Simple Informative Prior Distributions for Metrology

Anthony O'Hagan
University of Sheffield, UK

Maurice Cox
National Physical Laboratory, UK

November 17, 2021

Abstract

The result of a measurement, including the expression of uncertainty in the measurement, should represent a carefully considered opinion based on the metrologist's experience and expertise, as well as on the data and other information sources. This is the position of the *Guide to the expression of uncertainty in measurement* (GUM), and the requirement for such judgement is clear in the case of Type B evaluation. However, when making Type A evaluations, involving statistical analysis of data, the GUM and its various supplements implicitly consider the data to be the only relevant information. This is unfortunate, and arguably unscientific, when, as is frequently the case, the metrologist could bring other relevant information to bear.

Bayesian statistical methods allow the use of prior information in addition to the data in Type A evaluation, and have been advocated by several authors. However, prior information is in principle subjective; another metrologist may bring different prior information to the analysis, leading to a different measurement result. As in other fields, there is some resistance in the metrology community to embrace Bayesian methods using substantive, subjective prior probability distributions.

We identify four desiderata — justification, simplicity, sufficient benefit and verification — that a Bayesian method employing prior information should satisfy. We present two prior distributions for use in the most basic of all Type A evaluations, where the data comprise a sample of indications assumed to be normally distributed. They represent prior information about the variance of the normal errors, in a simple form that is readily justified in practice. We show the benefit of these prior distributions, both in the single Type A evaluation and in a more complex measurement model, and present simple guidance for verifying their validity.

1 Introduction

1.1 Metrology and the Guide to the expression of uncertainty in measurement

Measurement is an essential part of human activity. Estimates of quantities are required for a diverse range of applications and for each of these estimates a statement is needed about its quality. Such a statement is usually made with measurement uncertainty. The two components, the estimate and the associated uncertainty, together constitute a common way of reporting a measurement result [6].

Measurement uncertainty plays an important role in many areas such as assessing compliance with regulation, and in the calibration of measuring systems. Calibration and testing laboratories are obliged [1] to state the uncertainty of their measurement results, so that recipients can take that uncertainty into account when evaluating their own results.

The Joint Committee for Guides in Metrology (JCGM) is responsible for maintaining and promoting the use of the *Guide to the expression of uncertainty in measurement* (GUM) [3] and the *International vocabulary of basic and general terms in metrology* (VIM) [6]. The GUM (JCGM 100) has for a long time been the authoritative document concerned with the evaluation and expression of measurement uncertainty that attempts to meet this objective:

> This Guide establishes general rules for evaluating and expressing uncertainty in measurement that can be followed at various levels of accuracy and in many fields — from the shop floor to fundamental research. Therefore, the principles of this Guide are intended to be applicable to a broad spectrum of measurements ...

In addition to the GUM, the JCGM has produced two documents on the propagation of probability distributions, JCGM 101 [7] for a single *measurand* (the quantity intended to be measured) and JCGM 102 [4] for multivariate measurands. It has also produced a document JCGM 106 [5] on the use of measurement uncertainty in conformity assessment and, most recently, guidance (JCGM GUM-6) on developing and using models of measurement [8].

The GUM suite of documents uses the concept of *standard uncertainty*, which is specifically defined [3, clause 2.3.1] as the 'uncertainty of the result of a measurement expressed as a standard deviation'.

The GUM treats measurement as in general involving a *measurement model* relating the measurand Y (taken as a univariate quantity here) to input quantities X_i:

$$Y = f(X_1, \dots X_N).$$

Knowledge of Y can be determined given f and knowledge of the X_i. Typically, the GUM itself requires estimates x_i of the X_i, associated standard uncertainties $u(x_i)$ and possibly covariances between the X_i. The process of determining the estimate and standard uncertainty of a model input is itself a measurement, and in the context of such a measurement we will refer to the input quantity as the measurand.

Guidance is given in the GUM on estimating the input quantities and on Type A and Type B evaluation of their standard uncertainties. A Type A evaluation uses statistical methods such as taking the arithmetic mean of a set of readings obtained independently under the same conditions of measurement, and using the standard error of the mean as the standard uncertainty. A Type B evaluation uses a knowledge-based probability distribution for an input quantity, taking the standard deviation of the probability distribution as the according standard uncertainty.

The act of evaluating a quantity X_i (for instance, Type A evaluation) is measurement, and in the context of that measurement that quantity is the measurand. But that evaluation occurs prior to, and outside the context of a measurement model in which X_i is to be used as an input to a measurement of another quantity Y. Now X_i is no longer the subject of the measurement and, in the context of that measurement, Y is the measurand and X_i is just an input.

Degrees of freedom are assigned to inputs, along with estimates and standard uncertainties, as part of Type A and Type B evaluation. The GUM considers a probability distribution characterized by the measurement result in the form of a normal or a shifted and scaled Student's t from which a confidence interval for the measurand is obtained. The degrees of freedom of the t distribution is estimated in the GUM using the Welch-Satterthwaite formula [16].

Confidence intervals are expressed in terms of expanded uncertainty. So, a typical measurement result will involve an estimate, a standard uncertainty and an expanded uncertainty for a 95 % coverage.

Although the GUM still has enormous influence in metrology, with uncertainty in measurements being very widely and routinely evaluated exactly according to the procedure described above, several key components of that procedure have been challenged by various authors. In the following subsections we set out alternative elements that will be adopted in this article.

1.2 Two statistical paradigms

A long-running controversy in metrology concerns the underlying statistical methodology employed for the expression of uncertainty in measurement. The two principal paradigms in statistics, termed 'frequentist' and 'Bayesian', express uncertainty in different ways and using different formal definitions of probability.

- Frequentist methods are based on the frequency definition of probability, where the probability of an event is defined to be the frequency with which that event occurs in the long run, over many repetitions. The Type A procedures given in the GUM are based on frequentist statistical theory, and accordingly the resulting standard uncertainties quantify how variable the estimate of a measurand will be over many repetitions of the measurement process.
- Bayesian methods employ a subjective definition of probability, whereby the probability of an event is a subjective judgement representing a person's rational degree of belief that it will occur. Type B evaluation in the GUM is a subjective judgement and the resulting standard uncertainty quantifies the metrologist's uncertainty about the measurand.

When Type A and Type B evaluations are combined, the GUM is mixing frequentist and Bayesian concepts and has received considerable criticism for doing so (references include [12, 13, 15]). We believe that the only logical, coherent solution is to adopt the Bayesian paradigm consistently, including making Type A evaluations using Bayesian methods. This accords with the suggestion in the GUM [3, clause E.3.5] that disparate standard uncertainties can be combined because ultimately all expressions of uncertainty must be the metrologist's judgement and opinion, and with the use of Bayesian methods in JCGM 101.

1.3 Bayesian methods

From the Bayesian perspective, uncertainty in any quantity is expressed using probabilities, and a complete description of that uncertainty consists of a probability distribution. A Bayesian Type A evaluation for a quantity X will therefore result in a probability distribution for X. It should represent the metrologist's considered judgements about X based on all the available information. In Bayesian analysis, a distinction is made between the data, typically comprising the sample of experimental determinations of X for a Type A evaluation, and the prior information, comprising all other knowledge the metrologist may have, including experience with the measurement procedure and with quantities such as X in the past. The information in the data is represented through the same statistical model as would be used in frequentist Type A evaluation, but the Bayesian analysis also recognises the prior information represented as a prior distribution for X. The two sources of knowledge are synthesised in a very natural way using Bayes' theorem. The result is a probability distribution for X, the *posterior* distribution, that represents the sum of the metrologist's knowledge about X.

An important potential benefit of Bayesian methods in metrology is the ability to make use of more information. The addition of prior knowledge will typically result in less uncertainty regarding X than would have been obtained through use of the data alone. It is worth noting that day to day measurements in practising laboratories often involve Type A evaluations with very few experimental determinations; sample sizes as low as three or four are commonplace. In this context, the addition of prior information can offer substantial benefits.

Where a measurand is expressed as a function of various input quantities through a measurement model, these quantities will all have probability distributions in a Bayesian analysis. An input that is subject to Type A evaluation will have a posterior distribution. One that is subject to Type B evaluation will have a distribution expressed directly as the metrologist's judgement. Distributions for the inputs to a measurement model imply a probability distribution for the measurand, which may for instance be computed using the Monte Carlo method of GUM-S1. Probabilities and probability distributions are always to be understood as representing the considered opinion and judgement of the metrologist.

Bayesian methods are often unfamiliar to scientists whose training in statistics has been confined to the more common frequentist concepts and methods. And although Bayes' theorem is a standard statistical tool, it involves unfamiliar probability operations. Bayesian methods are therefore often regarded as more complex and more mathematical than frequentist methods, but this is an unfair perception. One purpose of this article is to demonstrate that Bayesian methods can be equally straightforward to apply in practice.

1.4 Prior information

The potential benefits of using Bayesian methods in metrology are two-fold. First, adopting the Bayesian framework provides a rigorous and legitimate way to combine Type A and Type B evaluations. Second, the incorporation of the metrologist's prior knowledge in Type A evaluation may strengthen the measurement and allow smaller uncertainties to be reported from a given sample of data.

Bayesian methods will require the specification of a prior distribution that represents the metrologist's judgement about the likely values of a quantity before seeing the data that will be used to obtain the measurement result, based on background knowledge and experience. As such, it is necessarily subjective; different metrologists in the same context might express different prior judgements, although it is the prior knowledge and professional judgement of the metrologist who is responsible for the measurement result that matters. Bayesian methods are widely used in almost all areas where statistical analysis is employed, but they often face resistance because of the subjective nature of the prior distribution. In metrology,

the use of a prior distribution may be viewed, we believe unfairly, as compromising or undermining the objectivity of the data.

To address concerns about subjectivity, some practitioners of Bayesian statistics use so-called *noninformative* prior distributions that are supposed to be objective representations of prior ignorance. In metrology, the standard deviation of the prior distribution expresses the strength of prior information. The larger is the characteristic uncertainty, the weaker is the prior information. A noninformative prior distribution should therefore have a standard deviation that is so large as to be effectively infinite. Using such a prior distribution, it is claimed, should gain the first benefits of Bayesian methods, namely a rigorous framework for combining Type A and Type B evaluations, without contaminating the data with the metrologist's subjective judgements.

Attractive though this approach might be, it is controversial for several reasons.

- Numerous different formulations have been proposed for representing the notion of prior ignorance, and in any given situation they may give different 'noninformative' distributions. There is no unique, objective distribution to represent a state of complete or near ignorance regarding a measurand.
- In practice, there is always some prior knowledge. For example, without some idea of likely values for a measurand it is not possible to devise a suitable measurement procedure, and there is always prior knowledge about the error characteristics of any instrument.
- Claiming prior ignorance when there is in fact some prior information implies failing to use all available information regarding the measurand. That may be seen as unscientific, and even a derogation of duty.

Ultimately, the notion that subjectivity is unacceptable in science is widespread but demonstrably false. Subjective judgement is a feature of all scientific activity: examples include formulating hypotheses, building models, designing experiments, choosing how to analyse data and interpreting data. Good science involves judgements and opinions being formed carefully and rigorously, *scientifically*, and being open to challenge in the forum of scientific debate and peer review.

It is our opinion, therefore, that where genuine prior knowledge exists it should be acknowledged and used, in the form of an informative prior distribution, and not denied by substituting a 'noninformative' distribution. An informative prior distribution will in general allow the reporting of a smaller measurement uncertainty and a correspondingly narrower 95 % coverage interval. However, these benefits depend on the prior distribution being realistic.

A 95 % interval for an unknown quantity should contain the true value of that quantity with probability 0.95. The practical meaning of this statement in metrology is that over a long period of time, when a metrologist constructs many 95 % intervals for many measurands, then approximately 95 % of those intervals should contain the true values of those measurands. This statement will be true when the intervals are constructed as *confidence intervals*, using the frequentist approach to statistics, provided that the statistical model used to construct the intervals is valid. It is straightforward to show (see Appendix A) that it is also true for Bayesian intervals, but only on the additional condition that the prior distributions for those measurands are realistic. That is, the true measurands should behave as if they have been sampled from their prior distributions. If, for instance, the true values of the measurands were mostly to lie in the upper tails of the metrologist's prior distributions, then those prior distributions would not be realistic. The metrologist's prior judgements would consistently underestimate the measurands, and we would not expect 95 % of the resulting intervals to contain the true measurand values.

1.5 Desiderata for informative prior distributions in metrology

We have argued that prior information can and should be used in metrology to enhance the specific data, and indeed that this is one of the important benefits of adopting a Bayesian paradigm. However, we have also seen that (a) metrologists have understandable concerns about the use of subjective prior information, (b) the performance of coverage intervals may be poor if the prior distribution is not valid, and (c) Bayesian methods are generally seen to be complex and mathematically or computationally demanding. These issues will need to be addressed if informative prior distributions are to find practical application in metrology.

We suggest the following list of desirable criteria, which all relate to research reproducibility [22].

- *Justification.* The prior distribution should be based on judgements that are open to scrutiny and justified by reference to prior information and experience.
- *Simplicity.* The Bayesian procedure that derives the distribution of the measurand and summaries such as standard or characteristic uncertainty should be documented in a peer-reviewed source. It should be simple to use and to be replicated.
- *Sufficient benefit.* The benefits of incorporating the prior information, for instance in terms of reduced measurement uncertainty, should be sufficient to warrant the use of the Bayesian procedure.
- *Verification.* The consistency of the prior distribution and the data should be capable of verification.

1.6 Outline of the paper

The organization of this paper is as follows, Section 2 considers the most widely used case of Type A evaluation, in which the data comprise a sample of independent determinations with normally distributed observation errors. We introduce two simple informative prior distributions that represent the kind of prior information that a metrologist will typically have, from previous experience and knowledge of the measurement procedure, regarding the magnitude of the observation errors (*Justification*). We provide simple formulae for deriving the posterior distribution and relevant summaries (*Simplicity*). We show how the addition of this information materially reduces uncertainty in the measurand (*Sufficient benefit*), and we also show how the validity of the prior information can be verified in practice (*Verification*).

In Section 3, we consider the case where a measurement model has multiple inputs, in some or all of which it is possible to apply the new informative priors. The approach is illustrated in a numerical example using a model with six inputs. The reduction in the characteristic uncertainty of the measurand achieved through using the two simple informative prior distributions is examined, and the validity of the prior information is tested.

Section 4 summarises the findings and conclusions of this article.

2 Type A evaluation

The canonical example of Type A uncertainty evaluation is as follows. We have a sample $x = (x_1, \ldots, x_n)$ of n observations, assumed to be distributed independently as $\mathrm{N}(\mu, \sigma^2)$, where μ is the (unknown) population mean, while σ^2 is the (unknown) population variance. We will consider measurement models in which a measurand is expressed in terms of two or more input quantities in Section 3, but here, in the context of its Type A evaluation/measurement we refer to μ as the measurand.

We denote the sample mean by $\bar{x}$ and the sample variance by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,.$$

2.1 Noninformative prior distribution

The standard non-Bayesian Type A evaluation for this problem, given in the GUM, estimates μ by $\bar{x}$ and expresses uncertainty through a standard uncertainty $u(\bar{x})$ and expanded uncertainty $U(\bar{x})$ given by

$$u(\bar{x}) = \frac{s}{\sqrt{n}} \,, \qquad U(\bar{x}) = k(n-1)\frac{s}{\sqrt{n}} \,,$$

where $k(d)$ is the 97.5 % quantile of the Student t distribution with d degrees of freedom.

Supplement 1 to the GUM (GUM-S1) [7] gives a Bayesian Type A evaluation for this problem in which the posterior distribution of μ is a scaled and shifted t distribution with mean $\bar{x}$, standard deviation

$$u(\mu) = \sqrt{\frac{n-1}{n-3}}\frac{s}{\sqrt{n}}$$

and degrees of freedom $n-1$.

Part of the controversy over frequentist versus Bayesian inference in metrology concerns the difference between the two standard uncertainties. The Bayesian $u(\mu)$ is larger than the frequentist $u(\bar{x})$ by the factor $\sqrt{(n-1)/(n-3)}$. However, the interval $\bar{x} \pm k(n-1)s/\sqrt{n}$ is a 95 % coverage interval in both analyses.

The authors have criticized the use of a standard deviation as a measure of uncertainty [10] on the grounds that (a) it has conceptually different interpretations in the frequentist and Bayesian paradigms, (b) it is neither helpful nor meaningful for the recipient of a measurement result, (c) it can be infinite and (d) it is unduly sensitive to the possibility of large measurement errors. They argue that it is more meaningful to tie the primary expression of uncertainty to the width of the 95 % coverage interval. They introduce the *characteristic uncertainty* $c(X)$ of a measurand X, defined to be such that there is a 95 % probability that X will lie within $\pm 2c(X)$ of the estimate. The characteristic uncertainty is well-defined, always finite and relatively insensitive to the chances of extreme measurement errors. Furthermore, the definition of the characteristic uncertainty makes it unnecessary to report both standard uncertainty and expanded uncertainty. For the Type A evaluation based on a normal sample,

$$c(\mu) = \frac{1}{2}k(n-1)\frac{s}{\sqrt{n}}\,.$$

Although not stated explicitly in GUM-S1, the posterior distribution of μ given therein derives from a noninformative prior distribution that is commonly used to represent no prior knowledge about either μ or σ^2. See Appendix B for details of this distribution.

The metrologist may of course have, on the basis of experience or judgement, subjective prior knowledge about the measurand, which can be formally incorporated into the analysis, but it will generally be thought controversial, inappropriate or even inadmissible to influence the estimated value of the measurand in this way. However, the metrologist will also typically have at least some knowledge about the precision of measuring instruments and procedures, based on manufacturers' specifications, data sheets and/or experience. Therefore, relatively uncontroversial prior information about σ^2 will very often exist and, if such information can be incorporated into the Type A evaluation through the use of an informative prior distribution, this should be reflected in reduced posterior uncertainty. Specifically, we should see a reduction in both standard uncertainty and characteristic uncertainty, allowing the metrologist to report a stronger measurement result.

2.2 Simple informative prior distributions

We now suppose that the metrologist can specify prior information about the error variance σ^2, but that the prior distribution for μ is still to be noninformative, reflecting no prior knowledge about the measurand's value. We recognise that Bayesian statistical methods are unfamiliar to most metrologists and can be complex to apply. For any such method to be adopted in regular metrological practice, it must satisfy the desirable criteria listed in Subsection 1.5.

The following two specific informative prior distributions, representing different strengths of prior knowledge, are developed in Appendix B.

- The Mildly Informative Prior (MIP) distribution. The weaker of the two distributions is denoted by MIP(v) and is recommended when the metrologist can be confident that σ^2 will lie within a factor of 9 either side of an estimate v.
- The more Strongly Informative Prior (SIP) distribution. The stronger of the two distributions is denoted by SIP(v) and is recommended when the metrologist can be confident that σ^2 will lie within a factor of 3 either side of an estimate v.

Appendix B shows that these prior distributions are easy to specify and to justify in practice, meeting the *Justification* criterion. Furthermore, they meet the *Simplicity* criterion because they have the property that estimates and uncertainty measures can be derived in closed form as simple formulae.

Full details of these distributions will be found in Appendix B. In each case, the use of the distribution requires only a prior estimate v. The choice of distribution, MIP or SIP, expresses the strength of prior

knowledge, through a judgement of confidence that σ^2 will lie within a factor 9 or 3, respectively, either side of v. Specifically, the metrologist should feel at least 95 % certain that σ^2 will be in that range.

The MIP distribution represents weak prior knowledge about a variance, and we suggest that at least this degree of prior knowledge will be justifiable in the great majority of problems in metrology. In many cases, the more informative WIP distribution should also be justifiable. Instances of the availability of prior information where these distributions could be applicable are in dimensional metrology [11], sludge, biowaste and soil sampling [14], and manufacturing metrology [21].

Other informative priors have been used for Type A uncertainty evaluation, but fail to satisfy our criteria. Cox and Shirono [9] use a Jeffreys' prior truncated above and below. Truncating the Jeffreys' prior [9] either involves an arbitrary choice of truncation points or else explicit judgements about where to place them, knowing that if the resulting prior is to have any informative value the answer will depend on those points. Such choices will be challenging to justify in practice, and the truncated distribution does not yield simple formulae for the estimate or uncertainty measures. Van der Veen [23] considers some weakly informative priors for various forms of Type A uncertainty evaluation, but there are again no simple formulae for the estimate or measures of uncertainty. Instead, calculations need to be carried out using Markov chain Monte Carlo methods.

2.3 The effect of additional information

A proposed new method should of course also offer some tangible advantage over an existing method (*Sufficient Benefit*), which in this case should be an expected reduction in uncertainty measures. The existing method is the Bayesian analysis with noninformative prior distribution, or equivalently the widely used non-Bayesian method of the GUM, under which $c(\mu)$ is as given in Subsection 2.1. We will compare the standard and characteristic uncertainties obtained using the noninformative prior (NIP) distribution with those obtained with an MIP or SIP distribution.

Appendix B.3 shows that for all three prior distributions the posterior distribution of μ is a scaled and shifted Student t distribution with mean $\bar{x}$. Hence all three distributions yield the same estimate of μ,

$$m(\mu) = \bar{x}\,.$$

However, they yield different standard uncertainties

$$u(\mu) = \sqrt{\frac{d^*}{d^* - 2}}\sqrt{\frac{v^*}{n}}$$

and characteristic uncertainties

$$c(\mu) = \frac{k(d^*)}{2}\sqrt{\frac{v^*}{n}},$$

where the values of d^* and v^* are given in Table 1.

Table 1: Posterior degrees of freedom d^* and variance factors v^* for three prior distributions

Prior	d^*	v^*
NIP	$n-1$	s^2
MIP(v)	$n+2$	$[3v + (n-1)s^2]/(n+2)$
SIP(v)	$n+7$	$[8v + (n-1)s^2]/(n+7)$

In effect, the prior information in the MIP and SIP distributions is equivalent to a pseudo-sample of 4 or 9 additional observations, respectively, in each case with a pseudo-sample variance of v. In the d^* column of Table 1, we see that the additional pseudo-sample size increases the sample degrees of freedom $n-1$ by 3 or 8, respectively. In the v^* column, the pseudo-sample variance v is combined with the sample variance s^2 in the natural way.

The effect of these informative prior distributions on both standard and characteristic uncertainty is seen primarily in the posterior degrees of freedom d^*. For the standard uncertainty, increasing d^* reduces the

factor $\sqrt{d^*/(d^*-2)}$, while for the characteristic uncertainty larger d^* decreases $k(d^*)$. In both cases, moving from NIP to MIP to SIP produces systematic reductions in these uncertainty factors.

The effect of the prior information on the v^* term is to pull the sample estimate s^2 of σ^2 towards the prior estimate v. The pull is stronger with the more informative SIP than with MIP. If s^2 is larger than v, then v^* will be smaller than s^2, again reducing the standard and characteristic uncertainties. Conversely, if s^2 is smaller than v the uncertainty will be increased. However, v^* is expected to be neither larger nor smaller than s^2; they are both estimates of σ^2. Overall, therefore, the informative prior distributions will reduce both measures of uncertainty, primarily by increasing d^*.

In practice, the reductions in standard and characteristic uncertainties are almost identical. For the remainder of this article we will focus on characteristic uncertainty, first because we believe it is a more meaningful expression of uncertainty [10], and second because it links directly to the 95 % interval and we also wish to study the effect of informative priors on coverage. Our conclusions will hold equally strongly for those who prefer to report standard uncertainties.

2.4 Numerical example 1

To illustrate the benefits of the informative MIP and SIP prior distributions over the noninformative NIP distribution, as well as to identify the price to be paid for those benefits, we present a simple example in the case of one normal sample.

We therefore consider a single normal sample as set out at the beginning of this section, and contrast the characteristic uncertainties and coverage probabilities from the noninformative NIP prior distribution, with those from the informative MIP and SIP distributions. We set the sample size to $n = 5$ and for MIP and SIP we set $v = 1$.

Figure 1 shows the expected characteristic uncertainty as a function of σ using the NIP (noninformative), MIP and SIP prior distributions. The expected characteristic uncertainty using the NIP distribution, which is equivalent to the standard frequentist analysis of the GUM, is linear in σ, while the values for MIP and SIP, computed by a Monte Carlo method with 10^6 samples as set out in Appendix C, show the influence of the prior information.

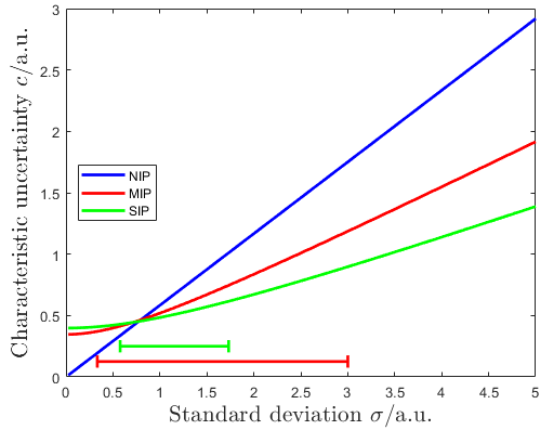

Figure 1: Expected characteristic uncertainty as a function of σ using the NIP (noninformative), MIP (mildly informative) and SIP (strongly informative) prior distributions for a sample of size 5

Considering first the SIP prior distribution, the metrologist's judgement here is that σ^2 is highly likely to lie within a factor 3 either side of the prior estimate $v = 1$, and hence that σ is highly likely to be in the range $3^{-1/2} = 0.577$ to $3^{1/2} = 1.732$. This range is shown in green in Figure 1. Although the prior distribution will yield a characteristic uncertainty that is higher on average than the standard GUM proposal (the NIP line in Figure 1), if σ is smaller than the metrologist expected, conversely it will give a lower characteristic uncertainty if σ is as the metrologist expected or larger. Overall, if σ^2 were indeed drawn randomly according to the metrologist's SIP prior distribution then with probability 0.9 this prior distribution would give a lower characteristic uncertainty than the noninformative prior distribution; the median percentage reduction (the reduction which is achieved or exceeded with probability 0.5) would

be 19.1 %.

Turning to the weaker MIP prior distribution, the metrologist's judgement in this case is that σ will lie in the range 0.333 to 3 with high probability. This range is also marked, in red, in Figure 1. We see a similar pattern to that observed with the SIP distribution. Overall, if σ^2 were indeed drawn randomly according to the metrologist's MIP prior distribution then this prior distribution would give a lower characteristic uncertainty than the noninformative prior distribution with probability 0.8, and the median percentage reduction would be 15.9 %.

These gains relative to the standard GUM formulation are achieved conditional on the metrologist's prior information being valid, in the sense that over many applications the true values of the underlying error variance σ^2 behave as if drawn randomly from the stated prior distribution. To assess the consequences of the prior information not being valid in this sense, we consider the coverage of the implied 95 % coverage interval $m(\mu) \pm 2c(\mu)$. Figure 2 shows the coverage probability as a function of σ for the three priors, and again the most likely ranges for σ according to the MIP and SIP prior distributions are shown.

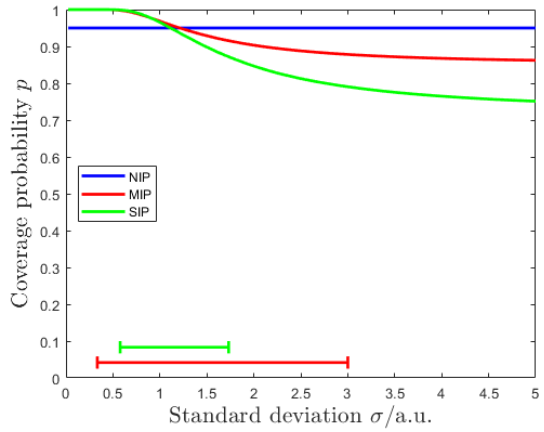

Figure 2: Coverage probability as a function of σ using the NIP (noninformative), MIP (mildly informative) and SIP (strongly informative) prior distributions for a sample of size 5

The coverage probability is identically 0.95 (95 %) by construction for the NIP prior distribution, for all σ^2. For both informative priors, the expected coverage is also 95 % if the prior distribution is valid. However, Figure 2 shows the consequence if the metrologist misjudges the likely values of σ^2. The coverage in both cases is a decreasing function of σ and for small values of σ approaches 100 %. Therefore, if the true error variance is appreciably smaller than the metrologist expects, then the informative prior will result in a conservative evaluation of uncertainty. That is, the metrologist will report a characteristic uncertainty implying a 95 % coverage interval that in fact is almost certain to contain the true value of μ.

Conversely, though, if the metrologist under-estimates the likely value of σ^2 the resulting characteristic uncertainty will also be under-estimated and the implied 95 % coverage interval will have an actual coverage probability appreciably below 95 %. If. for instance, despite specifying the SIP(1) prior distribution, and hence that the metrologist is confident that σ will not exceed 1.732, the true value of σ is 3 or more, then the coverage probability will be less than 80 %. Just as the gains in terms of reduced characteristic uncertainty are greater with the stronger SIP prior distribution, the consequences of mis-specifying the prior information, and in particular of under-estimating σ^2, are greater.

The above illustrations are for $v = 1$, but the benefits of utilising MIP and SIP prior distributions would be the same for any value of v if they are valid judgements, and the consequences of under-estimating σ^2 would be the same for any value of v. In particular, graphs of expected characteristic uncertainty and coverage probability for different values of v would be identical to Figures 1 and 2 except that the x-axis values would be multiplied by $\sqrt{v}$.

The illustrations are also for $n = 5$. The reduction in characteristic uncertainty will be smaller for larger values of n, but will also be even larger for $n < 5$, as shown in Table 2.

We suggest that for sample sizes of 8 or less, the informative prior distributions satisfy our criterion of *Sufficient benefit* by offering substantial reductions in reported uncertainty.

Table 2: Median percentage reduction in characteristic uncertainty using MIP and SIP prior distributions, relative to the noninformative NIP distribution, for different sample sizes n

n	MIP/%	SIP/%
2	72.9	75.5
3	37.9	42.0
4	23.1	26.9
5	15.9	19.1
6	11.9	14.6
7	9.3	11.6
8	7.6	9.5
9	6.4	8.1
10	5.5	7.0

2.5 Reality checking

However, there is an onus on the metrologist to justify the choice of an informative prior distribution. The primary requirement is to be able to document the evidence and experience in support of a prior estimate v for σ^2 such that the metrologist can be confident that σ^2 lies within a factor 3 (for SIP) or 9 (for MIP) either side of v. The sample data cannot be used to estimate or suggest a value for v; the evidence in support of the prior distribution must be *prior* knowledge.

Two suggestions can be made to assist with the choice.

First, it can be noted that it is safe to err on the side of expressing weaker prior information. If, for instance, the metrologist can justify confidence that σ^2 lies within a factor 5 of v, then the MIP prior is acceptable. The expected coverage of the resulting 95 % intervals with the weaker prior will actually be greater than 95 %, while there will still be some gain in terms of shorter intervals and reduced characteristic uncertainty compared with the standard frequentist method of the GUM.

Second, it is possible to test whether the observed sample variance s^2 is consistent with the stated prior distribution for σ^2. Formally, if the prior distribution is valid, the predictive distribution of the ratio s^2/v is $\mathrm{F}_{n-1,8}$ for the SIP prior, and $\mathrm{F}_{n-1,3}$ for the MIP prior, where F_{ν_1,ν_2} denotes the Snedecor F distribution with degrees of freedom ν_1 and ν_2. There would be concern about the validity of the prior distribution if the ratio is very large, since this would suggest that σ^2 is larger than expected by the prior distribution, and this is when the coverage decreases significantly. It is therefore suggested that a laboratory using these simple informative prior distributions should routinely compute s^2/v. Over time, these values should resemble random draws from the corresponding F distribution. A single particularly large value should be cause for investigation and possible choice of an alternative or weaker prior distribution.

The metrologist does not, however, need to be familiar with F distributions because Table 3 suffices to facilitate these checks. For each value of n from 2 to 10, and for each informative prior distribution, four percentiles are given, the 25th, 50th, 75th and 95th. Over time, when using these distributions for many measurements, the metrologist should find that approximately equal numbers of the computed s^2/v values lie below the corresponding 25th percentile, between the 25th and 50th, between the 50th and 75th, and above the 75th. Values above the 95th percentile should be found only occasionally (approximately once in 20 measurements).

If the four proportions are far from equal, or if values of s^2/v exceed the 95th percentile, the following actions may be suggested.

- If many more s^2/v values fall below the corresponding 50th percentiles than above, this suggests that the metrologist may be tending to give values of v that are too large, thereby overestimating the values of σ^2. Smaller characteristic uncertainty values could overall have been reported by better prior estimation of σ^2.
- Conversely, if many more s^2/v values fall above the corresponding 50th percentiles than below, this suggests that the metrologist may be tending to give values of v that are too small, thereby underestimating the values of σ^2. This is a more serious departure from the norm of equal propor-

Table 3: Percentiles of F distributions

	MIP				SIP			
n	25th	50th	75th	95th	25th	50th	75th	95th
2	0.12	0.59	2.02	10.13	0.11	0.50	1.54	5.32
3	0.32	0.88	2.28	9.55	0.30	0.76	1.66	4.46
4	0.42	1.00	2.36	9.28	0.41	0.86	1.67	4.07
5	0.49	1.06	2.39	9.12	0.49	0.91	1.66	3.84
6	0.53	1.10	2.41	9.01	0.53	0.95	1.66	3.69
7	0.56	1.13	2.42	8.94	0.56	0.97	1.65	3.58
8	0.58	1.15	2.43	8,89	0.59	0.99	1.64	3.50
9	0.60	1.16	2.44	8.85	0.61	1.00	1.64	3.44
10	0.61	1.17	2.44	8.81	0.63	1.01	1.63	3.39

tions, because it will mean that the metrologist's reported characteristic uncertainty values may have been overall too small.

- If, when using the MIP prior distribution, many more s^2/v values fall between the corresponding 25th and 75th percentiles than outside this range, this suggests that the metrologist could more often justify using the stronger SIP distribution.
- Conversely, if when using the SIP prior distribution many fewer s^2/v values fall between the corresponding 25th and 75th percentiles than outside that range, this suggests that the metrologist is often using the stronger prior distribution when only the weaker MIP distribution would be justified.
- A single value of s^2/v exceeding the 95th percentile is a cause for concern because it suggests that the informative prior distribution may not be valid, and that the characteristic uncertainty is likely to be underestimated. Such values can be expected to occur by chance, about once in every twenty measurements, even if the prior distribution is valid but should always cause the metrologist to check their justification.

Table 3 and the above check actions could be printed and provided as a standard laboratory reference document.

The ability to validate the prior distribution by checking its consistency with the data is an important practical feature of our proposed simple informative Bayesian methods, and satisfies our *Verification* criterion.

3 Informative prior distributions for model inputs

In Section 2 we considered a single Type A evaluation of a measurand X, but measurement often involves a measurement model to relate the measurand Y to a number of inputs $X_1, X_2, \ldots$, each of which may have Type A or Type B evaluation. The simple informative prior distributions proposed in Section 2.2 have a role to play here, too, since reduced uncertainty about any of the model inputs should lead to reduced uncertainty about the measurand. We will illustrate the extent of this reduction using an example in which there are six Type A evaluations.

3.1 Numerical example 2

The Standards Publication CEN/TR 16988:2016 [2] is entitled 'Estimation of uncertainty in the single burning item test'. Clause 2.5.13.2 deals with the uncertainty concerning an input described as the 'velocity profile correction factor', which we will denote by κ and which is expressed using the sub-model

$$\kappa = \frac{1}{5}\sum_{i=1}^{5}\frac{w_i}{w_c} \tag{1}$$

with six input quantities. w_i, $i = 1, \ldots, 5$, are measurements taken on five different radii and w_c is a central measurement. Each measurement is actually the average of four indications taken at 90° intervals. The six GUM Type A evaluations are reported in Table 4. The characteristic uncertainty of each input is the standard uncertainty multiplied by $k_3/2 = 1.591$. The same results would be obtained from Bayesian Type A evaluations using the noninformative NIP prior distribution in each case.

Table 4: GUM/NIP evaluations, Example 2

Quantity	Estimate /ms^{-1}	Standard uncertainty/ms^{-1}	Degrees of freedom	Characteristic uncertainty/ms^{-1}
w_1	7.00	1.132	3	1.801
w_2	9.39	0.412	3	0.656
w_3	10.62	0.531	3	0.845
w_4	11.25	0.180	3	0.286
w_5	12.37	0.233	3	0.355
w_c	12.39	0.636	3	1.012

Cox and O'Hagan [10] propagate these uncertainties through the model (1) using the Monte Carlo method, showing that the median estimate of κ is 0.817 and its characteristic uncertainty is 0.076. They also employ an approximate computation in which the law of propagation of uncertainty is used to propagate characteristic uncertainties through a linearised version of the model, obtaining the same median estimate and an approximate characteristic uncertainty of 0.075. We now consider the effect of introducing informative prior distributions.

We suppose that the metrologist provides the same prior distribution for the each input's σ^2 parameter, with an estimate of $v = 0.25\,\mathrm{m}^2/\mathrm{s}^2$. That is, in each case the metrologist estimates the standard deviation of the Gaussian sampling error to be $0.5\,\mathrm{m\,s}^{-1}$. We consider the effect of using independent MIP(v) prior distributions for all six inputs, and of using SIP(v) prior distributions instead.

The evaluations of the various inputs now yield scaled and shifted t distributions with parameters d^* and v^* as specified in Table 1. Thus the degrees of freedom are $d^* = 6$ and $d^* = 11$ with MIP and SIP prior distributions, respectively. In Table 5, the second and third columns give the corresponding values of v^*. Using the Monte Carlo method, we sample from these distributions and apply equation (1) to obtain a sample of κ, from which the characteristic uncertainties are found to be 0.052 with MIP prior distributions and 0.045 with SIP distributions.

Table 5: Computations with MIP and SIP prior distributions, Example 2

Input quantity	MIP v^*	SIP v^*	s^2/v
w_1	0.7657	0.5313	5.13
w_2	0.2099	0.2281	0.68
w_3	0.2660	0.2587	1.13
w_4	0.1412	0.1907	0.13
w_5	0.1521	0.1966	0.22
w_c	0.3272	0.2921	1.62

As expected, the additional prior information has reduced the measurement uncertainty regarding κ compared with the characteristic uncertainty of 0.076 obtained with noninformative prior distributions. The reductions achieved by the MIP and SIP distributions are substantial in this example, more than 30 % and 40 % respectively.

We now apply the reality check suggested in Subsection 2.5. The relevant row of Table 3 is for $n = 4$. We only have six values for s^2/v, but none of the checks suggested there indicate a problem with the prior distribution except the last one. If we use SIP prior distributions then one of the values (5.13) exceeds the 95th percentile (4.07). Although one such instance in six measurements is not particularly unexpected, the reality checks suggest that in this case the metrologist should use the MIP distribution.

The prior information in this example is of course not a genuine metrologist's opinion but arbitrarily chosen for the purpose of illustration. In practice, a metrologist having such a sample of just six s^2/v values might still use the SIP prior distribution if they felt it could be fully justified.

4 Conclusions

The metrologist frequently has prior knowledge concerning the likely magnitude of errors in the sample indications for a quantity subject to Type A evaluation. We have presented two simple prior distributions, the MIP and SIP distributions, to encode such prior information. We have shown how they can be rigorously justified in practice through specific prior judgements, presented simple formulae to incorporate them in a Bayesian Type A evaluation and given equally simple procedures to verify their validity over a series of measurements. We have presented examples illustrating the benefits of these prior distributions in terms of reduced measurement uncertainty in Type A evaluation, and showing how even larger reductions in uncertainty for a measurand can be achieved when several Type A evaluations contribute to a measurement model. Although those reductions in uncertainty have been presented in terms of characteristic uncertainty, almost identical reductions are achieved in standard uncertainty.

The mildly informative MIP and the more strongly informative SIP prior distributions have thereby been shown to satisfy the desiderata of *Justification, Simplicity, Sufficient benefit and Verification* for informative Bayesian methods to be acceptable for use in metrology.

Type A evaluation from a sample of indications assumed to be normally distributed is employed daily by metrologists in laboratories world-wide. In all these applications, the MIP and SIP prior distributions offer substantial reductions in measurement uncertainty over the existing GUM procedure, without requiring any more sophisticated computations.

Acknowledgement

This work was supported by an ISCF (Industrial Strategy Challenge Fund) Metrology Fellowship grant provided by the UK Government's Department for Business, Energy and Industrial Strategy (BEIS).

The authors benefited greatly from discussions with Alistair Forbes.

A Expectations of Bayesian coverage intervals

Here we prove the result stated in Section 1.4 concerning the expected coverage of Bayesian intervals.

In this appendix we will denote the measurand by θ and the data to be used in a Type A evaluation of θ by t. Bayesian methods derive the posterior distribution of θ, denoted by $p(\theta \mid t)$ by applying Bayes' theorem to combine the prior distribution $p(\theta)$ with the information in the data, represented by the *likelihood function* $p(t \mid \theta)$. Various forms of Bayesian inference may then be obtained from the posterior distribution. We focus on a 95 % coverage interval, usually referred to in Bayesian analysis as a *posterior credible interval* $\Theta(t)$ defined such that $\mathrm{P}(\theta \in \Theta(t) \mid t) = 0.95$. This is a conditional probability that applies for given data t. The claim in Section 1.4 is that the unconditional probability $\mathrm{P}(\theta \in \Theta(t))$ is also 0.95.

There are two random variables here, θ and t, and we are considering an event, $\theta \in \Theta(t)$, that depends on both. In general, consider an event F depending on two random variables Y and Z. It is a standard result in probability theory (known as the law of total probability, a special case of the law of iterated expectation) that

$$\mathrm{P}(F) = \mathrm{E}(\mathrm{P}(F \mid Z)) \,. \tag{2}$$

The interpretation here is that on the left-hand side the probability of F is unconditional, and therefore averaged over the joint distribution of both Y and Z. On the right-hand side, the term $\mathrm{P}(F \mid Z)$ is the conditional probability of F, averaged over the conditional distribution of Y given Z. This conditional probability is in general a function of Z, and we then take the expectation of this function, averaging with respect to the marginal distribution of Z.

We will apply the general result in two different ways. In both cases, we take F to be the event $\theta \in \Theta(t)$. First let Y to be the measurand θ and Z the data t; then the theorem says

$$\mathrm{P}(\theta \in \Theta(t)) = \mathrm{E}(\mathrm{P}(\theta \in \Theta(t) \mid t)),$$

but the Bayesian interval has the property that $\mathrm{P}(\theta \in \Theta(t) \mid t) = 0.95$, a constant, for all t, and the expectation of a constant is a constant. Therefore the unconditional probability $\mathrm{P}(\theta \in \Theta(t))$ is also 0.95.

However, it is important to recognise that the Bayesian posterior distribution is only a valid opinion for the metrologist regarding the measurand after seeing the data t if the prior distribution is a valid opinion before the data. Hence the statement $\mathrm{P}(\theta \in \Theta(t) \mid t) = 0.95$ and the above proof depends on the validity of the prior distribution.

The role of the prior distribution becomes clearer if we reverse the roles of Y and Z in equation (2) so that now Y is t and Z is θ. The theorem now says that

$$\mathrm{P}(\theta \in \Theta(t)) = \mathrm{E}(\mathrm{P}(\theta \in \Theta(t) \mid \theta)).$$

The probability $\mathrm{P}(\theta \in \Theta(t) \mid \theta)$ is the frequentist coverage probability, in which we consider the measurand θ to be fixed and compute the frequency with which $\theta \in \Theta(t)$ over an infinite sequence of random draws of the data t. For a frequentist 95 % interval, this coverage is 0.95 for all θ, and since this is constant the unconditional probability is also 0.95. In the case of a Bayesian interval, however, $\mathrm{P}(\theta \in \Theta(t) \mid \theta)$ depends on θ. We have proved that the unconditional probability is 0.95, and hence its frequentist coverage will average to 0.95 when averaged with respect to the prior distribution. The practical meaning of this is that over a long sequence of measurements the Bayesian intervals will contain the true measurand values 95 % of the time if, and only if, the corresponding θ values behave as if they are sampled from the metrologist's prior distribution

B Simple informative prior distributions

The MIP and SIP distributions are members of a larger class of distributions known as inverse-chi-square (ICS) distributions; these two specific members of that class are recommended for use by metrologists because of the ease with which they can be justified in everyday metrology. In this appendix, we present some theory of inverse-chi-square distributions and develop the MIP and SIP cases that are presented in Section 2.2 as simple 'default' choices for variances of measurement devices in metrology.

If a parameter z has the ICS distribution with degrees of freedom d and scale v, which we write as $\sigma^2 \sim vd\chi_d^{-2}$. Then the density function of z has the form

$$f(z) = \frac{(vd/2)^{d/2}}{\Gamma(d/2)} z^{-1-d/2} \exp(-vd/(2z)).$$

The expectation of z is $\mathrm{E}(z) = vd/(d-2)$ and the variance is $\mathrm{Var}(z) = 2v^2d^2/\left[(d-2)^2(d-4)\right] = 2\mathrm{E}(z)^2/(d-4)$, provided $d > 2$ and $d > 4$, respectively [20, section 11.5].

ICS distributions provide a flexible family to represent prior information about a variance parameter σ^2. They are defined for positive quantities, and through the choice of degrees of freedom and scale they can represent a wide range of prior knowledge about σ^2. For instance, if the analyst has a prior mean t for σ^2, with variance w, then this can be represented by an ICS prior distribution with degrees of freedom $d = 4 + 2t^2/w$ and scale $v = t(w + t^2)/(2w + t^2)$.

This is not a recommended way to specify a prior distribution, however, because judgements of mean and variance are unreliable. Indeed, research in psychology has identified numerous ways in which people make poor judgements regarding uncertain quantities [19]. Formal protocols for eliciting expert judgements, such as the SHELF protocol, [17, 18] are the gold standard for formulating prior distributions, but require resources and expertise that are generally unavailable to a laboratory making routine measurements.

The noninformative prior distribution for σ^2 that is implicitly used in GUM-S1 to derive the Bayesian analysis presented in Section 2.1 is a limiting case of an ICS distribution in which the degrees of freedom parameter tends to zero (referred to as NIP here).

As alternatives to the noninformative formulation, we propose two informative ICS distributions that can be assigned in practice based only on simple judgements, even when prior information is relatively weak.

The SIP(v) distribution has degrees of freedom 8 and scale v, while the MIP(v) distribution has degrees of freedom 3 and scale v.

B.1 The SIP(v) distribution

Figure 3 shows the SIP(v) distribution.

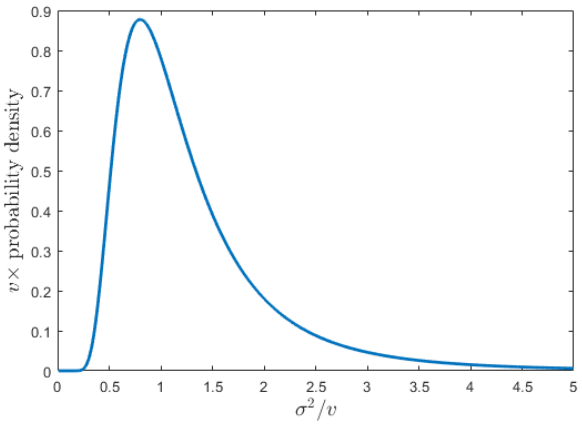

Figure 3: Density function of SIP(v)

As a prior distribution for a variance σ^2, it represents a judgement that σ^2 is most likely to be around the estimate v, and is highly likely to be in the range $v/3$ to $3v$.

Prior information about the magnitude of measurement errors is more naturally expressed in terms of the standard deviation than the variance. The distribution of σ when $\sigma^2 \sim$ SIP(v) is shown in Figure 4.

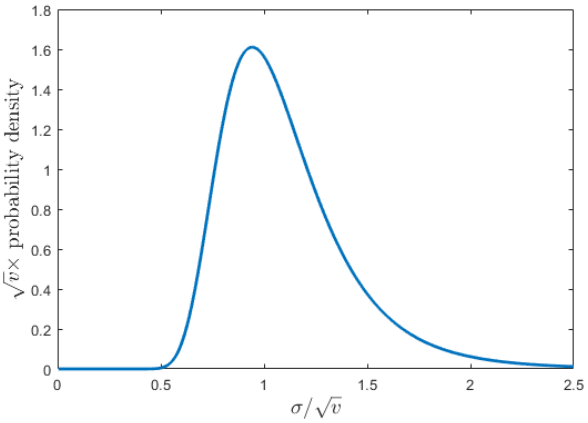

Figure 4: Density function of the square root of SIP(v)

Thus, the prior distribution $\sigma^2 \sim$ SIP(v) represents a belief that σ is most likely to be around its estimate of $\sqrt{v}$ and is highly likely to be within a factor $\sqrt{3} = 1.73$ of that value. It is judged almost certain to be in the range $\sqrt{v}/2$ to $2\sqrt{v}$. For an organisation carrying out regular testing with the same equipment, it is reasonable to suppose that there will be at least this level of knowledge of the standard deviation of measurement errors.

B.2 The MIP(v) distribution

The MIP(v) distribution is an alternative when there is less certainty about σ. Figure 5 shows the distribution of σ when $\sigma^2 \sim$ MIP(v).

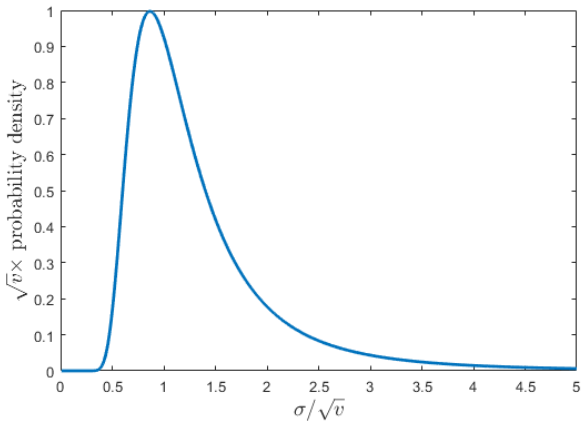

Figure 5: Density function of the square root of MIP(v)

Thus, the prior distribution $\sigma^2 \sim \text{MIP}(v)$ represents a belief that σ is most likely to be around its estimate of $\sqrt{v}$ and is highly likely to be within the range $\sqrt{v}/3$ to $3\sqrt{v}$.

B.3 One normal sample with ICS prior

We show here that ICS distributions are also a convenient choice because they allow a simple application of Bayes' theorem to combine the prior distribution with the information in the data. Formally, they are conjugate distributions for the variance of a normal sample.

Consider the case of a single normal sample, as in Section 2. Suppose that $\sigma^2 \sim v d \chi_d^{-2}$, and suppose that we assign a noninformative uniform prior to μ. Standard Bayesian analysis [20] yields a posterior distribution with the following features.

- $\sigma^2 \sim v^* d^* \chi_{d^*}^{-2}$, where $d^* = d + n - 1$ and $v^* = [vd + (n-1)s^2]/d^*$. Thus, the $n-1$ degrees of freedom in the sample is augmented by the d degrees of freedom in the prior distribution. The posterior scale parameter v^* is a weighted average of the prior scale parameter v and the sample variance s^2, with weights equal to their respective degrees of freedom.
- μ has a scaled and shifted Student t distribution with mean $\bar{x}$ and scale parameter v^*/n. Its variance is $v^* d^*/[n(d^* - 2)]$. As long as d^* is at least 3 (as it is with both the MIP and SIP prior distributions), then the variance exists, even for a sample of size 1.

Therefore, a metrologist using an ICS prior distribution such as MIP(v) or SIP(v) will assign the median estimate $\bar{x}$, just as in the original GUM analysis, with characteristic uncertainty

$$c(\mu) = \frac{k(d^*)}{2}\sqrt{\frac{v^*}{n}}\,. \tag{3}$$

Notice that when $d = 0$ the value of v is irrelevant: it has no effect on the posterior distribution of either μ or σ^2. This is why we omit v when designating the noninformative prior distribution as NIP.

C Computations for normal sample example

Section 2.4 presents results for a single normal sample, using NIP, MIP and SIP prior distributions for σ^2. In general, consider a sample of size n and a sample variance of s^2, and for convenience we write $W = (n-1)s^2/\sigma^2$. The characteristic uncertainty for the noninformative prior distribution (and for the frequentist analysis) is half of the expanded uncertainty:

$$c(\mu) = \frac{1}{2}k(n-1)\frac{s}{\sqrt{n}} = \frac{1}{2}k(n-1)\sqrt{\frac{\sigma^2 W}{n(n-1)}}.$$

For the MIP(v) prior, from (3),

$$c(\mu) = \frac{1}{2}k(2+n)\sqrt{\frac{3v+\sigma^2 W}{n(2+n)}}, \tag{4}$$

and for the SIP(v) prior this becomes

$$c(\mu) = \frac{1}{2}k(7+n)\sqrt{\frac{8v+\sigma^2 W}{n(7+n)}}. \tag{5}$$

Since $W \sim \chi^2_{n-1}$, the expected values of these characteristic uncertainties can all be computed for given σ^2 and n by numerical integration or by a Monte Carlo computation.

The graphs of expected characteristic uncertainty for MIP and SIP priors in Figure 1 were computed for $n = 5$, $v = 1$ and each value of σ by Monte Carlo, sampling 10^6 values of W.

Section 2.4 also reports the median percentage reduction in characteristic uncertainty obtained by the MIP and SIP prior distributions, relative to the NIP distribution, and the probability that the reduction is positive. In each case, the calculations assume that σ^2 is drawn from the metrologist's prior distribution, MIP or SIP respectively. These were computed by Monte Carlo again, sampling 10^6 values of both W and σ^2. For each sampled pair, the characteristic uncertainties were calculated as given above and the percentage reduction using the MIP or SIP prior was computed. The median reduction was then computed as the median of the 10^6 percentage values and the probability of a reduction was computed as the proportion of percentage reductions that were positive.

For each prior distribution, the 95 % credible interval is $\bar{x} \pm 2c(\mu)$. For the noninformative prior distribution, the coverage probability is 95 %, for all σ^2, but for the informative priors the coverage is a function of σ^2. To compute this, we note that the credible interval contains the measurand value μ if

$$(\bar{x} - \mu)^2 \leq 4c^2(\mu),$$

and we now let $V = n\,(\bar{x} - \mu)^2 / \sigma^2 \sim \chi^2_1$. The coverage probability can then be computed by a simple Monte Carlo computation. In the case of the MIP(v) prior, randomly draw a value V from the χ^2_1 distribution and a value W from the χ^2_{n-1} distribution, and then evaluate the condition

$$\frac{\sigma^2 V}{n} \leq \frac{[k(2+n)]^2\,(3v+\sigma^2 W)}{n(2+n)}$$

and therefore

$$fV - W \leq \frac{3v}{\sigma^2},$$

where

$$f = \frac{2+n}{[k(2+n)]^2}\ .$$

The coverage probability is then estimated by the proportion of times that this condition is satisfied in a large number of simulated draws of (V, W). The corresponding condition for the SIP(v) prior is readily derived.

The graphs of coverage probability for MIP and SIP priors in Figure 2 were computed for $n = 5$, $v = 1$ and each value of σ by Monte Carlo, sampling 10^6 values of V and W.

Note that the coverage condition does not depend on μ at all, and only depends on σ^2 and v through their ratio v/σ^2.

References

[1] ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories.

[2] PD CEN/TR 16988:2016, *Estimation of uncertainty in the single burning item test.*

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

[4] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, AND OIML. Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement" — Models with any number of output quantities. Joint Committee for Guides in Metrology, JCGM 102:2011.

[5] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, AND OIML. Evaluation of measurement data — The role of measurement uncertainty in conformity assessment. Joint Committee for Guides in Metrology, JCGM 106:2012.

[6] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, AND OIML. International Vocabulary of Metrology — Basic and General Concepts and Associated Terms. Joint Committee for Guides in Metrology, JCGM 200:2012.

[7] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, AND OIML. Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology, JCGM 101:2008, 2008.

[8] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, AND OIML. Guide to the expression of uncertainty in measurement — Part 6: Developing and using measurement models. Joint Committee for Guides in Metrology, GUM-6:2020, 2020.

[9] COX, M., AND SHIRONO, K. Informative Bayesian Type A uncertainty evaluation, especially applicable to a small number of observations. *Metrologia 54*, 5 (2017), 642–652.

[10] COX, M. G., AND O'HAGAN, A. Meaningful expressions of uncertainty in measurement. Tech. Rep. MS 27, National Physical Laboratory, 2021.

[11] ESTLER, W. T. Measurement as inference: Fundamental ideas. *CIRP Annals 48*, 2 (1999), 611–631.

[12] GLESER, L. J. Assessing uncertainty in measurement. *Stat. Sci. 13* (1998), 277–290.

[13] KACKER, R., AND JONES, A. On use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent. *Metrologia 40* (2003), 235–248.

[14] LAMBKIN, D., NORTCLIFF, S., AND WHITE, T. The importance of precision in sampling sludges, biowastes and treated soils in a regulatory framework. *TrAC Trends in Analytical Chemistry 23*, 10-11 (2004), 704–715.

[15] LIRA, I. Type A Evaluation of Measurement Uncertainty: Frequentist or Bayesian? In *2019 XXIX International Scientific Symposium "Metrology and Metrology Assurance" (MMA)* (2019), IEEE.

[16] NETER, J. *Applied linear statistical models: regression, analysis of variance, and experimental designs*. Irwin, Homewood, IL, 1990.

[17] OAKLEY, J. E., AND O'HAGAN, A. SHELF: the Sheffield Elicitation Framework (version 4). School of Mathematics and Statistics, University of Sheffield, UK. http://tonyohagan.co.uk/shelf, 2019. [Online; accessed 10-March-2021].

[18] O'HAGAN, A. Eliciting and using expert knowledge in metrology. *Metrologia 51*, 4 (2014), S237–S244.

[19] O'HAGAN, A., BUCK, C. E., DANESHKHAH, A., EISER, J. R., GARTHWAITE, P. H., JENKINSON, D. J., OAKLEY, J. E., AND RAKOW, T. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Ltd, 2006.

[20] O'HAGAN, A., AND FORSTER, J. *The Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*. Hodder Education Publishers, 2004.

[21] PAPANANIAS, M., MCLEAY, T. E., MAHFOUF, M., AND KADIRKAMANATHAN, V. A Bayesian framework to estimate part quality and associated uncertainties in multistage manufacturing. *Computers in Industry 105* (feb 2019), 35–47.

[22] Sené, M., Gilmore, I., and Janssen, J.-T. Metrology is key to reproducing results. *Nature 547*, 7664 (Jul 2017), 397–399.

[23] van der Veen, A. M. H. Bayesian methods for Type A evaluation of standard uncertainty. *Metrologia 55*, 5 (2018), 670–684.