaussian distribution $\mathrm{N}(a, V_a)$, $V_a = \left(C^{\mathrm{T}}C\right)^{-1}$.

Under the model (1), the sum of the squares of the weighted residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{v}) = \boldsymbol{y} - Ca$,

$$R^2 = R^2(\boldsymbol{v}) = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{r},$$

is a sample from a χ^2_{n-1} distribution with $n-1$ degrees of freedom. The degree of consistency of the data $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ and $\boldsymbol{u} = (u_1, \ldots, u_n)^{\mathrm{T}}$ can be assessed by evaluating $\Pr(\zeta^2 \geq R^2)$ where $\zeta^2 \sim \chi^2_{n-1}$. If $\Pr(\zeta^2 \geq R^2)$ is less than 5%, say, then there is doubt about consistency of the data [1]. If the data is judged to be inconsistent, the WLS estimate a might not be a reliable estimate of α in the

* Corresponding author: kavya.jagan@npl.co.uk

sense that the variance matrix V_a may underestimate the uncertainty associated with this estimate. In the presence of inconsistency, a procedure can be applied to adjust the uncertainties to achieve consistency, see, e.g., [2–10]. The underlying assumption behind these approaches is that some or all of the laboratories have underestimated or neglected some uncertainty contributions, sometimes referred to as 'dark uncertainty' [11], and the adjusted uncertainties implicitly provide an estimate of this dark uncertainty derived from the complete set of laboratory results.

The methods discussed in this paper are relevant to several types of ILCs such as those included in ISO standards and in the field of testing. For these types of ILC, a primary objective is to learn about laboratory effects from the information generated in the inter-comparison. They are however not directly relevant to the analysis of Key Comparisons (KC) as the type of data adjustments considered in this paper are not permitted.

1.1 When to adjust, how far to adjust

The issue of whether or not to adjust a set of input uncertainties can be thought of as a model selection problem involving two models. The first model M_1 assumes complete faith in the input data $\boldsymbol{x}$ and $\boldsymbol{u}_0$. In model M_1, any large value for R^2, say greater than the 95th percentile $\tilde{c}^2$, of the relevant χ^2 distribution is due to chance: we expect R^2 to exceed this threshold for 1 in 20 ILCs performed at 'random'. The second model M_2 assumes that one or more of the laboratories have underestimated their uncertainties and that some adjustment to the input uncertainties may be necessary to achieve a more consistent set of results.

The observed value of R^2 is used to decide which of these models is most supported by the data: we set $\tilde{c}^2$ so that if $R^2 \leq \tilde{c}^2$, we select M_1 and select model M_2 otherwise. For example, if $\tilde{c}^2$ corresponds to the 95th percentile, we are saying that there is a better than 1 in 20 chance that inconsistency is due to model M_2 than model M_1. (It is possible to provide a more formal methodology for such model selection problems, e.g., [12,13], but this is not the focus of this paper. See also [6,14] that discuss Bayesian model *averaging* approaches to the analysis of ILC data.)

We now turn to the question of how far to adjust. Prior to the analysis, we assume that a threshold value $\tilde{c}^2$ for R^2 has been assigned, say $\tilde{c}^2$ corresponding to the 95th percentile of χ^2_{n-1} and that $R^2_0 > \tilde{c}^2$, so that model M_2 is selected. We look for adjusted variances $\boldsymbol{v}$ such that the corresponding value of $R^2(\boldsymbol{v})$ is such that $R^2(\boldsymbol{v}) = c^2 < \tilde{c}^2$. The choice of the constraint value c^2 needs to be considered. A common choice is to set $c^2 = n-1$, the expected value of the distribution χ^2_{n-1}. In the example below, there are $n = 11$ laboratories. The 95th percentile for χ^2_{10} is $\tilde{c}^2 = 18.3$ while its mean is 10 which in fact corresponds to the 56th percentile, $P(\chi^2 < 10) = 0.56$. (The 50th percentile, in other words the median, of a χ^2_ν distribution is approximately $\nu(1-2/(9\nu))^3$ and is somewhat less than its mean, due to the skewness of the distribution.) Thus, if $R^2_0 = 18.5$, above the threshold, an adjustment procedure will be applied to achieve a value of $R^2(\boldsymbol{v}) = 10$, while if $R^2_0 = 18.1$, less than the threshold, no adjustment procedure is applied. Hence, a relatively small change in the input data can lead to relatively large changes in the inferences about $\boldsymbol{\alpha}$. We would prefer the adjustment procedures to be smooth with respect to changes in the input data and therefore argue that c^2 should be the same as $\tilde{c}^2$. It also seems somewhat illogical to tolerate a value of R^2_0 just below the threshold value but, if adjustment is deemed necessary, to adjust to a level of consistency much lower than the threshold value.

We return to the question of what is a good value for $\tilde{c}^2 (= c^2)$. While a value corresponding the 95th percentile is usual [5], other choices such as a value corresponding to, say, the 80th percentile may be appropriate, being a balance between only adjusting where there is evidence of inconsistency and achieving a level of consistency that is closer to the expected level. It also represents a compromise from current practice to set a threshold at the 95th percentile and adjust, if necessary, to the mean of the χ^2 distribution. In terms of model selection, a choice of the 80th percentile indicates that if $R^2 > \tilde{c}^2$ there is a better than 1 in 5 chance that inconsistency is due to model M_2 than model M_1.

1.2 Relative entropy as a measure of the extent of an adjustment

Our main concern is to design adjustment procedures that lead to adjusted $\boldsymbol{v}$ that represent a minimal departure from the stated $\boldsymbol{v}_0$ according to some measure. We regard the input $\boldsymbol{v}_0$ as being the best estimates that the participating laboratories can deliver, based on their knowledge of their own systems. If the complete set of results indicates inconsistency, then this new knowledge is available to help provide adjusted $\boldsymbol{v}$ that improves consistency, i.e., delivers estimates that are more likely, given the new knowledge but at the same time respects as much as possible the knowledge represented by the initial input data. In this paper, we discuss adjusted uncertainty measures derived by minimising the relative entropy or Kullback–Leibler divergence [15,16] a measure of how far $\boldsymbol{v}$ is from $\boldsymbol{v}_0$. We consider parameterising $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{b})$ as a function of parameters $\boldsymbol{b}$ with the input $\boldsymbol{v}_0 = \boldsymbol{v}(\boldsymbol{b}_0)$ for some $\boldsymbol{b}_0$ and we look to perturb $\boldsymbol{b}$ from $\boldsymbol{b}_0$ to bring about consistency. We now regard $a = a(\boldsymbol{b})$ and $R^2 = R^2(\boldsymbol{b})$ as functions of $\boldsymbol{b}$ through their dependence on $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{b})$.

Section 2 defines the relative entropy of an adjustment and Section 3 describes adjustment through minimising relative entropy. Section 4 describes a number of single parameter adjustment schemes in which the degree of adjustment is determined by the consistency constraint. Section 5 introduces the steepest descent (S-D) adjustment scheme. A comparison of adjustment procedures in terms of the degree of adjustment of ILC data is presented in Section 6. This data is used for illustrative purposes only as key comparison data cannot be adjusted in the manner described in the paper. Our concluding remarks are given in Section 7.

2 Relative entropy as a measure of degree of adjustment

Relative entropy, or Kullback–Leibler divergence [15,16], is a measure of the difference between two distributions. For continuous distributions it is given by

$$D_{\mathrm{KL}}(p(x)||p_0(x)) = \int_{x=-\infty}^{\infty} p(x)\log\left(\frac{p(x)}{p_0(x)}\right)dx.$$

For any $p=p(x)$ and $p_0=p_0(x)$, $D_{\mathrm{KL}}(p||p_0)\geq 0$ and $D_{\mathrm{KL}}(p||p_0)=0$ only if $p=p_0$. Note that $D_{\mathrm{KL}}(p||p_0)\neq D_{\mathrm{KL}}(p_0||p)$ and therefore relative entropy is not a norm. It is usually the case that p_0 is an approximate distribution for p and the relative entropy is a measure of information gained using p over p_0.

For multivariate Gaussian distributions,

$$D_{\mathrm{KL}}(\mathrm{N}(x,V)||\mathrm{N}(x_0,V_0)) = D+S,$$

where

$$D=\frac{1}{2}\log\left(|V_0V^{-1}|\right)+\frac{1}{2}\mathrm{Tr}\left(VV_0^{-1}\right)-\frac{1}{2}n,$$

and

$$S=\frac{1}{2}(x-x_0)^{\mathsf{T}}V_0^{-1}(x-x_0).$$

In the above $|A|$ is the determinant of a square matrix A and $\mathrm{Tr}\,(A)$ is the trace of A, the sum of its diagonal elements. For our application, with $\boldsymbol{x}$ the same for both distributions and diagonal variance matrices, the relative entropy is a function of $\boldsymbol{b}$ and $\boldsymbol{b}_0$ in terms of $\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{b})$ and $\boldsymbol{v}_0=\boldsymbol{v}(\boldsymbol{b}_0)$:

$$D_{\mathrm{KL}}(\boldsymbol{b}||\boldsymbol{b}_0)=\frac{1}{2}\sum_{j=1}^{n}\log\frac{v_{0,j}}{v_j}+\frac{1}{2}\sum_{j=1}^{n}\frac{v_j}{v_{0,j}}-\frac{1}{2}n.$$

3 Adjustment procedure based on minimising relative entropy

We are interested in adjustment procedures that represent as small a change as necessary, as measured by the relative entropy, to bring about consistency. This problem can be formulated as

$$\min_{\boldsymbol{b}} D_{\mathrm{KL}}(\boldsymbol{b}||\boldsymbol{b}_0)$$

subject to $R^2(\boldsymbol{b})\leq c^2$. The constraint is equivalent to $R^2(\boldsymbol{b})=c^2$ as this is a minimisation problem.

The problem can also be parametrised in terms of weights

$$w_{0,j}=\tfrac{1}{v_{0,j}},\ w_j=\tfrac{1}{v_j}$$

with

$$D_{\mathrm{KL}}(\boldsymbol{w}||\boldsymbol{w}_0)=\frac{1}{2}\left[\sum_{j=1}^{n}\log\frac{w_j}{w_{0,j}}+\sum_{j=1}^{n}\frac{w_{0,j}}{w_j}-n\right].$$

In terms of weights $\boldsymbol{w}$, the consistency constraint can be written explicitly as

$$E(\boldsymbol{w})=c^2S_w-S_wS_{wxx}+S_{wx}^2=0,$$

where $S_w=\sum_{j=1}^{n}w_j$, $S_{wx}=\sum_{j=1}^{n}w_jx_j$, $S_{wxx}=\sum_{j=1}^{n}w_jx_j^2$.

Thus, $E(\boldsymbol{w})$ is a quadratic function of the parameters $\boldsymbol{w}$, one of the advantages of working with $\boldsymbol{w}(\boldsymbol{b})$ rather than $\boldsymbol{v}(\boldsymbol{b})$. In terms of the weights $\boldsymbol{w}$, we arrive at the adjustment procedure

$$\min_{\boldsymbol{b}} D_{\mathrm{KL}}(\boldsymbol{w}||\boldsymbol{w}_0)$$

subject to $E(\boldsymbol{w})=0$, $0\leq w_j\leq w_{0,j}$, $j=1,\ldots,n$.

The last set of inequalities imposes the constraint that we only look for an adjustment that increases the input uncertainties.

3.1 Newton scheme to determine $\boldsymbol{w}$

If we assume that $E(\boldsymbol{w}_0)<0$, that is, $R^2(\boldsymbol{v}_0)>c^2$, the optimal $\boldsymbol{w}$ automatically satisfies the constraints $w_j\leq w_{0,j}$, and then the optimal $\boldsymbol{w}$ is a critical point of the Lagrangian function

$$L(\boldsymbol{w},\kappa)=D_{\mathrm{KL}}(\boldsymbol{w}||\boldsymbol{w}_0)+\kappa E(\boldsymbol{w}).$$

This leads to a square system of $n+1$ equations

$$\begin{aligned}\nabla_{\boldsymbol{w}}D_{\mathrm{KL}}+\kappa\nabla_{\boldsymbol{w}}E&=0,\\ E&=0,\end{aligned}$$

that can be solved iteratively using Newton's method, for example [17]. A starting estimate $\boldsymbol{w}(\boldsymbol{b})$ could be $\boldsymbol{w}(\boldsymbol{b}_0)$ with κ defined by the least squares solution of the first n equations above. The Newton scheme can be implemented in stages for decreasing values c_k^2 of c^2, starting with a value of c^2 close to ${c_0}^2$ determined by the input $\boldsymbol{v}_0$, and ending up with the desired value for c^2. The staged approach may help issues with divergence of a simple Newton scheme if the starting point is far from the solution.

4 Single parameter adjustment schemes

There are number of adjustment procedures that are commonly used for which $\boldsymbol{v}=\boldsymbol{v}(\lambda)$ depends on a single parameter λ and the consistency constraint $R^2(\lambda)=c^2$ defines the value of λ.

The Birge procedure [18] corresponds to the model

$$\boldsymbol{v}(\lambda)=\lambda\boldsymbol{v}_0,$$

while the M-P procedure [8] corresponds to model

$$\boldsymbol{v}(\lambda)=\boldsymbol{v}_0+\lambda\boldsymbol{e},$$

where $e_j=1$, $j=1,\ldots,n$. The M-P procedure can be motivated by a model in which all the participants have omitted the same additive influence factor from their uncertainty budget. The variance of this influence factor is

estimated from the consistency constraint. It is not so easy to associate the Birge procedure with a plausible underlying statistical model.

5 Adjustment procedures based on a steepest descent vector

Inconsistency is indicated if the sum of squared residuals $R^2(\boldsymbol{b}_0)$ is too high. This suggests adjusting $\boldsymbol{b}_0$ to $\boldsymbol{b}$ along a descent direction $\boldsymbol{g}$ so that, at least for small $\lambda > 0$, $R^2(\boldsymbol{b}_0 + \lambda \boldsymbol{g}) < R^2(\boldsymbol{b}_0)$. Since, we are interested in determining $\boldsymbol{b}$ that achieves consistency with an adjustment that is minimal in some way, we set $\boldsymbol{g}$ to be the steepest decent (S-D) direction:

$$\boldsymbol{g} = -\nabla_{\boldsymbol{b}} R^2(\boldsymbol{b})|_{\boldsymbol{b}=\boldsymbol{b}_0}.$$

A small step in the direction of $\boldsymbol{g}$ will bring about the swiftest reduction in $R^2(\boldsymbol{b})$ relative to the size of the change in $\boldsymbol{b}$.

The S-D direction depends on how $\boldsymbol{v}(\boldsymbol{b})$ is parametrised in terms of $\boldsymbol{b}$. There are a number of candidates, (i) $v_j(\boldsymbol{b}) = b_j$ so that b_j represents the variance, (ii) $v_j(\boldsymbol{b}) = b_j^2$ so that b_j represents the standard uncertainty, (iii) $v_j(\boldsymbol{b}) = \frac{1}{b_j}$ so that b_j represents the precision or (iv) $v_j(\boldsymbol{b}) = \frac{1}{b_j^2}$, where b_j is square root of the precision.

We may also choose the parametrisation

$$v_j(\boldsymbol{b}) = v_{0,j} e^{b_j}, \quad b_{0,j} = 0. \tag{2}$$

For this parametrisation, we have that

$$\frac{\delta R^2}{\delta b_j}|_{b_{0,j}=0} = \frac{\left(x_j - a_0\right)^2}{v_{0,j}} = \left(\frac{x_j - a_0}{u_{0,j}}\right)^2. \tag{3}$$

Thus, the jth element of the S-D vector depends on the difference between x_j and the weighted least-squares solution a_0 relative to the input uncertainty $u_{0,j}$. This means that the degree of adjustment for a laboratory depends on the extent to which its result is deemed to be outlying, relative to the weighted least-squares solution.

For the parametrisation (2), we have

$$\begin{aligned} D_{\mathrm{KL}}(\boldsymbol{b}||\boldsymbol{b}_0) &= \frac{1}{2}\left[-\sum_{j=1}^{n} b_j + \sum_{j=1}^{n} e^{b_j} - n\right], \\ &= \frac{1}{2}\sum_{j=1}^{n}(e^{b_j} - b_j - 1), \\ &= \frac{1}{2}\left[\sum_{j=1}^{n}\left(\frac{b_j^2}{2!} + \frac{b_j^3}{3!} + \ \ldots\right)\right]. \end{aligned}$$

Thus, for $b_j \approx 0$, $j = 1, \ldots, n$, $D_{\mathrm{KL}}(\boldsymbol{b}||\boldsymbol{b}_0) \propto ||\boldsymbol{b} - \boldsymbol{b}_0||^2$. This means that for the parametrisation (2), a small step along the S-D direction derived from (3) brings about the largest reduction in $R^2(\boldsymbol{b})$ relative to the change in $D_{\mathrm{KL}}(\boldsymbol{b}||\boldsymbol{b}_0)$.

5.1 Stepped approach

With the S-D approach, if the estimate x_j provided by the jth laboratory happens to be very close to the consensus value a_0 determined by the WLS approach (using the input $\boldsymbol{v}(\boldsymbol{b}_0)$), then the S-D vector will have only a very small component in the direction of increasing v_j so that v_j remains essentially unadjusted. If the consensus value a given by the adjusted $\boldsymbol{v}(\boldsymbol{b})$ is significantly different from the initial consensus value, there may be evidence the v_j should also undergo some adjustment as x_j is no longer so close to the new consensus value. This suggests implementing a staged approach determining S-D adjustments for a sequence of values c_k^2 descending from some large c_0^2 (just shy of the χ^2 value determined by the initial data, say) down to the desired c^2.

6 Example calculations

We present calculations for data derived from the Consultative Committee for Length key comparison CCL-K1 [19], calibration of gauge blocks by interferometry. This data is used for illustrative purposes only as according to the technical protocol, key comparison data cannot be adjusted in the manner described in this paper. The measured values x_j and associated standard uncertainties $u_{0,j}$ are from 11 participating laboratories measuring 9 tungsten carbide gauge blocks of nominal lengths ranging from 0.5 mm to 100 mm ([20], Tab. 5). The laboratory results and uncertainty bars $x_j \pm 2u_j$ are represented in Figure 1 given by the first (red) uncertainty bar in each group. The figure shows that the results from laboratories 6 and 7 and, to a lesser extent, laboratory 11 deviated from the consensus value given by the weighted mean, the bottom (solid red) horizontal line. The analysis of the data for the 100 mm gauge block found that the $R^2(\boldsymbol{b}_0)$ value was the 98th percentile of the χ^2 distribution indicating inconsistency of the data: if we simulated a large number of data generated according to the model (1), only 2% would result in values of R^2 greater than or equal to $R^2(b_0)$.

6.1 Adjustment to different consistency levels

As mentioned in Section 1, adjustments are often made to the mean of the χ^2 distribution. We first apply the relative entropy approach to determine adjustments to the 95th and 80th percentiles and to the mean. Table 1 shows the adjusted uncertainties for each level while Figure 1 shows the resulting uncertainty bars $x_j \pm 2u_j$, along with the corresponding weighted means. It is seen that, for the relative entropy method, the main adjustments are made to the results of laboratories 6, 7 and 11, with adjustments increasing as the level of consistency is increased. Of interest is the fact that even for the adjustment from the 98th percentile (the original data) to the 95th percentile, the weighted mean changes.

6.2 Comparison of adjustment procedures

We now leave aside the question of which level of consistency we should adjust and examine the degree

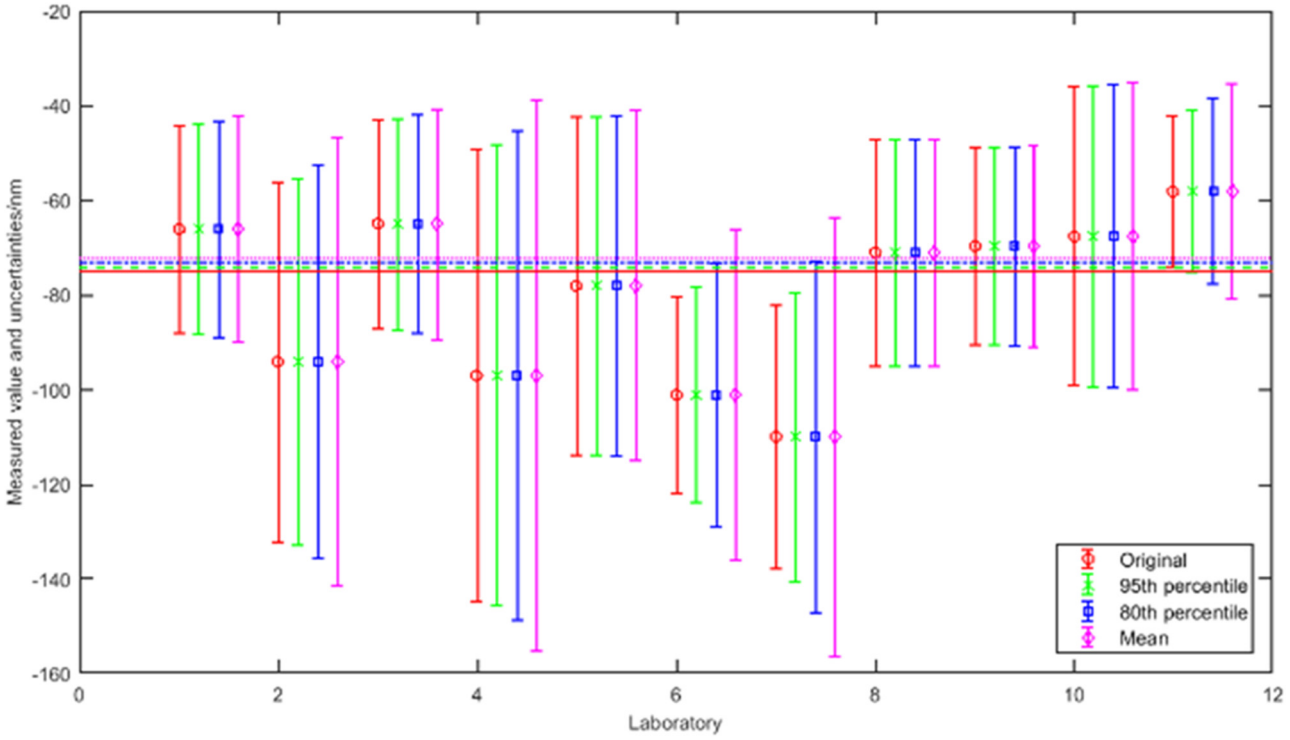

Fig. 1. Estimates and 95% confidence intervals of a 100 mm gauge block. The graph shows the uncertainty bars $x_j \pm 2u_j$ for the input uncertainties $\boldsymbol{u}_0$, the first set of uncertainty bars in each group, and for the adjusted uncertainties based on the relative entropy method $\boldsymbol{u}$ at the 95th percentile, second set, 80th percentile, third set and the mean, fourth set. The horizontal lines are the weighted means determined from the input uncertainties (red solid line), adjusted uncertainties to the 95th percentile (green dashed line), adjusted uncertainties to the 80th percentile (blue dotted and dashed line) and adjusted uncertainties to the mean (magenta dotted line).

Table 1. Uncertainties adjusted to various levels using the relative entropy method.

Lab	Original	95	80	Mean
1	8.00	8.00	8.00	8.02
2	14.00	14.04	14.07	14.12
3	10.00	11.92	13.31	15.47
4	14.00	14.04	14.07	14.12
5	9.40	10.23	10.95	12.18
6	7.00	8.25	9.34	11.03
7	8.00	14.48	17.52	21.66
8	9.00	11.84	13.6	16.16
9	8.60	8.60	8.61	8.66
10	10.00	10.59	11.30	12.65
11	5.50	5.60	5.66	5.77

of adjustment of different adjustment strategies as measured by relative entropy. For these comparisons, the adjustment is determined for c^2 set at the 80th percentile value of the χ^2 distribution. As shown in Table 1, the adjustments made to the uncertainties are also moderate compared to those that are adjusted to the mean of the χ^2 distribution.

Figure 2 shows the estimates $\boldsymbol{x}$ and the uncertainty bars $x_j \pm 2u_j$ for the input uncertainties $\boldsymbol{u}_0$ and for the adjusted uncertainties $\boldsymbol{u}$ from the Birge, M-P, S-D and minimum relative entropy (DKL) procedures.

The Birge and M-P schemes each apply the same adjustment to the uncertainties from all the laboratories, either multiplicatively or additively. The degree of adjustment for the S-D approach depends on how far the laboratory estimate is from the consensus value a_0 determined from the initial WLSs analysis. If x_j is close to a_0 relative to $u_{j,0}$, then u_j remains close to $u_{j,0}$. For laboratories 6, 7 and 11, the S-D approach makes a larger adjustment to the uncertainties compared to the other laboratories as the respective estimates are further away from a_0 relative to the laboratories' initial uncertainty. Visually, there is evidence that the S-D method brings about a smaller adjustment compared to the other two single parameter adjustment schemes, which is expected as shown by the calculations in Section 5. The extent of adjustment using each of these methods can be quantified using the Kullback–Leibler divergence, the lower the divergence, the smaller the adjustment. Table 2 shows that among the single parameter adjustment schemes, the S-D method has the lowest Kullback–Leibler divergence. It is next only to the minimum relative entropy adjustment scheme. Like the S-D scheme, the minimum relative entropy method

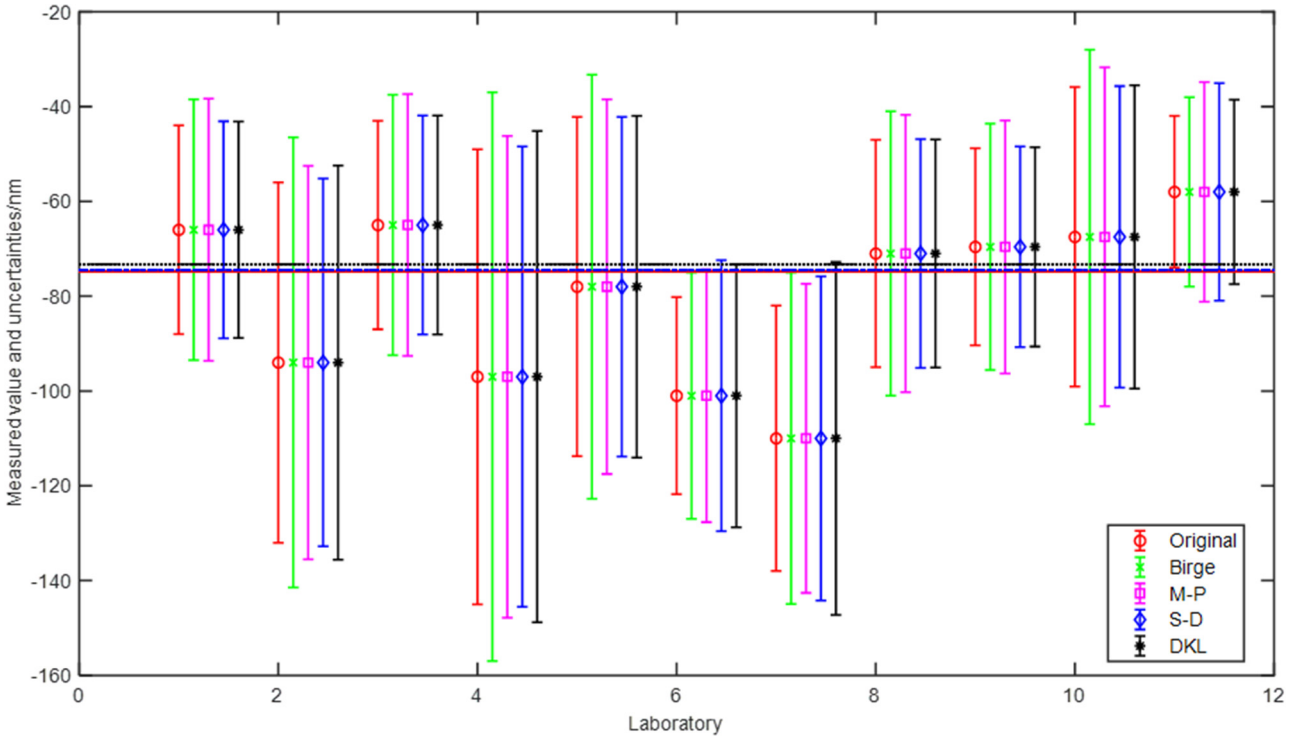

Fig. 2. Estimates and 95% confidence intervals of a 100 mm gauge block for various adjustment procedures. The graph shows the uncertainty bars $x_j \pm 2u_j$ for the input uncertainties $\boldsymbol{u}_0$, the first set of uncertainty bars in each group, and for the adjusted uncertainties $\boldsymbol{u}$ from the Birge, second set, Mandel-Paule (M-P), third set, steepest decent (S-D), fourth set and minimum relative entropy (DKL) procedures, fifth set. The horizontal lines are the weighted means determined from the input uncertainties (solid red), steepest descent (blue dashed) and minimum relative entropy (black dotted).

Table 2. $D_{\mathrm{KL}}\,(\boldsymbol{b}||\boldsymbol{b}_0)$ for various adjustment schemes.

	$D_{\mathrm{KL}}\,(\boldsymbol{b}\|\|\boldsymbol{b}_0)$
Birge	0.64
Mandel–Paule	0.56
Steepest descent	0.36
Relative entropy	0.27

Table 3. Consensus values with their associated uncertainty.

	Estimate/nm	Uncertinity/nm
WLS	−74.8	3.7
Birge	−74.8	4.6
Mandel–Paule	−76.4	4.6
Steepest descent	−75.0	4.2
Relative entropy	−73.2	4.1

only makes a substantial adjustment to the uncertainties of laboratories 6, 7 and 11.

Consensus values and their associated standard uncertainties are shown in Table 3. All the adjustment schemes result in a larger uncertainty compared to that for WLS, which is a direct consequence of adjusting each laboratory's uncertainty $u_{0,j}$. The minimum relative entropy adjustment method yields the smallest uncertainty of all the adjustment procedures as a result of the minimal adjustment to $u_{0,j}$.

There are a number of points of interest associated with Tables 2 and 3. In Table 2, we see that the S-D method makes a smaller change to the input distributions as measured by the relative entropy compared to the Birge and M-P procedures. Both the Birge and M-P procedures provide an adjustment direction that does not depend on how the input estimates x_j relate to the consensus value a_0, whereas the S-D procedure does reflect this information with the uncertainties associated with estimates further from the consensus value undergoing more adjustment. Of course, the fact that there is inconsistency means that use of the initial consensus estimate a_0 to determine the degree of adjustment has to be undertaken with caution. An advantage of the relative entropy approach (as well as guaranteeing a minimal adjustment according to the criterion) is that the initial consensus estimate plays a significantly lesser role in the adjustment procedure. A

feature of Table 3 is the spread of the consensus estimates from −73.2 nm to −76.4 nm, a range of 3.2 nm that can be compared to a standard uncertainty of the order of 4 nm. This shows that the method of adjustment can have a significant impact on the reported consensus value.

6.3 Bayesian approaches to adjustment

The adjustment procedures we have considered so far have the feature that the input uncertainties are adjusted to achieve a desired degree of consistency and the adjusted uncertainties are then treated as known reliably as far as the determination of the consensus value is concerned. It seems more reasonable to regard the input uncertainties as estimates of the true uncertainties and then use the ensemble of information provided by the ILC to update these estimates in a Bayesian framework. Such an approach recognises that the updated estimates are not known exactly and this lack of complete knowledge about them will contribute an uncertainty to the consensus value derived from them. A Bayesian version of the Birge and M-P procedures is considered in [18], for example, and leads to t-distributions associated with the consensus value. The Bayesian approach can be developed further [21] in a hierarchical model that leads to a consensus distribution that is a mixture of t-distributions (in the same way that a t-distribution is a mixture of Gaussians). The analysis of the gauge block data in [21] led to a consensus estimate of −73.3 nm with associated standard uncertainty of 3.8 nm which is more in line with the relative entropy estimate in Table 3. We must bear in mind that the relative entropy approach associates a Gaussian distribution to the consensus estimate while the Bayesian approaches assign heavier-tailed distributions.

7 Concluding remarks

This paper has been concerned with the analysis of ILC data and approaches to resolving inconsistency in such data. When data is judged to be inconsistent, the WLS estimate of the consensus value of the quantity being measured might not be reliable. Several adjustment schemes have been discussed. The underlying assumption behind adjustment procedures is that some or all of the laboratories have underestimated or neglected some uncertainty contributions, sometimes referred to as 'dark uncertainty', and the adjusted uncertainties provide an estimate of this dark uncertainty derived from the complete set of laboratory results. We have been interested in procedures that make minimal adjustments to the input uncertainties. One way of quantifying the extent of adjustment is the Kullback–Leibler divergence or relative entropy. We have compared adjustment schemes such as the Birge and M-P methods with a S-D adjustment and relative entropy adjustment that are specifically designed to minimise the degree of adjustment. The Birge and M-P methods tend to make large adjustments compared to those made using a minimal relative entropy adjustment. The S-D method tends to compare well with the relative entropy adjustment and is not as numerically demanding as it involves only solving a single nonlinear equation. However, the S-D method is dependent on the initial consensus estimate. Adopting a stepped approach could help alleviate this issue. An advantage of the relative entropy approach (as well as guaranteeing a minimal adjustment according to the criterion) is that the initial consensus estimate, determined from inconsistent data, plays a significantly lesser role in the adjustment procedure.

This work was supported by the UK's National Measurement System programme for Data Science. We thank our colleague Dr. Peter Harris, NPL, for comments on a draft of this paper.

References

1. M.G. Cox, The evaluation of key comparison data, Metrologia **39**, 589–595 (2002)
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, W. Wöger, Analysis of key comparison data and laboratory biases, Metrologia **45**, 211–216 (2008)
4. M.G. Cox, A.B. Forbes, J. Flowers, P.M. Harris, Least squares adjustment in the presence of discrepant data, in *Advanced Mathematical and Computational Tools in Metrology VI*, edited by P. Ciarlini, M.G. Cox, F. Pavese, G.B. Rossi (World Scientific, Singapore, 2004), pp. 37–51
5. M.G. Cox, The evaluation of key comparison data: determining the largest consistent subset, Metrologia **44**, 187–200 (2007)
6. C. Elster, B. Toman, Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using Bayesian model averaging, Metrologia **47**, 113–119 (2010)
7. A.B. Forbes, C. Perruchet. Measurement systems analysis: concepts and computational approaches, in *IMEKO World Congress, Rio de Janeiro, September 18–22, 2006*
8. R.C. Paule, J. Mandel, Consensus values and weighting factors, J. Res. Natl. Bur. Stand. **87**, 377–385 (1982)
9. K. Weise, W. Woeger, Removing model and data non-conformity in measurement evaluation, Meas. Sci. Technol. **11**, 1649–1658 (2000)
10. R. Willink, Statistical determination of a comparison reference value using hidden errors, Metrologia **39**, 343–354 (2002)
11. S. Thompson, S. Ellison, Dark uncertainty, Accredit. Qual. Assur. **16**, 483–487 (2011)
12. J.O. Berger, in *Statistical decision theory and Bayesian analysis*, 2nd edn. (Springer, New York, 1985)
13. H. Chipman, E.I. George, R.E. McCulloch, *The practical implementation of Bayesian model selection* (Institute of Mathematical Statistics, Beachwood, Ohio, 2001)
14. A.B. Forbes, Traceable measurements using sensor networks, Trans. Mach. Learning Data Mining **8**, 77–100 (2015)

15. S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. **55**, 79–86 (1951)
16. D.J.C. MacKay, *Information theory, inference and learning algorithms* (Cambridge University Press, Cambridge, 2003)
17. P.E. Gill, W. Murray, M.H. Wright, *Practical optimization* (Academic Press, London, 1981)
18. O. Bodnar, C. Elster, J. Fischer, A. Possolo, B. Toman, Evaluation of uncertainty in the adjustment of fundamental constants, Metrologia **53**, S46–S54 (2016)
19. Thalmann R. CCL key comparison: calibration of gauge blocks by interferometry, Metrologia **39**, 165 (2002)
20. R. Thalmann. CCL key comparison: calibration of gauge blocks by interferometry, Metrologia **39**, 165–177 (2002)
21. A.B. Forbes, A hierarchical model for the analysis of inter-laboratory comparison data, Metrologia **53**, 1295–1305 (2016)

Cite this article as: Kavya Jagan, Alistair B. Forbes, Assessing interlaboratory comparison data adjustment procedures, Int. J. Metrol. Qual. Eng. **10**, 3 (2019)