Metrologia **45** (2008) 512–523 doi:10.1088/0026-1394/45/5/004

Bayesian posterior predictive *p*-value of statistical consistency in interlaboratory evaluations

Raghu N Kacker[1], Alistair Forbes[2], Rüdiger Kessel[1] and Klaus-Dieter Sommer[3]

[1] National Institute of Standards and Technology, Gaithersburg, MD 20899-8910, USA
[2] National Physical Laboratory, Teddington, Middlesex, TW11 0LW, UK
[3] Physikalisch-Technische Bundesanstalt, D-38116 Braunschweig, Germany

E-mail: raghu.kacker@nist.gov, alistair.forbes@npl.co.uk, ruediger.kessel@nist.gov and klaus-dieter.sommer@ptb.de

Received 11 April 2008
Published 17 September 2008
Online at stacks.iop.org/Met/45/512

Abstract

The results from an interlaboratory evaluation are said to be statistically consistent if they fit a normal (Gaussian) consistency model which postulates that the results have the same unknown expected value and stated variances–covariances. A modern method for checking the fit of a statistical model to the data is posterior predictive checking, which is a Bayesian adaptation of classical hypothesis testing. In this paper we propose the use of posterior predictive checking to check the fit of the normal consistency model to interlaboratory results. If the model fits reasonably then the results may be regarded as statistically consistent. The principle of posterior predictive checking is that the realized results should look plausible under a posterior predictive distribution. A posterior predictive distribution is the conditional distribution of potential results, given the realized results, which could be obtained in contemplated replications of the interlaboratory evaluation under the statistical model. A systematic discrepancy between potential results obtained from the posterior predictive distribution and the realized results indicates a potential failing of the model. One can investigate any number of potential discrepancies between the model and the results. We discuss an overall measure of discrepancy for checking the consistency of a set of interlaboratory results. We also discuss two sets of unilateral and bilateral measures of discrepancy. A unilateral discrepancy measure checks whether the result of a particular laboratory agrees with the statistical consistency model. A bilateral discrepancy measure checks whether the results of a particular pair of laboratories agree with each other. The degree of agreement is quantified by the Bayesian posterior predictive p-value. The unilateral and bilateral measures of discrepancy and their posterior predictive p-values discussed in this paper apply to both correlated and independent interlaboratory results. We suggest that the posterior predicative p-values may be used to assess unilateral and bilateral degrees of agreement in International Committee of Weights and Measures (CIPM) key comparisons.

1. Introduction

A question that is often asked about the results from an interlaboratory evaluation is whether they are statistically consistent. Indeed some interlaboratory evaluations are conducted for the primary purpose of determining whether statistically consistent results can be obtained. The summary data from an interlaboratory evaluation consists of n paired results and standard uncertainties $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$. The results $x_1, \ldots, x_n$ need not be the measured values of the same measurand; however, they are deemed to be suitable for comparison, otherwise a check of their consistency would be meaningless. The uncertainties $u(x_1), \ldots, u(x_n)$ are assumed to be non-zero and they may be unequal. The traditional

0026-1394/08/050512+12$30.00 © 2008 BIPM and IOP Publishing Ltd Printed in the UK

concept of statistical consistency motivated by the Birge test [1] is based on regarding the results $x_1, \ldots, x_n$ as realizations of random variables with sampling probability distributions with unknown expected values. To apply the Birge test, the standard uncertainties $u(x_1), \ldots, u(x_n)$ are regarded as the known standard deviations of those sampling distributions. A test of statistical consistency checks whether the expected values of the sampling probability distributions of the results may be regarded as approximately equal.

1.1. The Birge test of consistency

The Birge test [1, 2] is based on the following three assumptions: (i) The results $x_1, \ldots, x_n$ may be regarded as realizations of random variables, also denoted by $x_1, \ldots, x_n$, with sampling probability density functions (pdfs) of unknown expected values. (ii) The standard uncertainties $u(x_1), \ldots, u(x_n)$ may be regarded as the known standard deviations of the pdfs of $x_1, \ldots, x_n$, respectively. (iii) The sampling pdfs of $x_1, \ldots, x_n$ may be regarded as normal (Gaussian) and mutually independent. The following test statistic, called the Birge test statistic is calculated from the data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$:

$$R^2 = \sum_{i=1}^{n} w_i (x_i - x_{\mathrm{W}})^2/(n-1), \qquad (1)$$

where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, and $x_{\mathrm{W}} = \sum_i w_i x_i / \sum_i w_i$ is the weighted mean of the results $x_1, \ldots, x_n$. If the expected values $E(x_1), \ldots, E(x_n)$ of the results are all equal to some unknown value μ, then the expected value of the test statistic R^2 is one[4]. Thus if the calculated (realized) value of R^2 is substantially larger than one or equivalently the calculated value of $(n-1)$ $R^2 = \sum_i w_i (x_i - x_{\mathrm{W}})^2$ is substantially larger than $(n-1)$, then the results $x_1, \ldots, x_n$ are declared to be inconsistent.

To quantify the largeness of a calculated value of $(n-1)$ R^2, a statistical approach is needed. Suppose the calculated value of R^2 is R_0^2. Now suppose that $\chi_\nu^2[1-\alpha]$ is the 100 $(1-\alpha)$th percentile of the chi-square probability distribution, χ_ν^2, with degrees of freedom ν; that is, $\Pr\{\chi_\nu^2 \leqslant \chi_\nu^2 [1-\alpha]\} = 1-\alpha$, for $0 < \alpha < 1$. A traditional value of α is 0.05, which corresponds to the 95th percentile $\chi_\nu^2[0.95]$. The calculated value $(n-1)$ R_0^2 is compared with the 95th percentile $\chi_{(n-1)}^2[0.95]$ of the chi-square distribution $\chi_{(n-1)}^2$ with degrees of freedom $(n-1)$. The 95th percentile $\chi_{(n-1)}^2[0.95]$ is substantially larger than $(n-1)$; therefore, if $(n-1)$ R_0^2 is larger than $\chi_{(n-1)}^2[0.95]$ then it would be substantially larger than $(n-1)$. Suppose the calculated value $(n-1)$ R_0^2 is larger than $\chi_{(n-1)}^2[0.95]$; that is, the event $\{(n-1)R_0^2 > \chi_{(n-1)}^2[0.95]\}$ occurs. Then the results $x_1, \ldots, x_n$ are declared to be statistically inconsistent with 95% confidence. Statistical inconsistency implies that the expected values $E(x_1), \ldots, E(x_n)$ of the results may not be regarded as equal.

1.2. Statistical consistency defined as not excessive dispersion in the results

We had previously used in [3] the following definition of statistical consistency: the results $x_1, \ldots, x_n$ are said to be statistically consistent, relative to their stated variances $u^2(x_1), \ldots, u^2(x_n)$ and covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$, if their dispersion *is not greater than* what can be expected from a normal statistical consistency model which postulates that $x_1, \ldots, x_n$ have a joint n-variate normal sampling pdf with a common unknown expected value μ, known variances $u^2(x_1), \ldots, u^2(x_n)$, and known covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$. In the Birge test of consistency, all covariances are assumed to be zero.

In matrix form, the normal consistency model postulates that the random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ has an n-variate normal distribution, $N(\mathbf{1}\mu, \boldsymbol{D})$, with expected value $\mathbf{1}\mu$ and variance–covariance matrix (dispersion matrix) $\boldsymbol{D}$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$, the variances $u^2(x_1), \ldots, u^2(x_n)$ are diagonal elements of $\boldsymbol{D}$, and the covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$ are off-diagonal elements of $\boldsymbol{D}$. The superscript t in the definitions of $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ and $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$ indicates transpose of a vector or of a matrix. We can express the normal consistency model as the linear statistical model[5]:

$$\boldsymbol{x} = \mathbf{1}\mu + \boldsymbol{e}, \qquad \text{where } \boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{D}). \qquad (2)$$

By the relational symbol $\sim$ used in (2) we mean that the random vector $\boldsymbol{e} = (e_1, \ldots, e_n)^{\mathrm{t}}$ has the joint n-variate normal probability distribution $N(\mathbf{0}, \boldsymbol{D})$. Using the notation $u(x_i, x_i) = u^2(x_i)$ for $i = 1, 2, \ldots, n$, we can express the variance–covariance matrix $\boldsymbol{D}$ as $[u(x_i, x_j)]$. In the linear statistical model (2), the dispersion matrix $\boldsymbol{D}$ is assumed to be *known* and *positive definite* [3]. In terms of the model (2), we had previously used in [3] the following definition of statistical consistency.

Definition 1. The results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ are said to be statistically consistent relative to the variance–covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$, if their dispersion *is not greater than* what can be expected from the normal consistency model (2).

1.3. Statistical interpretation of the Birge test and its generalized version

Reference [3] interprets the Birge test of consistency as a classical test of the null hypothesis H_0 that the variances of the sampling pdfs of the results $x_1, \ldots, x_n$ are less than or equal to their stated values $u^2(x_1), \ldots, u^2(x_n)$ against the alternative hypothesis H_1 that the variances of the sampling pdfs of $x_1, \ldots, x_n$ are greater than $u^2(x_1), \ldots, u^2(x_n)$. A modern statistical protocol for hypothesis testing is to calculate the classical p-value. The classical p-value is the maximum probability under the null hypothesis H_0 of realizing in conceptual replications of the interlaboratory evaluation a

[4] This particular statement does not require that the forms of the sampling distributions of $x_1, \ldots, x_n$ be normal. However, the normal distribution is required to make a probabilistic statement about statistical consistency.

[5] Statisticians often express a general linear statistical model as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, where $\boldsymbol{e} \sim N(\mathbf{0}, \tau^2\boldsymbol{\Sigma})$. The model (2) is a special case in which $\boldsymbol{X} = \mathbf{1}$, $\boldsymbol{\beta} = \mu$, $\boldsymbol{\Sigma} = \boldsymbol{D}$ and $\tau^2 = 1$.

value of the test statistic R^2 that is equal to or larger than its realized (observed, calculated) value. The null hypothesis H_0 is rejected when the p-value is too small. Reference [3] gives an expression for the classical p-value of the calculated Birge test statistic R_0^2. Then it is shown in [3] that the classical p-value of the Birge test statistic R^2 is equal to the Bayesian posterior probability corresponding to the null hypothesis of statistical consistency based on non-informative improper prior distributions for the unknown statistical parameters.

Occasionally the interlaboratory results are correlated and it is necessary to check for their consistency. Reference [3] presents a general test of consistency for both uncorrelated and correlated results, of which the Birge test is a special case. Then it is shown in [3] that the classical p-value of the general test statistic is equal to the Bayesian posterior probability of the null hypothesis based on using non-informative prior distributions. The general test makes it possible to check the consistency of correlated results from interlaboratory evaluations.

1.4. Statistical consistency defined as fitting the normal (Gaussian) consistency model

In this paper we propose an improved definition of statistical consistency in interlaboratory results than the definition used earlier in [3]. Suppose the expected values $E(x_1), \ldots, E(x_n)$ are unequal; that is, the differences between them are of practical importance. If the variances $u^2(x_1), \ldots, u^2(x_n)$ were stated to be too large and consequently the calculated value of the Birge test statistic R_0^2 were sufficiently small, then the results $x_1, \ldots, x_n$ would appear to be statistically consistent. This apparent statistical consistency would be an artefact of stating the variances $u^2(x_1), \ldots, u^2(x_n)$ to be too large. Thus, if a metrologist overstates the uncertainties then he (she) may mask inequality of the expected values $E(x_1), \ldots, E(x_n)$; that is, wrongly declare the results with unequal expected values to be consistent.

A review of the Birge test in [2] notes that if the value of R_0^2 is substantially less than one[6], then the stated variances $u^2(x_1), \ldots, u^2(x_n)$ may well be too large. To prevent unreal pronouncements of statistical consistency arising from excessively overstating the variances, in this paper we use the following definition of consistency.

Definition 2. The results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ are said to be statistically consistent relative to the variance–covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$, if the normal consistency model (2), *reasonably fits* the results $x_1, \ldots, x_n$.

A modern Bayesian method for checking the fit of a statistical model to the data is posterior predictive checking [4, chapter 6]. In this paper we use posterior predictive checking to check the fit of the normal consistency model (2) to the results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$. If the model (2) reasonably fits, then the results $x_1, \ldots, x_n$ may be regarded as statistically consistent.

1.5. Bayesian posterior predictive checking of the fit of statistical consistency model

Concept of posterior predictive checking. The concept of posterior predictive checking of the fit of a statistical model to the realized data is a Bayesian adaptation of the classical (frequentist sampling) method of hypothesis testing [4]. A statistical model is a description of the sampling probability distribution attributed to the data; it describes the probabilities of obtaining various data values conditional on the values of certain parameters in contemplated replications of the data generation process. The aim in classical hypothesis testing is to assess whether the unknown values of the parameters of a sampling distribution belong to a set specified by the null hypothesis. A suitable test statistic is determined. A test statistic is a criterion to check the discrepancy between the realized data and the other data which might be obtained under the set of parameter values specified by the null hypothesis. The classical p-value of a test statistic is the maximum probability of obtaining a value of the test statistic more extreme than its realized value in contemplated replications according to the sampling distributions under the null hypothesis. A small classical p-value indicates that the realized data have low probability of occurrence under the sampling distributions specified by the null hypothesis. Therefore if the classical p-value is too small then the null hypothesis is rejected.

All Bayesian statistical inferences are conditional on the realized data. A posterior predictive distribution is like a sampling distribution of the data except that it is conditioned on the realized results rather than conditioned on unknown or hypothesized values of certain statistical parameters. A posterior predictive distribution of potential data is the integral of the sampling distribution with respect to the Bayesian posterior distributions of the parameters conditioned on the realized data. A discrepancy measure is a measure of the discrepancy between the statistical model and the data. It plays the same role in Bayesian posterior predictive model checking that a test statistic plays in classical hypothesis testing. The posterior predictive p-value of a discrepancy measure is the probability of obtaining a value of the discrepancy measure more extreme than its realized value in contemplated replications according to the posterior predictive distribution conditioned on the realized data. An extreme posterior predictive p-value (one which is close to 0 or close to 1) indicates that the realized data have a low probability of occurring under the postulated statistical model. That is, the statistical model does not appear to fit the realized data.

In this section we discuss application of the concept of posterior predictive checking to assess the fit of the statistical consistency model (2) to the results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$. Let $\boldsymbol{x}^{\mathrm{rep}} = (x_1^{\mathrm{rep}}, \ldots, x_n^{\mathrm{rep}})^{\mathrm{t}}$ represent potential results that could be obtained in a contemplated replication of the interlaboratory evaluation under the normal consistency model (2). The sampling distributions of $\boldsymbol{x}^{\mathrm{rep}}$ and $\boldsymbol{x}$, conditional on the unknown parameter μ, are identical; that is, $\boldsymbol{x} \sim N(\mathbf{1}\mu, \boldsymbol{D})$ as well as $\boldsymbol{x}^{\mathrm{rep}} \sim N(\mathbf{1}\mu, \boldsymbol{D})$. The values of the parameters μ and $\boldsymbol{D} = [u(x_i, x_j)]$ remain unchanged in contemplated replications. The state of knowledge concerning the

[6] If the value R_0^2 is substantially larger than one, then either the expected values $E(x_1), \ldots, E(x_n)$ are not equal or the stated variances $u^2(x_1), \ldots, u^2(x_n)$ are too small.

unknown value of μ is described by the Bayesian posterior probability distribution of μ conditional on the realized results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\rm t}$. When there is no *a priori* knowledge about the value of μ, a non-informative improper prior distribution is used to determine the posterior distribution of μ conditional on $\boldsymbol{x}$.

Posterior predictive distribution of the potential results. The posterior predictive distribution of $\boldsymbol{x}^{\rm rep}$ is the conditional distribution of $\boldsymbol{x}^{\rm rep}$ given the realized results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\rm t}$. Let $f\,(\boldsymbol{x}^{\rm rep}|\mu)$ be the sampling pdf of $\boldsymbol{x}^{\rm rep}$ conditional on the unknown parameter μ postulated by the normal consistency model (2). Now suppose, $p(\mu|\boldsymbol{x})$ is the posterior pdf of μ given the realized results $\boldsymbol{x}$, then the posterior predictive pdf of $\boldsymbol{x}^{\rm rep}$ given $\boldsymbol{x}$, $p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x})$, is the integral

$$p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x}) = \int f\,(\boldsymbol{x}^{\rm rep}|\mu)\,p(\mu|\boldsymbol{x})\,{\rm d}\mu. \qquad (3)$$

The posterior predictive pdf $p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x})$ is the average of the conditional pdfs $f\,(\boldsymbol{x}^{\rm rep}|\mu)$ of $\boldsymbol{x}^{\rm rep}$ over the posterior distribution of μ. It is the prediction of the potential results $\boldsymbol{x}^{\rm rep}$ that could be obtained in contemplated replications of the interlaboratory evaluation under the normal consistency model (2) conditioned on the realized results $\boldsymbol{x}$.

If the statistical model (2) fits, then the replicated results $\boldsymbol{x}^{\rm rep}$ obtained from this model should look similar to the realized results $\boldsymbol{x}$ [4, section 6.3]. Another way of saying this is that the realized results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\rm t}$ should look plausible under the posterior predictive pdf $p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x})$. A systematic discrepancy between the results $\boldsymbol{x}^{\rm rep}$ obtained from the posterior predictive pdf and the realized results $\boldsymbol{x}$ indicates a failing of the statistical model (2).

Discrepancy measures. A discrepancy measure $\mathrm{T}(\boldsymbol{x}^{\rm rep})$ is a measure of the discrepancy that one wishes to check between the statistical model and the data [4, section 6.3]. Unlike a classical test statistic which depends only on the data, a Bayesian discrepancy measure may depend in addition to the data on the model parameters under their posterior distribution. Since, in model (2) there is only one unknown parameter; therefore, the discrepancy measures depend only on the data, such as classical test statistics. We will discuss several discrepancy measures for checking the fit of statistical consistency model (2) to the results $\boldsymbol{x}$.

Posterior predictive p-values of discrepancy measures. The Bayesian posterior predictive p-value, $p_{\rm P}$, of a realized value $\mathrm{T}(\boldsymbol{x})$ of the discrepancy measure $\mathrm{T}(\boldsymbol{x}^{\rm rep})$ is the probability of realizing in contemplated replications a value of the discrepancy measure $\mathrm{T}(\boldsymbol{x}^{\rm rep})$ more extreme than its realized value $\mathrm{T}(\boldsymbol{x})$; that is,

$$p_{\rm P} = \Pr\{T\,(\boldsymbol{x}^{\rm rep}) \geqslant T\,(\boldsymbol{x})|\boldsymbol{x}\}, \qquad (4)$$

where the probability is defined with respect to the posterior predictive distribution of $\boldsymbol{x}^{\rm rep}$ conditioned on the realized results $\boldsymbol{x}$. The statistical model (2) is suspect if the posterior predictive p-value $p_{\rm P}$ of a discrepancy measure $\mathrm{T}(\boldsymbol{x})$ is extreme (that is, close to 0 or close to 1), thereby indicating that the realized results $\boldsymbol{x}$ are not very likely to be seen in contemplated replications if the statistical model (2) were true.

1.6. Outline

In section 2, we discuss an overall measure of discrepancy for checking the statistical consistency of interlaboratory results. Then we determine the posterior predictive p-value of obtaining a value of the overall discrepancy measure more extreme than its realized value. In section 3, we discuss two sets of unilateral and bilateral measures of discrepancy. A unilateral discrepancy measure checks whether the result from a particular laboratory agrees with the statistical consistency model. A bilateral discrepancy measure checks whether the results from a particular pair of laboratories agree with each other. Then we determine the posterior predictive p-values of obtaining values of unilateral and bilateral discrepancy measures more extreme than their realized values. A p-value close to zero or close to one suggests discrepancy. In section 4, we illustrate the calculation of posterior predicative p-values of discrepancy measures. In section 5, we suggest that the posterior predicative p-values may be used to assess the degrees of agreement in the results from an International Committee of Weights and Measures (CIPM) key comparison. A brief summary appears in section 6.

2. Overall measure of discrepancy for checking consistency

The normal consistency model (2) postulates that the sampling pdf $f\,(\boldsymbol{x}|\mu)$ of $\boldsymbol{x}$ given μ is

$$f\,(\boldsymbol{x}|\mu) = (2\pi)^{-n/2}|\boldsymbol{D}|^{-1/2} \times \exp\{-\tfrac{1}{2}(\boldsymbol{x} - \mathbf{1}\mu)^{\rm t}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mathbf{1}\mu)\}. \qquad (5)$$

We will determine a Bayesian posterior pdf $p(\mu|\boldsymbol{x})$ for the unknown parameter μ in (5) using a non-informative improper prior distribution for μ. Then we will use the integral (3) to determine the posterior predictive pdf $p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x})$ of $\boldsymbol{x}^{\rm rep}$ given $\boldsymbol{x}$. We will introduce an overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}^{\rm rep})$ for checking the fit of the normal consistency model (2) to the realized results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\rm t}$. It turns out that the posterior predictive pdf of the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}^{\rm rep})$ has a simple and well-known form. Thus, the posterior predictive p-value, $p_{\rm P}$, of the realized value of the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x})$ can be analytically determined.

2.1. Bayesian posterior distribution of the common expected value

The generalized least squares estimate (GLSE) m of the single unknown parameter μ in (5) is that value m for which the quadratic form $(\boldsymbol{x} - \mathbf{1}\mu)^{\rm t}\boldsymbol{D}^{-1}$ $(\boldsymbol{x} - \mathbf{1}\mu)$ is minimum. The GLSE m has the following properties, which are special cases (corresponding to $\tau^2 = 1$) of the properties derived in [3, appendices B and E]:

(i) The GLSE estimate[7] m of μ in (5) is $m = \boldsymbol{B}^{\rm t}\boldsymbol{x}$ where $\boldsymbol{B}^{\rm t} = (\mathbf{1}^{\rm t}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\rm t}\boldsymbol{D}^{-1}$.

[7] The minimization of the quadratic form $(\boldsymbol{x} - \mathbf{1}\mu)^{\rm t}\boldsymbol{D}^{-1}$ $(\boldsymbol{x} - \mathbf{1}\mu)$ in (5) is equivalent to maximization of the pdf $f\,(\boldsymbol{x}|\mu)$ interpreted as a likelihood function of μ. Therefore, $m = \boldsymbol{B}^{\rm t}\boldsymbol{x}$ is both the GLSE and the maximum likelihood estimate (MLE) of μ.

(*ii*) The sampling distribution of $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}$ is normal with expected value $E(m) = \mu$ and variance $V(m) = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$.

(*iii*) The quadratic form $(\boldsymbol{x} - \mathbf{1}\mu)^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x} - \mathbf{1}\mu)$ in (5) can be parsed as

$$(\boldsymbol{x} - \mathbf{1}\mu)^{\mathrm{t}} D^{-1}(\boldsymbol{x} - \mathbf{1}\mu) = (\boldsymbol{x} - \mathbf{1}m)^{\mathrm{t}} D^{-1}(\boldsymbol{x} - \mathbf{1}m) + \frac{(m-\mu)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}. \qquad (6)$$

This expression shows that the minimum value[8] of the quadratic form $(\boldsymbol{x} - \mathbf{1}\mu)^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x} - \mathbf{1}\mu)$ is $(\boldsymbol{x} - \mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x} - \mathbf{1}m)$.

(*iv*) The sampling distribution of the minimum quadratic form $(\boldsymbol{x}-\mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}-\mathbf{1}m)$ is the chi-square distribution $\chi^2_{(n-1)}$ with degrees of freedom $(n - 1)$.

(*v*) The sampling distributions of m and $(\boldsymbol{x}-\mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x}-\mathbf{1}m)$ are independent.

The estimate m is the minimum variance unbiased estimate of μ in the sense that it is unbiased, that is $E(m) = \mu$, and it has the smallest variance among all unbiased estimates of μ [5, section 5a.2]. Thus it is a statistically optimum estimate of the parameter μ in model (2).

As discussed in appendix A, the Bayesian posterior distribution of μ given $\boldsymbol{x}$, based on using a well-known non-informative improper prior distribution $p(\mu)$ for μ, is normal, $N(m, (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1})$, with expected value $m = \boldsymbol{B}^{\mathrm{t}}\ \boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and variance $(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$ having the pdf

$$p(\mu|\boldsymbol{x}) = (2\pi)^{-1/2}[(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}]^{-1/2} \times \exp\left\{-\frac{1}{2}\frac{(\mu-m)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right\}. \qquad (7)$$

2.2. *Posterior predictive distribution of the results*

The sampling pdf $f(\boldsymbol{x}^{\mathrm{rep}}|\mu)$ of $\boldsymbol{x}^{\mathrm{rep}}$ conditional on the unknown parameter μ postulated by the normal consistency model (2) is

$$f(\boldsymbol{x}^{\mathrm{rep}}|\mu) = (2\pi)^{-n/2}|\boldsymbol{D}|^{-1/2} \times \exp\{-\tfrac{1}{2}(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\mu)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\mu)\}. \qquad (8)$$

As discussed in appendix B, the posterior predictive distribution of $\boldsymbol{x}^{\mathrm{rep}}$ conditional on the given results $\boldsymbol{x}$ determined from the integral (3) is the n-variate normal distribution with the pdf

$$p(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2} \times \exp\{-\tfrac{1}{2}(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\boldsymbol{B}^{\mathrm{t}}\boldsymbol{x})\boldsymbol{V}^{-1}(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\boldsymbol{B}^{\mathrm{t}}\boldsymbol{x})\}, \qquad (9)$$

having the expected value $E(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \mathbf{1}\boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and the variance–covariance matrix $\boldsymbol{V}(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \boldsymbol{V} = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$. Since the results $\boldsymbol{x}$ are known, both parameters, expected value and variance, of the posterior predictive pdf $p(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x})$ are known.

2.3. *Overall measure of discrepancy and its posterior predictive p-value*

A generic measure of discrepancy between the model (2) and the results $\boldsymbol{x}^{\mathrm{rep}}$ (motivated by generalized least squares theory [5]) is the quadratic form $(\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}\mu)^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}\mu)$, which depends on the unknown parameter μ. Therefore we use its minimum value $(\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}m^{\mathrm{rep}})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}m^{\mathrm{rep}})$ as an overall measure of discrepancy, $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}})$, where $m^{\mathrm{rep}} = \boldsymbol{B}^{\mathrm{t}}\ \boldsymbol{x}^{\mathrm{rep}}$ and $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}$ [3, section 2]; thus, $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}}) = (\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}m^{\mathrm{rep}})^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}m^{\mathrm{rep}})$. The realized value of the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}})$, calculated from the results $\boldsymbol{x}$, is $\mathrm{T_c}(\boldsymbol{x}) = (\boldsymbol{x} - \mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mathbf{1}m)$, where $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$.

We show in appendix C that the posterior predictive distribution of the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on the realized results $\boldsymbol{x}$ is the chi-square distribution $\chi^2_{(n-1)}$ with degrees of freedom $(n-1)$. In symbols,

$$p(T_{\mathrm{c}}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) \sim \chi^2_{(n-1)}. \qquad (10)$$

Therefore the posterior predictive p-value p_{P} of $\mathrm{T}(\boldsymbol{x})$ is

$$p_{\mathrm{P}} = \Pr\{T_{\mathrm{c}}(\boldsymbol{x}^{\mathrm{rep}}) \geqslant T_{\mathrm{c}}(\boldsymbol{x})|\boldsymbol{x}\} = \Pr\{\chi^2_{(n-1)} \geqslant T_{\mathrm{c}}(\boldsymbol{x})\}. \qquad (11)$$

Thus, to check the fit of model (2) to the results $\boldsymbol{x}$, we calculate the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}) = (\boldsymbol{x} - \mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}\ (\boldsymbol{x} - \mathbf{1}m)$ where $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$. Then we determine the posterior predictive p-value of $\mathrm{T_c}(\boldsymbol{x})$ from the expression (11). A posterior predictive p-value p_{P} that is extreme (close to 0 or 1) indicates that the statistical consistency model (2) does not fit the results $\boldsymbol{x}$; that is, the results may not be regarded as statistically consistent.

A posterior predictive p-value close to 0 (say, less than 0.05) indicates that either the expected values $E(x_1), \ldots, E(x_n)$ are not equal or the stated variances $u^2(x_1), \ldots, u^2(x_n)$ are too small. If the stated variances are believed to be reliable then one may conclude that the expected values $E(x_1), \ldots, E(x_n)$ appear to be unequal. A posterior predictive p-value close to 1 (say, greater than 0.95) indicates that the stated variances $u^2(x_1), \ldots, u^2(x_n)$ may be too large; thus, they are not a reliable basis for assessing the equality of the expected values $E(x_1), \ldots, E(x_n)$. Thus a posterior predictive p-value that is close to 0 or close to 1 indicates that the results $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{t}}$ may not be regarded as consistent.

2.4. *Birge test of consistency for independent results*

If the results $x_1, \ldots, x_n$ are mutually independent then the covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$ are all zero. Thus the variance–covariance matrix $\boldsymbol{D}$ reduces to the diagonal matrix $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$ with inverse $\boldsymbol{D}^{-1} = \mathrm{Diag}[w_1, \ldots, w_n]$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$. Thus, $\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1} = \sum_i w_i$, $(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1} = 1/\sum_i w_i$, and $\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1} = \sum_i w_i x_i$, and $\boldsymbol{x}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \sum_i w_i x_i^2$. The GLSE of μ reduces to $m = \boldsymbol{B}^{\mathrm{t}}\ \boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \sum_i w_i x_i / \sum_i w_i = x_{\mathrm{W}}$, the weighted mean. Consequently, the sampling distribution of $m = x_{\mathrm{W}}$ is normal with expected value μ and variance $V(x_{\mathrm{W}}) = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1} = 1/\sum_i w_i$. We will use the symbol $u^2(x_{\mathrm{W}})$ for the variance $V(x_{\mathrm{W}}) = 1/\sum_i w_i$ of the weighted mean x_{W}.

[8] The minimum quadratic form $(\boldsymbol{x} - \mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x} - \mathbf{1}m)$ is the residual sum of squares for the model (2) [5, table 4.a.7, column 4].

When the results are independent, the quadratic form $(\boldsymbol{x}-\boldsymbol{1}\mu)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}-\boldsymbol{1}\mu)$ reduces to $\sum_i (x_i-\mu)^2/u^2(x_i) = \sum_i w_i (x_i-\mu)^2$ with minimum value $(\boldsymbol{x}-\boldsymbol{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}-\boldsymbol{1}m) = \sum_i w_i (x_i - x_{\mathrm{W}})^2 = (n-1)R^2$. Thus, the sampling distribution of $\sum_i w_i\ (x_i - x_{\mathrm{W}})^2 = (n-1)R^2$ is the $\chi^2_{(n-1)}$ distribution. Also, the sampling distributions of x_{W} and $(n-1)\ R^2$ are independent [3, appendix B].

If the results $x_1, \ldots, x_n$ are mutually independent, then the posterior pdf $p(\mu|\boldsymbol{x})$ given in (7) reduces to

$$p(\mu|\boldsymbol{x}) = (2\pi)^{-1/2}[u^2(x_{\mathrm{W}})]^{-1/2} \exp\left\{-\frac{1}{2}\frac{(\mu - x_{\mathrm{W}})^2}{u^2(x_{\mathrm{W}})}\right\}. \tag{12}$$

Further, the posterior predictive distribution $p(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x})$ of $\boldsymbol{x}^{\mathrm{rep}}$ conditional on the results $\boldsymbol{x}$ given in (9) reduces to an n-variate normal distribution, $N(\boldsymbol{1}x_{\mathrm{W}}, \boldsymbol{D}+u^2(x_{\mathrm{W}})\boldsymbol{1}\boldsymbol{1}^{\mathrm{t}})$, with the expected value $E(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \boldsymbol{1}(\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \boldsymbol{1}x_{\mathrm{W}}$ and the variance–covariance matrix $V(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \boldsymbol{D}+\boldsymbol{1}(\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\mathrm{t}} = \boldsymbol{D} + u^2(x_{\mathrm{W}})\boldsymbol{1}\boldsymbol{1}^{\mathrm{t}}$, where $\boldsymbol{D} = \mathrm{Diag}[u^2(x_1), \ldots, u^2(x_n)]$.

The overall discrepancy measure introduced in section 2.3 reduces to $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}}) = \sum_i w_i\ (x_i^{\mathrm{rep}} - x_{\mathrm{W}}^{\mathrm{rep}})^2$, where $x_{\mathrm{W}}^{\mathrm{rep}} = \sum_i w_i x_i^{\mathrm{rep}}/\sum_i w_i$ [3, appendix B]. The realized value of this discrepancy measure is $\mathrm{T_c}(\boldsymbol{x}) = \sum_i w_i(x_i - x_{\mathrm{W}})^2 = (n-1)R_0^2$, where R_0^2 is the realized (calculated) value of the Birge test statistic (1). The posterior predictive distribution of the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}}) = \sum_i w_i(x_i^{\mathrm{rep}} - x_{\mathrm{W}}^{\mathrm{rep}})^2$ is the chi-square distribution $\chi^2_{(n-1)}$ with degrees of freedom $(n-1)$ (section 2.3). The posterior predictive p-value of the realized discrepancy measure $\mathrm{T_c}(\boldsymbol{x}) = \sum_i w_i(x_i - x_{\mathrm{W}})^2 = (n-1)R_0^2$ is

$$p_{\mathrm{P}} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)R_0^2\}. \tag{13}$$

As discussed in [3, appendix D], the event $\{(n-1)R_0^2 > \chi^2_{(n-1)}[0.95]\}$ is equivalent to the event that $p_{\mathrm{P}} = \Pr\{\chi^2_{(n-1)} \geqslant (n-1)\ R_0^2\} < 0.05$. Thus a comparison of the posterior predictive p-value p_{P} relative to 0.05 is equivalent to the Birge test of consistency discussed in section 1.1.

3. Unilateral and bilateral measures of discrepancy for individual laboratories

A great advantage of the posterior predictive checking is that there is no limit on the number of potential discrepancies between the statistical model and the data that may be investigated. For example, one may investigate whether the result from a particular laboratory agrees with the statistical consistency model (2) and whether the results from a particular pair of laboratories agree with each other [10]. The result x_i from the laboratory labelled i agrees with the statistical model (2) if the difference between x_i and the prediction of x_i based on the model is not too large in view of the stated variance of the difference, for $i = 1, 2, \ldots, n$. The results x_i and x_j from the laboratories labelled i and j agree with each other if their difference is not too large in view of the stated variance of the difference, for $i,\ j = 1, 2, \ldots, n$ and $i \neq j$. In this section we introduce two sets of unilateral and bilateral discrepancy measures which are suitable for investigating such discrepancies.

A basic statistical measure of discrepancy between a statistical model and the data is the vector of residuals which are differences between the data and their predicted values determined by replacing the unknown statistical parameters in the model by their estimates [11, chapter 3]. If all residuals are zero, the model fits the data perfectly. If the residuals show a random pattern, then the model may be useful for statistical prediction. On the other hand, a systematic pattern in the residuals indicates inadequacies of the model in fitting the data. The unilateral and bilateral discrepancy measures discussed below are linear functions of the residuals.

We will determine the residuals and their posterior predictive distributions. Then we will discuss the discrepancy measures. A statistically optimum estimate of the parameter μ in the model (2) is $m^{\mathrm{rep}} = (\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}^{\mathrm{rep}}$ (section 2.1). The corresponding prediction of the potential result x_i^{rep} is m^{rep}, for $i = 1, 2, \ldots, n$. The difference $r_i(\boldsymbol{x}^{\mathrm{rep}}) = x_i^{\mathrm{rep}} - m^{\mathrm{rep}}$ between the result x_i^{rep} and its prediction m^{rep} is the residual, for $i = 1, 2, \ldots, n$. The residual $r_i(\boldsymbol{x}^{\mathrm{rep}})$ indicates discrepancy of the result x_i^{rep} from the model (2). The n residuals $r_i(\boldsymbol{x}^{\mathrm{rep}})$ form the vector $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}}) = (r_1(\boldsymbol{x}^{\mathrm{rep}}), \ldots, r_n(\boldsymbol{x}^{\mathrm{rep}}))^{\mathrm{t}} = (x_1^{\mathrm{rep}} - m^{\mathrm{rep}}, \ldots, x_n^{\mathrm{rep}} - m^{\mathrm{rep}})^{\mathrm{t}}$ with realized value $\boldsymbol{r}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{1}m) = (r_1(\boldsymbol{x}), \ldots, r_n(\boldsymbol{x}))^{\mathrm{t}} = (x_1 - m, \ldots, x_n - m)^t$, where $m = (\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$.

As discussed in appendix D, the posterior predictive distribution $p(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$ of the vector $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ of residuals is the n-variate normal distribution, $N(\boldsymbol{0}, \boldsymbol{D} - \boldsymbol{1}(\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\mathrm{t}})$, with expected value $E(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = \boldsymbol{0}$ and variance–covariance matrix $V(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = \boldsymbol{D} - \boldsymbol{1}(\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}^{\mathrm{t}}$. Since $(\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1} = V(m)$ (section 2.1); therefore, $V(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = \boldsymbol{D} - V(m)\boldsymbol{1}\boldsymbol{1}^{\mathrm{t}} = [u(x_i, x_j) - V(m)]$. Thus, $V(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$ is a matrix whose ijth element is $u(x_i, x_j) - V(m)$. It follows that the covariance of $r_i(\boldsymbol{x}^{\mathrm{rep}})$ and $r_j(\boldsymbol{x}^{\mathrm{rep}})$ is $C(r_i(\boldsymbol{x}^{\mathrm{rep}}), r_j(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = C(x_i^{\mathrm{rep}} - m^{\mathrm{rep}}, x_j^{\mathrm{rep}} - m^{\mathrm{rep}}|\boldsymbol{x}) = u(x_i, x_j) - V(m)$, and the variance of $r_i(\boldsymbol{x}^{\mathrm{rep}})$ is $V(r_i(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = V(x_i^{\mathrm{rep}} - m^{\mathrm{rep}}|\boldsymbol{x}) = u^2(x_i) - V(m)$ for $i, j = 1, 2, \ldots, n$. In symbols,

$$p(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) \sim N(\boldsymbol{0}, [u(x_i, x_j) - V(m)]), \tag{14}$$

where $u(x_i, x_j) - V(m)$ is the ijth element of the variance–covariance matrix $V(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$ and $V(m) = (\boldsymbol{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{1})^{-1}$ for $i, j = 1, 2, \ldots, n$.

3.1. Unilateral discrepancy measures

As a unilateral measure of discrepancy, denoted by $\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})$, between the result x_i^{rep} from the laboratory labelled i and the statistical consistency model (2), we can use the residual $r_i(\boldsymbol{x}^{\mathrm{rep}})$; that is, $\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}}) = r_i(\boldsymbol{x}^{\mathrm{rep}}) = x_i^{\mathrm{rep}} - m^{\mathrm{rep}}$, for $i = 1, 2, \ldots, n$. From (14), the posterior predictive distribution, $p(\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$, of $\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})$ conditional on the realized results $\boldsymbol{x}$ is normal, $N(0, u^2(x_i) - V(m))$, with expected value 0 and variance $u^2(x_i) - V(m)$; that is,

$$p(\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) \sim N(0, u^2(x_i) - V(m)). \tag{15}$$

Thus the posterior predictive distribution of the ratio $\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})/\sqrt{(u^2(x_i) - V(m))}$ conditional on $\boldsymbol{x}$ is the standard normal distribution with expected value 0 and variance 1.

The realized value of $\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})$ is $\mathrm{T}_i(\boldsymbol{x}) = r_i(\boldsymbol{x}) = (x_i - m)$, for $i = 1, 2, \ldots, n$. The posterior predictive p-value p_{P} of the realized discrepancy measure $\mathrm{T}_i(\boldsymbol{x}) = (x_i - m)$ is

$$p_{\mathrm{P}} = \Pr\{\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}}) \geqslant \mathrm{T}_i(\boldsymbol{x})|\boldsymbol{x}\}$$

$$= \Pr\left\{Z \geqslant \frac{(x_i - m)}{\sqrt{u^2(x_i) - V(m)}}\right\}, \tag{16}$$

where Z is a variable that has the standard normal distribution with expected value 0 and variance 1. If the results $x_1, \ldots, x_n$ are mutually independent then $m = x_{\mathrm{W}}$, and $V(m) = V(x_{\mathrm{W}}) = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1} = 1/\sum_i w_i$ denoted by $u^2(x_{\mathrm{W}})$. In that case the posterior predictive p-value p_{P} of the realized discrepancy measure $\mathrm{T}_i(\boldsymbol{x}) = r_i(\boldsymbol{x}) = (x_i - x_{\mathrm{W}})$ reduces to

$$p_{\mathrm{P}} = \Pr\left\{Z \geqslant \frac{(x_i - x_{\mathrm{W}})}{\sqrt{u^2(x_i) - u^2(x_{\mathrm{W}})}}\right\}, \tag{17}$$

for $i = 1, 2, \ldots, n$. A posterior predictive p-value that is close to 0 or close to 1 indicates that the result x_i from the laboratory labelled i does not agree with the statistical consistency model (2), for $i = 1, 2, \ldots, n$.

3.2. *Bilateral discrepancy measures*

As a bilateral measure of discrepancy, denoted by $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})$, between the results x_i^{rep} and x_j^{rep} from two particular laboratories labelled i and j, we can use the difference between the residuals $r_i(\boldsymbol{x}^{\mathrm{rep}})$ and $r_j(\boldsymbol{x}^{\mathrm{rep}})$; that is, $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}}) = T_i(\boldsymbol{x}^{\mathrm{rep}}) - T_j(\boldsymbol{x}^{\mathrm{rep}}) = r_i(\boldsymbol{x}^{\mathrm{rep}}) - r_j(\boldsymbol{x}^{\mathrm{rep}}) = (x_i^{\mathrm{rep}} - m^{\mathrm{rep}}) - (x_j^{\mathrm{rep}} - m^{\mathrm{rep}}) = (x_i^{\mathrm{rep}} - x_j^{\mathrm{rep}})$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. From (14), the posterior predictive distribution, $p(\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$, of $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})$, conditional on the realized results $\boldsymbol{x}$ is normal, $N(0, u^2(x_i) + u^2(x_j) - 2u(x_i, x_j))$, with expected value 0 and variance $u^2(x_i) + u^2(x_j) - 2u(x_i, x_j)$; that is

$$p(\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) \sim N(0, u^2(x_i) + u^2(x_j) - 2u(x_i, x_j)). \tag{18}$$

Thus the posterior predictive distribution of the ratio $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})/\sqrt{(u^2(x_i) + u^2(x_j) - 2u(x_i, x_j))}$ conditional on $\boldsymbol{x}$ is the standard normal distribution with expected value 0 and variance 1.

The realized value of $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})$ is $\mathrm{T}_{i-j}(\boldsymbol{x}) = r_i(\boldsymbol{x}) - r_j(\boldsymbol{x}) = (x_i - x_j)$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The posterior predictive p-value p_{P} of the realized discrepancy measure $\mathrm{T}_{i-j}(\boldsymbol{x}) = (x_i - x_j)$ is

$$p_{\mathrm{P}} = \Pr\{\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}}) \geqslant \mathrm{T}_{i-j}(\boldsymbol{x})|\boldsymbol{x}\}$$

$$= \Pr\left\{Z \geqslant \frac{(x_i - x_j)}{\sqrt{u^2(x_i) + u^2(x_j) - 2u(x_i, x_i)}}\right\}, \tag{19}$$

where Z is a variable that has the standard normal distribution with expected value 0 and variance 1. If the results $x_1, \ldots, x_n$ are mutually independent then all covariances $u(x_{1,}x_2), \ldots, u(x_{n-1}, x_n)$ are zero and the posterior predictive p-value p_{P} of the realized discrepancy measure $\mathrm{T}_{i-j}(\boldsymbol{x}) = r_i(\boldsymbol{x}^{\mathrm{rep}}) - r_j(\boldsymbol{x}^{\mathrm{rep}}) = (x_i - x_j)$ reduces to

$$p_{\mathrm{P}} = \Pr\left\{Z \geqslant \frac{(x_i - x_j)}{\sqrt{u^2(x_i) + u^2(x_j)}}\right\}, \tag{20}$$

for $i, j = 1, 2, \ldots, n$ and $i \neq j$. A posterior predictive p-value that is close to 0 or close to 1 indicates that the results x_i and x_j from the laboratories labelled i and j do not agree with each other, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. In interlaboratory evaluations between particular national measurement institutes (NMIs), complete bilateral consistency between all pairs may be required to assure that the international measurement system is working properly.

Table 1. Mean relative differences $x_1, \ldots, x_{16}$ from the BIPM measurements for the wavelength 514.536 nm and their associated standard uncertainties $u(x_1), \ldots, u(x_{16})$, reproduced from the BIPM Report [12, table 65, columns 6 and 7].

Indices i	$x_i \times 10^4$	$u(x_i) \times 10^4$
1	−0.20	1.30
2	1.10	1.70
3	2.00	1.40
4	−0.30	2.50
5	13.10	4.90
6	1.70	2.70
7	−11.00	6.80
8	0.00	2.20
9	0.30	1.30
10	−5.10	2.40
11	5.90	3.20
12	−1.10	2.60
13	1.30	1.10
14	5.30	3.40
15	2.90	2.90
16	−1.00	5.10

4. Calculation of posterior predictive *p*-values

To illustrate the calculation of posterior predictive p-values of the realized discrepancy measures, $\mathrm{T}_{\mathrm{c}}(\boldsymbol{x})$, $\mathrm{T}_i(\boldsymbol{x})$, and $\mathrm{T}_{i-j}(\boldsymbol{x})$ for $i, j = 1, 2, \ldots, n$ and $i \neq j$, we have used a small subset of the data from a BIPM Report [12] on the supplementary comparison CCPR-S3 of cryogenic radiometers carried out by the Consultative Committee for Photometry and Radiometry of the CIPM. A set of three transfer standard detectors was calibrated by the cryogenic radiometer of the laboratory labelled i (for $i = 1, 2, \ldots, n$, where $n = 16$) and also by the radiometer of the BIPM (which served as a reference laboratory). The result x_i is the arithmetic mean of the relative differences[9] between the 'responsivities' of the three detectors calibrated at the laboratory labelled i and the same three detectors calibrated at the BIPM for $i = 1, 2, \ldots, 16$. The responsivity of a detector depends on the wavelength for the laser source. The results $x_1, \ldots, x_{16}$ and the corresponding standard uncertainties $u(x_1), \ldots, u(x_{16})$ for the wavelength 514.536 nm from the BIPM Report [12, table 65, columns 6 and 7] are reproduced in table 1. For our discussion, the identities of the laboratories are not relevant so we display only the indices, $i = 1, 2, \ldots, 16$, for the laboratories. The results (mean relative differences) $x_1, \ldots, x_{16}$ are regarded in the BIPM Report [12] as mutually independent. Also, the

[9] If R_i and $\mathrm{R}_{\mathrm{BIPM}}$ are responsivities from the laboratories labelled i and the BIPM, respectively, for $i = 1, 2, \ldots, n$, then the relative difference is $(R_i - R_{\mathrm{BIPM}})/R_{\mathrm{BIPM}}$.

Table 2. Realized values of the unilateral discrepancy measures $\mathrm{T}_i(\boldsymbol{x}) = x_i - x_\mathrm{W}$, standard deviations $S(\mathrm{T}_i(\boldsymbol{x}^\mathrm{rep}))$, and posterior predictive p-values p_P for $i = 1, 2, \ldots, 16$.

Indices i	$\mathrm{T}_i(\boldsymbol{x}) \times 10^4$	$S(\mathrm{T}_i(\boldsymbol{x}^\mathrm{rep})) \times 10^4$	p_P
1	−1.01	1.20	0.80
2	0.29	1.63	0.43
3	1.19	1.31	0.18
4	−1.11	2.45	0.67
5	12.29	4.88	**0.01**
6	0.89	2.65	0.37
7	−11.81	6.78	**0.96**
8	−0.81	2.14	0.65
9	−0.51	1.20	0.66
10	−5.91	2.35	**0.99**
11	5.09	3.16	0.05
12	−1.91	2.55	0.77
13	0.49	0.98	0.31
14	4.49	3.36	0.09
15	2.09	2.86	0.23
16	−1.81	5.08	0.64

BIPM Report regards the variances $u^2(x_1), \ldots, u^2(x_{16})$ as the known variances of the sampling pdfs of the results.

The weighted mean of the $n = 16$ results $x_1, \ldots, x_{16}$ shown in table 1 is $x_\mathrm{W} = 0.81 \times 10^{-4}$ with standard uncertainty $u(x_\mathrm{W}) = 0.49 \times 10^{-4}$ units. The realized (calculated) value of the Birge test statistic (1) is $R_0^2 = 1.53$. The realized value of the overall discrepancy measure is $\mathrm{T}_\mathrm{c}(\boldsymbol{x}) = \sum_i w_i(x_i - x_\mathrm{W})^2 = (n-1)R_0^2 = 22.98$. The posterior predictive distribution of the overall discrepancy measure $\mathrm{T}_\mathrm{c}(\boldsymbol{x}^\mathrm{rep})$ conditional on the realized results $\boldsymbol{x} = (x_1, \ldots, x_{16})^\mathrm{t}$ is the chi-square distribution with degrees of freedom $(n-1) = 15$ (sections 2.3 and 2.4). The posterior predictive p-value of obtaining in contemplated replications of the CCPR-S3 a value of $\mathrm{T}_\mathrm{c}(\boldsymbol{x}^\mathrm{rep})$ more extreme than $\mathrm{T}_\mathrm{c}(\boldsymbol{x}) = 22.98$ is $p_\mathrm{P} = 0.08$. Relative to the traditional benchmark 0.05 for p_P the realized discrepancy measure $\mathrm{T}_\mathrm{c}(\boldsymbol{x}) = 22.98$ is not extreme. Thus solely based on the overall discrepancy measure $\mathrm{T}_\mathrm{c}(\boldsymbol{x})$, the statistical consistency model (2) appears to fit the results $\boldsymbol{x}$.

The overall discrepancy measure $\mathrm{T}_\mathrm{c}(\boldsymbol{x})$ does not give a sufficiently detailed picture of the fit of model (2) to the results $\boldsymbol{x}$. A more informative picture is provided by the unilateral and bilateral discrepancy measures. Table 2 displays the realized values of the unilateral discrepancy measures, $\mathrm{T}_i(\boldsymbol{x}) = x_i - x_\mathrm{W}$, for $i = 1, 2, \ldots, 16$. Table 2 also displays the standard deviations $S(T_i(\boldsymbol{x}^\mathrm{rep})) = \sqrt{u^2(x_i) - u^2(x_\mathrm{W})}$ of the posterior predictive distributions of $\mathrm{T}_i(\boldsymbol{x}^\mathrm{rep})$ conditional on the results $\boldsymbol{x}$ and the posterior predictive p-values of $\mathrm{T}_i(\boldsymbol{x})$, for $i = 1, 2, \ldots, 16$. The extreme posterior predictive p-values relative to the traditional benchmarks of 0.05 and 0.95 are shown in bold type. Thus the results from the laboratories labelled 5, 7, and 10 do not agree with the statistical consistency model (2). (We note that the BIPM Report [12] regards the laboratories 5 and 7 as discrepant but not the laboratory 10 with a more extreme p-value than the laboratory 7.)

Table 3 displays the realized values of the bilateral discrepancy measures, $\mathrm{T}_{i-j}(\boldsymbol{x}) = x_i - x_j$, expressed as 10^{-4} units, for $i, j = 1, 2, \ldots, 16$ and $i \neq j$. Table 4 displays the standard deviations $S(\mathrm{T}_{i-j}(\boldsymbol{x}^\mathrm{rep})) = \sqrt{u^2(x_i) + u^2(x_j)}$ of the posterior predictive distributions of the bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x}^\mathrm{rep})$ conditional on the realized results $\boldsymbol{x}$, expressed as 10^{-4} units, for $i, j = 1, 2, \ldots, 16$ and $i \neq j$. Table 5 displays the posterior predictive p-values of the realized bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x})$, for $i, j = 1, 2, \ldots, 16$ and $i \neq j$. The extreme posterior predictive p-values relative to the traditional benchmarks of 0.05 and 0.95 are shown in bold type. We note that the posterior predictive p-values of 62 of the $16^2 - 16 = 240$ bilateral discrepancy measures $x_i - x_j$, for $i \neq j$, are extreme. That is, 26% of the bilateral discrepancy measures have extreme p-values.

If the statistical consistency model (2) reasonably fits the results $x_1, \ldots, x_n$ then they are declared to be consistent. However, a check of the fitness depends on the criterion used for checking. We discussed three criteria: overall discrepancy, unilateral discrepancy and bilateral discrepancy. The unilateral discrepancy measures $\mathrm{T}_i(\boldsymbol{x}) = x_i - x_\mathrm{W}$ and the bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x}) = x_i - x_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$, give a more detailed picture of the fit of the statistical consistency model to the interlaboratory results $x_1, \ldots, x_n$ than the overall discrepancy measure $\mathrm{T}_\mathrm{c}(\boldsymbol{x})$. A set of results may be consistent according to one criterion but not according to a more stringent criterion. For example, from tables 1 and 2 we note that the results from the supplementary comparison CCPR-S3 may be regarded as statistically consistent according to the overall discrepancy measure $\mathrm{T}_\mathrm{c}(\boldsymbol{x}) = \sum_i w_i(x_i - x_\mathrm{W})^2 = (n-1)R_0^2$ but not according to the unilateral discrepancy measures $\mathrm{T}_i(\boldsymbol{x}) = x_i - x_\mathrm{W}$, for $i = 1, 2, \ldots, 16$, because three of the results (from laboratories 5, 7 and 10) have extreme posterior predictive p-values. If the laboratories 5, 7 and 10 are removed then the results from the remaining 13 laboratories are consistent according to the unilateral discrepancy measures. However, from table 5, we note that the results from the following two pairs of laboratories have extreme posterior predictive p-values: (1, 11) and (11, 12). Thus statistical consistency of all unilateral discrepancy measures does not imply statistical consistency of all bilateral discrepancy measures.

5. Use of posterior predictive p-values to assess the degrees of agreement in CIPM key comparisons

CIPM key comparisons are interlaboratory evaluations between national metrology institutes (NMIs) conducted by the consultative committees of the CIPM. They serve as technical bases for international Mutual Recognition Arrangements (MRA) [13]. A CIPM key comparison is expected to yield as output a key comparison reference value (KCRV), unilateral and bilateral degrees of equivalence (DOE), and their associated standard uncertainties. When the results $x_1, \ldots, x_n$ are judged to be overall consistent, the KCRV is generally set as the weighted mean x_W with standard uncertainty $u(x_\mathrm{W})$ [10]. Then the unilateral DOE are defined as $d_i = x_i - x_\mathrm{W}$ with uncertainties $u(d_i) = \sqrt{u^2(x_i) - u^2(x_\mathrm{W})}$, for $i = 1, 2, \ldots, n$, and the bilateral DOE are defined as $d_{i-j} = x_i - x_j$ with uncertainties $u(d_{i-j}) = \sqrt{u^2(x_i) + u^2(x_j)}$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. These uncertainties are determined in [10] from the statistical consistency

Table 3. Realized values of the bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x}) = x_i - x_j$, expressed as 10^{-4} units, for $i, j = 1, 2, \ldots, 16$ and $i \neq j$.

Indices	Indices j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		−1.3	−2.2	0.1	−13.3	−1.9	10.8	−0.2	−0.5	4.9	−6.1	0.9	−1.5	−5.5	−3.1	0.8
2	1.3		−0.9	1.4	−12.0	−0.6	12.1	1.1	0.8	6.2	−4.8	2.2	−0.2	−4.2	−1.8	2.1
3	2.2	0.9		2.3	−11.1	0.3	13.0	2.0	1.7	7.1	−3.9	3.1	0.7	−3.3	−0.9	3.0
4	−0.1	−1.4	−2.3		−13.4	−2.0	10.7	−0.3	−0.6	4.8	−6.2	0.8	−1.6	−5.6	−3.2	0.7
5	13.3	12.0	11.1	13.4		11.4	24.1	13.1	12.8	18.2	7.2	14.2	11.8	7.8	10.2	14.1
6	1.9	0.6	−0.3	2.0	−11.4		12.7	1.7	1.4	6.8	−4.2	2.8	0.4	−3.6	−1.2	2.7
7	−10.8	−12.1	−13.0	−10.7	−24.1	−12.7		−11.0	−11.3	−5.9	−16.9	−9.9	−12.3	−16.3	−13.9	−10.0
8	0.2	−1.1	−2.0	0.3	−13.1	−1.7	11.0		−0.3	5.1	−5.9	1.1	−1.3	−5.3	−2.9	1.0
9	0.5	−0.8	−1.7	0.6	−12.8	−1.4	11.3	0.3		5.4	−5.6	1.4	−1.0	−5.0	−2.6	1.3
10	−4.9	−6.2	−7.1	−4.8	−18.2	−6.8	5.9	−5.1	−5.4		−11.0	−4.0	−6.4	−10.4	−8.0	−4.1
11	6.1	4.8	3.9	6.2	−7.2	4.2	16.9	5.9	5.6	11.0		7.0	4.6	0.6	3.0	6.9
12	−0.9	−2.2	−3.1	−0.8	−14.2	−2.8	9.9	−1.1	−1.4	4.0	−7.0		−2.4	−6.4	−4.0	−0.1
13	1.5	0.2	−0.7	1.6	−11.8	−0.4	12.3	1.3	1.0	6.4	−4.6	2.4		−4.0	−1.6	2.3
14	5.5	4.2	3.3	5.6	−7.8	3.6	16.3	5.3	5.0	10.4	−0.6	6.4	4.0		2.4	6.3
15	3.1	1.8	0.9	3.2	−10.2	1.2	13.9	2.9	2.6	8.0	−3.0	4.0	1.6	−2.4		3.9
16	−0.8	−2.1	−3.0	−0.7	−14.1	−2.7	10.0	−1.0	−1.3	4.1	−6.9	0.1	−2.3	−6.3	−3.9	

Table 4. Standard deviations $S(\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})) = \sqrt{u^2(x_i) + u^2(x_j)}$ of the posterior predictive distributions of the bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on $\boldsymbol{x}$, expressed as 10^{-4} units, for $i, j = 1, 2, \ldots, 16$ and $i \neq j$.

Indices	Indices j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		2.1	1.9	2.8	5.1	3.0	6.9	2.6	1.8	2.7	3.5	2.9	1.7	3.6	3.2	5.3
2	2.1		2.2	3.0	5.2	3.2	7.0	2.8	2.1	2.9	3.6	3.1	2.0	3.8	3.4	5.4
3	1.9	2.2		2.9	5.1	3.0	6.9	2.6	1.9	2.8	3.5	3.0	1.8	3.7	3.2	5.3
4	2.8	3.0	2.9		5.5	3.7	7.2	3.3	2.8	3.5	4.1	3.6	2.7	4.2	3.8	5.7
5	5.1	5.2	5.1	5.5		5.6	8.4	5.4	5.1	5.5	5.9	5.5	5.0	6.0	5.7	7.1
6	3.0	3.2	3.0	3.7	5.6		7.3	3.5	3.0	3.6	4.2	3.7	2.9	4.3	4.0	5.8
7	6.9	7.0	6.9	7.2	8.4	7.3		7.1	6.9	7.2	7.5	7.3	6.9	7.6	7.4	8.5
8	2.6	2.8	2.6	3.3	5.4	3.5	7.1		2.6	3.3	3.9	3.4	2.5	4.0	3.6	5.6
9	1.8	2.1	1.9	2.8	5.1	3.0	6.9	2.6		2.7	3.5	2.9	1.7	3.6	3.2	5.3
10	2.7	2.9	2.8	3.5	5.5	3.6	7.2	3.3	2.7		4.0	3.5	2.6	4.2	3.8	5.6
11	3.5	3.6	3.5	4.1	5.9	4.2	7.5	3.9	3.5	4.0		4.1	3.4	4.7	4.3	6.0
12	2.9	3.1	3.0	3.6	5.5	3.7	7.3	3.4	2.9	3.5	4.1		2.8	4.3	3.9	5.7
13	1.7	2.0	1.8	2.7	5.0	2.9	6.9	2.5	1.7	2.6	3.4	2.8		3.6	3.1	5.2
14	3.6	3.8	3.7	4.2	6.0	4.3	7.6	4.0	3.6	4.2	4.7	4.3	3.6		4.5	6.1
15	3.2	3.4	3.2	3.8	5.7	4.0	7.4	3.6	3.2	3.8	4.3	3.9	3.1	4.5		5.9
16	5.3	5.4	5.3	5.7	7.1	5.8	8.5	5.6	5.3	5.6	6.0	5.7	5.2	6.1	5.9	

model (2), which postulates that the results $x_1, \ldots, x_n$ have independent normal sampling probability distributions with known variances $u^2(x_1), \ldots, u^2(x_n)$, respectively.

It is not clear what the 'degrees of equivalence' are meant to be. The expected values of the sampling distributions of $d_i = x_i - x_{\mathrm{W}}$ and $d_{i-j} = x_i - x_j$ are zero, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. This implies that their realized values are statistical estimates of zero. Therefore, $d_i = x_i - x_{\mathrm{W}}$ do not quantify the agreements of individual results with the weighted mean x_{W} (regarded as the KCRV), and $d_{i-j} = x_i - x_j$ do not quantify the agreements between pairs of individual results, for $i, j = 1, 2, \ldots, n$ and $i \neq j$ [14].

The variances of the differences $d_1, \ldots, d_n$ and d_{i-j}, for $i, j = 1, 2, \ldots, n$ and $i \neq j$, are generally unequal; so, they cannot be directly compared. To compare the differences, they should be transformed into another metric that takes their unequal variances into account. One simple way is to divide $d_1, \ldots, d_n$ and d_{i-j}, for $i, j = 1, 2, \ldots, n$ and $i \neq j$, by their standard deviations (square roots of their variances). The ratios $d_1/u(d_1) \ldots, d_n/u(d_n)$ can be directly compared. Likewise, the ratios $d_{i-j}/u(d_{i-j})$, where $i, j = 1, 2, \ldots, n$ and $i \neq j$, can be directly compared.

The term 'degrees of equivalence' suggests a quantitative scale of measurement for displaying the degrees of agreement. We suggest that $d_i = x_i - x_{\mathrm{W}}$ may be regarded as realized values of the unilateral discrepancy measures $\mathrm{T}_i(\boldsymbol{x}^{\mathrm{rep}})$ with variances $u^2(d_i) = u^2(x_i) - u^2(x_{\mathrm{W}})$, for $i = 1, 2, \ldots, n$. We suggest that $d_{i-j} = x_i - x_j$ may be regarded as realized values of the bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x}^{\mathrm{rep}})$ with variances $u^2(d_{i-j}) = u^2(x_i) + u^2(x_j)$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The posterior predictive p-values p_{P} of $d_i = x_i - x_{\mathrm{W}}$, for $i = 1, 2, \ldots, n$, may be used to assess the degrees of

Table 5. Posterior predictive p-values p_P of the realized bilateral discrepancy measures $\mathrm{T}_{i-j}(\boldsymbol{x}) = x_i - x_j$, for $i, j = 1, 2, \ldots, 16$ and $i \neq j$. The extreme p-values p_P are shown in bold type.

Indices	Indices j															
i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		0.73	0.88	0.49	**1.00**	0.74	0.06	0.53	0.61	**0.04**	**0.96**	0.38	0.81	0.93	0.84	0.44
2	0.27		0.66	0.32	**0.99**	0.57	**0.04**	0.35	0.35	**0.02**	0.91	0.24	0.54	0.87	0.70	0.35
3	0.12	0.34		0.21	**0.99**	0.46	**0.03**	0.22	0.19	**0.01**	0.87	0.15	0.35	0.82	0.61	0.29
4	0.51	0.68	0.79		**0.99**	0.71	0.07	0.54	0.58	0.08	0.94	0.41	0.72	0.91	0.80	0.45
5	**0.00**	**0.01**	**0.01**	**0.01**		**0.02**	**0.00**	**0.01**	**0.01**	**0.00**	0.11	**0.01**	**0.01**	0.10	**0.04**	**0.02**
6	0.26	0.43	0.54	0.29	**0.98**		**0.04**	0.31	0.32	**0.03**	0.84	0.23	0.45	0.80	0.62	0.32
7	0.94	**0.96**	**0.97**	0.93	**1.00**	**0.96**		0.94	0.95	0.79	**0.99**	0.91	**0.96**	**0.98**	**0.97**	0.88
8	0.47	0.65	0.78	0.46	**0.99**	0.69	0.06		0.55	0.06	0.94	0.37	0.70	0.90	0.79	0.43
9	0.39	0.65	0.81	0.42	**0.99**	0.68	0.05	0.45		**0.02**	0.95	0.32	0.72	0.92	0.79	0.40
10	**0.96**	**0.98**	**0.99**	0.92	**1.00**	**0.97**	0.21	0.94	**0.98**		**1.00**	0.87	**0.99**	**0.99**	**0.98**	0.77
11	**0.04**	0.09	0.13	0.06	0.89	0.16	**0.01**	0.06	0.05	**0.00**		**0.04**	0.09	0.45	0.24	0.13
12	0.62	0.76	0.85	0.59	**0.99**	0.77	0.09	0.63	0.68	0.13	**0.96**		0.80	0.93	0.85	0.51
13	0.19	0.46	0.65	0.28	**0.99**	0.55	**0.04**	0.30	0.28	**0.01**	0.91	0.20		0.87	0.70	0.33
14	0.07	0.13	0.18	0.09	0.90	0.20	**0.02**	0.10	0.08	**0.01**	0.55	0.07	0.13		0.30	0.15
15	0.16	0.30	0.39	0.20	**0.96**	0.38	**0.03**	0.21	0.21	**0.02**	0.76	0.15	0.30	0.70		0.25
16	0.56	0.65	0.71	0.55	**0.98**	0.68	0.12	0.57	0.60	0.23	0.87	0.49	0.67	0.85	0.75	

agreement between the individual results and the statistical consistency model (2). The posterior predictive p-values p_P of $d_{i-j} = x_i - x_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$, may be used to assess the degrees of agreement between pairs of results.

6. Summary

A set of the results $\boldsymbol{x} = (x_1, \ldots, x_n)^\mathrm{t}$ from an interlaboratory evaluation with a stated (known) variance–covariance matrix $V(\boldsymbol{x}) = \boldsymbol{D} = [u(x_i, x_j)]$ is said to be statistically consistent if the model $\boldsymbol{x} \sim N(\mathbf{1}\mu, \boldsymbol{D})$ reasonably fits the results $\boldsymbol{x}$. We refer to the model $\boldsymbol{x} \sim N(\mathbf{1}\mu, \boldsymbol{D})$ as the normal statistical consistency model. Statistical consistency implies that the expected values $E(x_1), \ldots, E(x_n)$ of the results are equal (to some unknown constant μ), at least approximately; that is, the results $x_1, \ldots, x_n$ agree with each other. A modern method for checking the fit of a statistical model to the data is Bayesian posterior predictive checking. It is a Bayesian adaptation of the classical (frequentist sampling) method of hypothesis testing. We used posterior predictive checking to check the fit of the normal consistency model, $N(\mathbf{1}\mu, \boldsymbol{D})$, to the interlaboratory results $\boldsymbol{x}$.

The principle of posterior predictive checking is that if a statistical model reasonably fits then the realized results should look plausible under the posterior predictive distribution of potential results that could be obtained in contemplated replications of the interlaboratory evaluation under that model. A posterior predictive distribution is the integral of the sampling distribution of potential data with respect to the Bayesian posterior distributions of the parameters conditioned on the realized data. A systematic discrepancy between the results obtained from the posterior predictive distribution and the realized results indicates misfit. A discrepancy measure is a measure of the discrepancy that one wishes to investigate between the statistical model and the results. A Bayesian posterior predictive p-value, p_P, of a discrepancy measure is the probability of realizing in contemplated replications a value of the discrepancy measure more extreme than its realized (observed) value. The statistical model is suspect if the posterior predictive p-value p_P is close to 0 or close to 1, thereby indicating that the realized results are unlikely to be seen in contemplated replications if the statistical model were true.

We discussed an overall measure of discrepancy for checking the overall fit of the normal consistency model $N(\mathbf{1}\mu, \boldsymbol{D})$ to the interlaboratory results $\boldsymbol{x}$. The posterior predictive distribution of the overall discrepancy measure conditional on the realized results $\boldsymbol{x}$ is a chi-square distribution; therefore, the corresponding posterior predictive p-value is analytically determined. We also discussed two sets of unilateral and bilateral measures of discrepancy. A unilateral discrepancy measure checks whether the result from a particular laboratory agrees with the statistical consistency model. A bilateral discrepancy measure checks whether the results from a particular pair of laboratories agree with each other. The particular laboratories of interest could be any of the participating laboratories. The posterior predictive distributions of the proposed unilateral and bilateral measures of discrepancy are normal with zero expected values and known variances; therefore, the corresponding posterior predictive p-values are analytically determined. A posterior predictive p-value p_P that is extreme (close to 0 or 1) indicates that the normal consistency model does not fit the results $\boldsymbol{x}$.

We have illustrated the calculation of posterior predictive p-values. The numerical example indicates that the unilateral and bilateral discrepancy measures give a better (more detailed) picture of the fit of the normal consistency model to the interlaboratory results than the overall discrepancy measure. The overall, unilateral and bilateral discrepancy measures and their posterior predictive p-values apply to both correlated and independent interlaboratory results. We proposed that the posterior predictive p-values of the realized unilateral and bilateral discrepancy measures may be used to assess the degrees of agreement in the results from a CIPM key comparison.

Acknowledgment

We thank Bert Rust and Howard Yoon for their review and comments.

Appendix A

The Bayesian posterior distribution of μ given $\boldsymbol{x}$ is normal, $N(m, (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1})$, with expected value $m = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and variance $(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}$.

The variance–covariance matrix $\boldsymbol{D} = [u(x_i, x_j)]$ is known; therefore, by substituting (6) in (5) we have

$$f(\boldsymbol{x}|\mu) = (2\pi)^{-n/2}|\boldsymbol{D}|^{-1/2} \times \exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\mathbf{1}m)^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}-\mathbf{1}m) + \frac{(m-\mu)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right]\right\}, \tag{21}$$

where $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$. The expression (21) regarded not as a function of $\boldsymbol{x}$ but as a function of μ is the likelihood function, $l\,(\mu|\boldsymbol{x})$, of μ given $\boldsymbol{x}$; thus,

$$l(\mu|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}\frac{(\mu-m)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right\}. \tag{22}$$

The symbol $\propto$ in (22) stands for 'is proportional to'. As a prior distribution $p(\mu)$ for μ we can use a non-informative improper function that is uniform in μ [6, p 53]; that is, $p(\mu) \propto 1$. According to Bayes's theorem [6], the posterior pdf, $p(\mu|\boldsymbol{x})$, of μ given $\boldsymbol{x}$ is proportional to the product of the likelihood function $l(\mu|\boldsymbol{x})$ and the prior distribution $p(\mu)$. Thus

$$p(\mu|\boldsymbol{x}) \propto l(\mu|\boldsymbol{x}) \times p(\mu) \propto \exp\left\{-\frac{1}{2}\frac{(\mu-m)^2}{(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}}\right\}. \tag{23}$$

By normalizing the expression (23) we get the posterior pdf $p(\mu|\boldsymbol{x})$ given in (7).

Appendix B

The posterior predictive distribution of $\boldsymbol{x}^{\mathrm{rep}}$ conditional on the given results $\boldsymbol{x}$ is normal, $N(\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}, \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}})$, with expected value $\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and variance–covariance matrix $\boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$.

The sampling pdf $f(\boldsymbol{x}^{\mathrm{rep}}|\mu)$ of $\boldsymbol{x}^{\mathrm{rep}}$ given μ is given in (8) and the posterior pdf $p(\mu|\boldsymbol{x})$ of μ given $\boldsymbol{x}$ is given in (7). Thus the integrand $f(\boldsymbol{x}^{\mathrm{rep}}|\mu)\ p(\mu|\boldsymbol{x})$ in the integral (3) may be expressed as

$$(2\pi)^{-(n+1)/2}\begin{vmatrix}\boldsymbol{D} & \mathbf{0}\\ \mathbf{0}^{\mathrm{t}} & (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\end{vmatrix}^{-1/2} \exp\left\{-\frac{1}{2}\begin{pmatrix}\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}\mu\\ \mu-m\end{pmatrix}^{\mathrm{t}} \times \begin{pmatrix}\boldsymbol{D} & \mathbf{0}\\ \mathbf{0}^{\mathrm{t}} & (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{x}^{\mathrm{rep}}-\mathbf{1}\mu\\ \mu-m\end{pmatrix}\right\}. \tag{24}$$

The integrand $f(\boldsymbol{x}^{\mathrm{rep}}|\mu)\ p(\mu|\boldsymbol{x})$ in (3) expressed as (24) may be regarded as the pdf of the $(n+1)$-variate normal distribution for the random vector $((\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\mu)^{\mathrm{t}}, \mu)^{\mathrm{t}}$ with expected value $(\mathbf{0}^{\mathrm{t}}, m)^{\mathrm{t}}$ and variance–covariance matrix

$$\begin{pmatrix}\boldsymbol{D} & \mathbf{0}\\ \mathbf{0}^{\mathrm{t}} & (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\end{pmatrix}, \tag{25}$$

where $m = \boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$. Thus the integral (3) is the marginal pdf of $\boldsymbol{x}^{\mathrm{rep}}$ determined from the joint pdf (24) of the $(n+1)$-variate normal distribution for $((\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\mu)^{\mathrm{t}}, \mu)^{\mathrm{t}}$.

The marginal pdf of $\boldsymbol{x}^{\mathrm{rep}}$ can be easily obtained from the following well-known theorem concerning the distribution of linear functions of a multivariate normal distribution [7].

Theorem. *If* $\boldsymbol{y}$ *has a multivariate normal distribution with expected value* $\boldsymbol{\eta}$ *and variance–covariance matrix* $\boldsymbol{\Sigma}$*, then the distribution of* $\boldsymbol{Ky}$ *is multivariate normal with expected value* $\boldsymbol{K\eta}$ *and variance–covariance matrix* $\boldsymbol{K\Sigma K}^{\mathrm{t}}$.

By applying this theorem with $\boldsymbol{y} = ((\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\mu)^{t}, \mu)^{\mathrm{t}}$, $\boldsymbol{\eta} = (\mathbf{0}^{\mathrm{t}}, m)^{\mathrm{t}}$, $\boldsymbol{\Sigma}$ given in (25), $\boldsymbol{K} = [\boldsymbol{I}, \mathbf{1}]$, we have $\boldsymbol{Ky} = (\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}\mu) + \mathbf{1}\mu = \boldsymbol{x}^{\mathrm{rep}}$, $\boldsymbol{K\eta} = \mathbf{1}m = \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$, and $\boldsymbol{K\,\Sigma K}^{\mathrm{t}} = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$. Thus, the marginal pdf of $\boldsymbol{x}^{\mathrm{rep}}$ is normal with expected value $\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and variance–covariance matrix $\boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$.

Therefore the posterior predictive pdf $p(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x})$ defined by the integral (3) is an n-variate normal distribution with expected value $E(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \mathbf{1}\boldsymbol{B}^{\mathrm{t}}\boldsymbol{x} = \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and variance–covariance matrix $V(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$. That is

$$p(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) \sim N(\mathbf{1}\boldsymbol{B}^{\mathrm{t}}\boldsymbol{x}, \boldsymbol{V}). \tag{26}$$

where $\boldsymbol{B}^{\mathrm{t}} = (\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}$ and $\boldsymbol{V} = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$.

Appendix C

The posterior predictive distribution of $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on the results $\boldsymbol{x}$ is the chi-square distribution, $\chi^2_{(n-1)}$, with degrees of freedom $n-1$.

As discussed in [3, appendix B], $(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}m^{\mathrm{rep}}) = [\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}^{\mathrm{rep}}] = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\boldsymbol{x}^{\mathrm{rep}}$. Consequently, we can express the overall discrepancy measure $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}})$ as $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}}) = (\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}m^{\mathrm{rep}})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}m^{\mathrm{rep}}) = (\boldsymbol{x}^{\mathrm{rep}})^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x}^{\mathrm{rep}}$, where $\boldsymbol{A} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$. The distribution of $\boldsymbol{x}^{\mathrm{rep}}$ conditional on $\boldsymbol{x}$ is given in appendix B. We seek to determine the pdf of $\mathrm{T_c}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on $\boldsymbol{x}$. We will use the following theorem from [8, section 2.5].

Theorem. *If the distribution of* $\mathbf{y}$ *is normal* $N(\boldsymbol{\mu}, \mathbf{V})$ *then the distribution of* $\mathbf{y}^{\mathrm{t}}\mathbf{A}\mathbf{y}$ *is non-central chi-square with degrees of freedom equal to rank of* $\mathbf{A}$ *and non-centrality parameter* (1/2) $\boldsymbol{\mu}^{t}\mathbf{A}\boldsymbol{\mu}$ *if and only if* $\mathbf{AV}$ *is idempotent, that is* $\mathbf{AVAV} = \mathbf{AV}$.

Let us consider this theorem with $\boldsymbol{y} = \boldsymbol{x}^{\mathrm{rep}}$, $\boldsymbol{\mu}= \mathbf{1}$ $(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$, $\boldsymbol{V} = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$, and $\boldsymbol{A} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]$. Now $\boldsymbol{AV} = [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}][\boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}] = [\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}]$, and $\boldsymbol{AVAV} = [\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}]\ [\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}] = [\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}]$. So $\boldsymbol{AV} = \boldsymbol{AVAV}$; that is, $\boldsymbol{AV}$ is idempotent. The rank of an idempotent matrix is equal to its trace [9, p 134]. The trace of $\boldsymbol{AV}$ is $\mathrm{tr}(\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}) = \mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}(\boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}) = \mathrm{tr}(\boldsymbol{I}) - \mathrm{tr}((\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1}) = n - 1$; therefore, the rank of $\boldsymbol{AV}$ is also $n - 1$. Since $\boldsymbol{D}$ is positive definite and hence non-singular, the matrix $\boldsymbol{V} = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$ is also non-singular [9, theorem 18.1.1]. So the rank of $\boldsymbol{AV}$ is the rank of $\boldsymbol{A}$. Thus the rank of $\boldsymbol{A}$ is $n - 1$.

Now $\boldsymbol{\mu}^{t}\boldsymbol{A\mu} = \boldsymbol{x}^{\mathrm{t}}[\boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}]\ [\boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\ [\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\ \boldsymbol{x} = \mathbf{0}$. Thus, $\boldsymbol{AV}$ is an idempotent matrix, $\boldsymbol{\mu}^{t}\boldsymbol{A\mu} = \mathbf{0}$, and rank of $\boldsymbol{A}$ is $n - 1$. Therefore the posterior predictive distribution of the discrepancy measure $\mathrm{T}_{\mathrm{c}}(\boldsymbol{x}^{\mathrm{rep}}) = (\boldsymbol{x}^{\mathrm{rep}})^{\mathrm{t}}\boldsymbol{A}\boldsymbol{x}^{\mathrm{rep}} = (\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}m^{\mathrm{rep}})^{\mathrm{t}}\boldsymbol{D}^{-1}(\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}m^{\mathrm{rep}})$ is the chi-square distribution, $\chi^2_{(n-1)}$, with degrees of freedom $n - 1$.

Appendix D

The posterior predictive distribution $p(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$ of the residuals $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ is the n-variate normal distribution, $N(\mathbf{0}, \boldsymbol{D} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}})$, with expected value $\mathbf{0}$ and variance–covariance matrix $\boldsymbol{D} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$.

The posterior predictive distribution $p(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x})$ of $\boldsymbol{x}^{\mathrm{rep}}$ conditional on the given results $\boldsymbol{x}$ is a fully specified n-variate normal distribution with expected value $E(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}$ and variance–covariance matrix $V(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = \boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$ (appendix B). Since $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}}) = (\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}m^{\mathrm{rep}}) = [\boldsymbol{x}^{\mathrm{rep}} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x}^{\mathrm{rep}}] = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\boldsymbol{x}^{\mathrm{rep}}$, the vector of residuals $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ is a linear function of $\boldsymbol{x}^{\mathrm{rep}}$. Therefore, the posterior predictive distribution $p(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$ of the residuals $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on $\boldsymbol{x}$ is also normal, which is fully described by its expected values and variance–covariance matrix. The expected value of the residuals $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on $\boldsymbol{x}$ is $E(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\ \boldsymbol{D}^{-1}]E(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x}) = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\ \mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\ \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\boldsymbol{x} = \mathbf{0}$. The variance–covariance matrix of the residuals $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ conditional on $\boldsymbol{x}$ is $V(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\ \boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]V(\boldsymbol{x}^{\mathrm{rep}}|\boldsymbol{x})\ [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\ \boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]^{\mathrm{t}} = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\ [\boldsymbol{D} + \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}]\ [\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}] = [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}]\ [\boldsymbol{D}][\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}] + [\boldsymbol{I} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}][\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}][\boldsymbol{I} - \boldsymbol{D}^{-1}\mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\ \mathbf{1}^{\mathrm{t}}] = [\boldsymbol{D} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}] + \mathbf{0}$. Thus the posterior predictive distribution $p(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x})$ of the residuals $\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})$ is the n-variate normal distribution with expected value $E(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = \mathbf{0}$ and variance–covariance matrix $V(\boldsymbol{r}(\boldsymbol{x}^{\mathrm{rep}})|\boldsymbol{x}) = \boldsymbol{D} - \mathbf{1}(\mathbf{1}^{\mathrm{t}}\boldsymbol{D}^{-1}\mathbf{1})^{-1}\mathbf{1}^{\mathrm{t}}$.

References

[1] Birge R T 1932 The calculation of errors by the method of least squares *Phys. Rev.* **40** 207–27

[2] Taylor B N, Parker W H and Langenberg D N 1969 Determination of *e/h*, using macroscopic quantum phase coherence in superconductors: implications for quantum electrodynamics and the fundamental physical constants *Rev. Mod. Phys.* **41** 375–496

[3] Kacker R N, Forbes A B, Kessel R and Sommer K-D 2008 Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations *Metrologia* **45** 257–64

[4] Gelman A, Carlin J B, Stern H S and Rubin D B 2004 *Bayesian Data Analysis* 2nd edn (London: Chapman and Hall)

[5] Rao C R 1973 *Linear Statistical Inference and its Application* 2nd edn (New York: Wiley)

[6] Lee P M 1997 *Bayesian Statistics, An Introduction* 2nd edn (Oxford: Oxford University Press)

[7] Evans M, Hastings N and Peacock B 2000 *Statistical Distributions* 3rd edn (New York: Wiley)

[8] Searle S R 1971 *Linear Models* (New York: Wiley)

[9] Harville D A 1997 *Matrix Algebra from a Statistician's Perspective* (Berlin: Springer)

[10] Cox M G 2002 The evaluation of key comparison data *Metrologia* **39** 589–95 (This paper is the work of an international advisory group on uncertainties commissioned by the director of the BIPM)

[11] Draper N R and Smith H 1981 *Applied Regression Analysis* 2nd edn (New York: Wiley)

[12] Goebel R, Stock M and Köhler R 2000 Report on the international comparison of cryogenic radiometers based on transfer detectors, BIPM-2000/9 (Sèvres: Bureau International des Poids et Mesures) http://www.bipm.org/pdf/RapportBIPM/2000/09.pdf

[13] International Committee of Weights and Measures (CIPM) 1999 *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued By National Metrology Institutes* http://www.bipm.org/utils/en/pdf/mra_2003.pdf

[14] Kacker R N, Datla R U and Parr A C 2004 Statistical analysis of CIPM key comparisons based on the ISO Guide *Metrologia* **41** 340–52