Model selection in the average of inconsistent data: an analysis of the measured Planck-constant values

G Mana¹, E Massa¹ and M Predescu²

¹ INRIM—Istituto Nazionale di Ricerca Metrologica, str. delle Cacce 91, 10135 Torino, Italy
² Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy

E-mail: g.mana@inrim.it

Received 25 September 2011, in final form 24 April 2012
Published 29 May 2012
Online at stacks.iop.org/Met/49/492

Abstract
When the data do not conform to the hypothesis of a known sampling variance, the fitting of a constant to a set of measured values is a long debated problem. Given the data, fitting would require one to find what measurand value is the most trustworthy. Bayesian inference is reviewed here, to assign probabilities to the possible measurand values. Different hypotheses about the data variance are tested by Bayesian model comparison. Eventually, model selection is exemplified in deriving an estimate of the Planck constant.

1. Introduction
Given a set of measured values of a constant and the associated uncertainties, the Gauss–Markov theorem states that the weighted arithmetic mean is the unbiased minimum-variance estimator of the measurand [1]. The uncertainty of the mean, which is smaller than the smallest data uncertainty, does not depend on data spread. This is a consequence of the assumption that the variance of the sampling distribution of each datum is exactly known. In practice, this hypothesis is often false and the inconsistency of the data—quantified, for example, by the χ² or the Birge-ratio values—suggests that the uncertainties associated with data are merely lower bounds to the standard deviations of their sampling distributions. In this case, the Gauss–Markov theorem is of no help and the choice of an optimal measurand value is a long debated issue. To average inconsistent data, a number of approaches have been considered; for an overview the reader can refer to [2–9] and the included references. Data inconsistency suggests many different hypotheses: for instance, Gaussian sampling having known variance, a missing term in the error budget or a sampling-variance underestimate. Traditional tests assess the mismatch between model and data: for instance, via χ² calculation. Although a small model likelihood corresponds to a large χ² value, to test a model from data, we must calculate the probability that the model is correct, not the misfit.

Decision theory, probability calculus and Bayesian inference help one to deal with the discrepancy between quoted uncertainties and data scatter. Given a data model, the first step is to find the post-data probability density of the measurand values, which is summarized by the product of data sampling distribution and the pre-data distribution of the measurand values. The normalizing constant of this product, which is named evidence, measures how much the data support the underlying model. Different models can thus be compared by the calculation of evidence. Once the best model has been selected, the optimal choice of the measurand value minimizes the expected loss over the probabilities of the measurand values. Bayesian inference allows different hypotheses to be compared and makes it no longer necessary to exclude the data disagreeing with the majority, as well as to arbitrarily scale the uncertainties to make the data consistent.

After a review of data averaging, the assumption that data are sampled from independent Gaussian distributions having known variance is compared against alternative hypotheses,
for example, that the uncertainties are only lower bounds to the standard deviations or that there are undiscovered errors. As an example, hypothesis testing is carried out using the measured values of the Planck constant. The best data model, of those considered, is identified and a Planck-constant value is estimated.

2. Problem statement

Let us consider a set of \( N \) measured values \( x_i \) of a measurand \( h \), where \( u_i \) is the uncertainty associated with the datum \( x_i \). The data are judged to be mutually inconsistent when combined in a least squares analysis. Hence, they are supposed independently sampled from Gaussian distributions having unknown variance \( \sigma^2 \). The problem is to find optimal estimates of the measurand value and of the confidence intervals.

We cannot rely on the weighted mean, because the actual variances of the sampling distributions are not known. On the other hand, we cannot go back to the arithmetic mean, because it leaves out significant information given by the data uncertainties.

Different hypotheses can be made about the variance of data. The purpose is to choose between the hypotheses and to find an optimal value of the measurand. There is no constraint about hypotheses, nor is there on their number. As an illustration we shall consider: (H0) the data collection is unbiased. Hence, they are supposed independently sampled from Gaussian distributions having unknown variance \( \sigma^2 \). The problem is to find optimal estimates of the standard deviations of the sampling distributions \[9\]. In the same way, we may test the hypothesis, whose lower bounds for the sampling distribution are not known. On the other hand, we cannot go back to the arithmetic mean, because the actual variances of the sampling distributions are not known. On the other hand, we cannot go back to the arithmetic mean, because it leaves out significant information given by the data uncertainties.

To choose from different data models, we must compare the probabilities of each of them being true, given the data set \( \{x_i\} \). The post-data probability of \( H \), \( \text{Prob}(H|x_i) \), where \( H \) indicates the hypothesis being tested, is started from the product rule

\[
\text{Prob}(A, B) = \text{Prob}(A|B)\text{Prob}(B) = \text{Prob}(B|A)\text{Prob}(A),
\]

where \( A \) and \( B \) are propositions—e.g. affirming that a measurement result is \( x_i \), that the measurand value is \( h \) or that the hypothesis \( H \) is true—and \( A|B \) implies that \( B \) is true. Hence,

\[
\text{Prob}(H, x_i) = \text{Prob}(H|x_i) W(x_i) = Z(x_i|H)\text{Prob}(H)
\]

where

\[
W(x_i) = \sum_H \text{Prob}(H, x_i) = \sum_H Z(x_i|H)\text{Prob}(H) \tag{3}
\]

is the probability density of data regardless of any hypothesis and \( \text{Prob}(H) \) is the probability of \( H \) regardless of data. The post-data probability of \( H \),

\[
\text{Prob}(H|x_i) = \frac{Z(x_i|H)\text{Prob}(H)}{W(x_i)}, \tag{4}
\]

is determined by two factors: \( \text{Prob}(H) \), the pre-data probability of \( H \), and the data-dependent term \( Z_H(x_i) = Z(x_i|H) \), which is named evidence for \( H \). The \( Z(x_i|H) \) evidence depends on both the data and \( H \). For a fixed model, \( Z(x_i|H) \) is the data distribution regardless of the model parameters. For fixed data, \( Z(x_i|H) \) is the evidence for the data model. If we are only interested in comparing \( H \) against another hypothesis, \( W(x_i) \), which is a normalizing factor, can be ignored. In addition, on the assumption that, before the data are available, the probabilities of different hypotheses are the same, also \( \text{Prob}(H) \) can be ignored. Therefore, to find the most probable hypothesis, we can identify \( Z(x_i|H) \) with \( \text{Prob}(H|x_i) \), the calculation of which is thus central.

The calculation of evidence is based again on Bayesian analysis. To examine this point, let us consider the general case where \( H = H(h, \lambda) \) is a class of hypotheses having \( h \) and \( \lambda \) as parameters. The post-data probability density of \( H(h, \lambda|\xi) \) of the model parameters is obtained by application of

\[
P(h, \lambda, \xi|H) = \Theta_H(h, \lambda; \xi)Z_H(\xi) = L_H(h, \lambda; \xi)\pi(h, \lambda), \tag{5}
\]

which is the data-dependent term. For fixed \( h \) and \( \lambda \), \( L_H(h; \lambda; \xi) \) is the data sampling distribution; for fixed data, \( L_H(h; \lambda; \xi) \) is the likelihood of the parameter values. Here and in the following, in the function argument, the semicolon separates the independent variables from the fixed parameters.

Model selection within the \( H \) class is equivalent to an estimation problem. From (5), for a fixed \( \lambda \), the evidence for the \( H(\lambda) \) model,

\[
Z_H(\lambda) = \int_{-\infty}^{+\infty} L_H(h; \lambda; \xi)\pi(h, \lambda) \, dh \tag{6}
\]

is the marginalization of \( L_H(h; \lambda; \xi)\pi(h, \lambda) \) with respect to \( h \), that is, the non-normalized posterior distribution of the \( \lambda \) values, regardless of the measurand value. Similarly, the evidence for the whole \( H \) class,

\[
Z_H = \int_{-\infty}^{+\infty} L_H(h; \lambda; \xi)\pi(h, \lambda) \, dh \, d\lambda = \int_{-\infty}^{+\infty} Z_H(\lambda) \, d\lambda \tag{7}
\]

is the normalizing factor of \( L_H(h; \lambda; \xi)\pi(h, \lambda) \). If we only wish to choose an optimal measurand value, we can forget \( Z_H(\lambda) \); however, in model selection problems, it is of the utmost importance.

4. Measurand estimate

Given the data and their associated uncertainty, the Bayesian solution to the estimate problem makes it necessary to assign probabilities to the \( h \) values and to use these probabilities...
to minimize any given loss function [10–12]. The post-data distribution \( \Theta_H(h; x_i) \) is central to this analysis. Given the loss \( \mathcal{L}(h_0 - h) \) due to a wrong estimate \( h_0 \), the optimal choice of the \( h \) value minimizes

\[
(\mathcal{L}(h_0 - h)) = \int_{-\infty}^{+\infty} \mathcal{L}(h_0 - h) \Theta_H(h) \, dh.
\]

(8)

With a quadratic loss, the optimal estimate is the mean; with a linear loss, it is the median. With a loss independent of error, it is the mode. Confidence intervals are easily calculable from \( \Theta_H(h; \lambda) \).

When \( H \) is a class of hypotheses, the result of the Bayesian analysis is the joint probability density \( \Theta_H(h, \lambda; x_i) \), where \( \lambda \) is the class parameter. In this case the usual approach to measurand prediction uses the probability density \( \Theta_H(h, \lambda_0; x_i) \) with a fixed \( \lambda_0 \) value, e.g. the most probable \( \lambda \) value. More accurate predictions are obtained by marginalization,

\[
\Theta_H(h) = \int_{-\infty}^{+\infty} \Theta_H(h, \lambda) \, d\lambda,
\]

(9)

which takes account of the uncertainty of the \( \lambda \) value.

5. Model selection

5.1. \( H_0 \) Hypothesis

If the data uncertainties are the standard deviations, that is, if the data are independently sampled from Gaussian distributions having \( \sigma_i^2 = u_i^2 \) variance, \( L_H(h; x_i) \) is the Gaussian likelihood,

\[
L_{H_0}(h) = \frac{1}{\sqrt{(2\pi)^N \prod_{i=1}^{N} u_i}} \exp\left[ -\frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - h)^2}{2u_i^2} \right]
\]

\[
= \frac{1}{\sqrt{(2\pi)^N \prod_{i=1}^{N} u_i}} \exp\left[ -\frac{\chi^2}{2} \right]
\]

\[
\times \exp\left[ -\frac{(\overline{x} - h)^2}{2\sigma^2} \right],
\]

(10)

where

\[
\overline{x} = \sigma_\overline{x}^2 \sum_{i=1}^{N} \frac{x_i}{u_i^2}
\]

(11a)

is the weighted arithmetic mean of data,

\[
\sigma^2 \equiv \frac{1}{\sum_{i=1}^{N} 1/u_i^2}
\]

(11b)

is the \( \overline{x} \) variance, and

\[
\chi^2 = \sum_{i=1}^{N} \frac{x_i^2}{u_i^2} - \overline{x}^2/\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{u_i^2}
\]

(11c)

is the sum of the squared residuals.

To calculate the post-data probability density of the \( h \) value and the evidence for the \( H_0 \) hypothesis, we must assign pre-data probabilities to the \( h \) values. In the absence of any additional information, we suppose these pre-data probabilities to be independent of the unit-scale origin, which implies the uniform distribution

\[
\pi_H(h) = \frac{\text{If}(\Xi/2 < h < \Xi/2)}{\Xi}
\]

(12)

where \( \text{If}(A) \) gives 1 if \( A \) is true and 0 if it is not. The interval \([-\Xi/2, \Xi/2]\) includes the unknown measurand value. In the following, we shall always use this prior measurand distribution. Accordingly, the evidence for the \( H_0 \) model is

\[
Z_{H_0} = \frac{1}{\Xi} \int_{-\Xi/2}^{\Xi/2} L_{H_0}(h) \, dh = \frac{\sigma_{\overline{x}}}{\Xi \sqrt{2\pi} \sigma_{\overline{x}} \prod_{i=1}^{N} u_i} \exp\left[ -\frac{\chi^2}{2} \right],
\]

(13)

where \( \Xi \) has been chosen large enough that the integration limits can be extended to infinity. The post-data probability density of the \( h \) values is the Gaussian distribution

\[
\Theta_{H_0}(h) = \frac{1}{\sqrt{2\pi} \sigma_{\overline{x}}} \exp\left[ -\frac{(h - \overline{x})^2}{2\sigma_{\overline{x}}^2} \right].
\]

(14)

The mean and most probable \( h \) values are both equal to the weighted mean (11a), whereas \( \sigma_{\overline{x}}^2 \) is the \( h \) variance. In this case, the Bayesian analysis replicates the least-squares analysis. The novelty is the calculation of the evidence for the \( \sigma_i = u_i \) hypothesis: it allows the probability of this assumption to be compared against alternatives.

5.2. \( H_1 \) Hypothesis

The \( H_1 \) hypothesis corresponds to \( \sigma_i = \lambda u_i \), where \( 0 < \lambda < \Lambda \). Since \( H_1 \) is a hypothesis class, we must specify the joint pre-data probability density \( \pi(h, \lambda) = \pi_h(h)\pi_{\lambda}(\lambda) \) of both the \( h \) and \( \lambda \) parameters. This density summarizes the knowledge about \( h \) and \( \lambda \) before considering data. Unlike Dose—who investigated this case extensively in [8]—in the absence of any information about the \( \lambda \) magnitude prior to considering data, we use the bounded uniform density

\[
\pi_{\lambda}(\lambda) = \frac{\text{If}(0 < \lambda < \Lambda)}{\Lambda}.
\]

(15)

Our choice is motivated by the \( H_1 \) data model including the \( u_i \) values; therefore, the uncertainties associated with data are supposed to be available before the data themselves and fix the scale of the problem. For this reason, rather than \( \pi_{\lambda}(\lambda) \) invariance with respect to scale transformations, namely, \( \lambda = a\lambda' \) and \( 0 < \lambda' < \Lambda/a \), we suppose invariance with respect to translations, namely, \( \lambda = \lambda_0 + \lambda' \) and \(-\lambda_0 < \lambda' < \Lambda - \lambda_0 \). The first invariance implies the Jeffreys prior distribution used in [8, 9]; the second implies (15). We will use (12) and (15) throughout the paper.

The parameter likelihood,

\[
L_{H_1}(h, \lambda) = \frac{1}{\sqrt{(2\pi)^N \lambda^N \prod_{i=1}^{N} u_i}} \exp\left[ -\frac{\chi^2}{2\lambda^2} \right]\]

\[
\times \exp\left[ -\frac{(\overline{x} - h)^2}{2\lambda^2\sigma_{\overline{x}}^2} \right],
\]

(16a)

G Mana et al

Metrologia, 49 (2012) 492–500
and the \( \lambda \) evidence,
\[
Z_{H1}(\lambda) = \frac{\sigma_\pi \Gamma(N/2 - 1)}{2 \lambda \Xi \sqrt{2\pi^N N^{-1} \prod_{i=1}^{N} u_i}} \exp\left(-\frac{\chi^2}{2\lambda^2}\right),
\]
(16b)
are obtained from (10) and (13), where \( \Xi \) has been chosen large enough that the integration limits can be extended to infinity.

Equation (16b) expresses the relative probability of different \( \lambda \) values; the most probable one,
\[
R_B = \sqrt{\frac{\chi^2}{N - 1}},
\]
(17)
is the Birge ratio. If \( R_B \) is not equal to 1, the data are inconsistent: the measurement uncertainties have been underestimated and the most probable standard deviations, \( R_B u_i \), differ from the uncertainties associated with data. The data are inconsistent also if \( R_B < 1 \). In this case, the measurement uncertainties have been overestimated.

The Bayesian derivation of (17) allows an expression of the Birge ratio to be found also when the data are correlated. By observing that the \( \chi^2 \) value is invariant with respect to orthogonal transformations of the \( x = [x_1, x_2, \ldots, x_N]^T \) data set, if the data are correlated, the sum of the squared residuals is \( \chi^2 = x^TC_{xx}^{-1}x \), where \( C_{xx} \) is the covariance matrix. Hence, in order to make data consistent, the covariance matrix must be scaled to \( R_B C_{xx} \), where \( R_B \) is still given by (17).

With a fixed \( \lambda \), say, \( \lambda = R_B \), the post-data distribution of the measurand values is
\[
\Theta_{H1}(h; R_B) = N(h; \bar{x}, \sigma_B) = \frac{1}{\sqrt{2\pi} \sigma_B} \exp\left[-\frac{(h - \bar{x})^2}{2\sigma^2_B}\right],
\]
(18)
where \( \sigma_B = R_B \sigma \) and \( N(h; \bar{x}, \sigma_B) \) is a normal distribution having mean \( \bar{x} \) and variance \( \sigma_B^2 \).

According to (7), the evidence for the H1 class is
\[
Z_{H1} = \int_{0}^{\Lambda} Z_{H1}(\lambda) d\lambda = \frac{\sigma_\pi \Gamma(N/2 - 1)}{2 \lambda \Xi \sqrt{2\pi^N N^{-1} \prod_{i=1}^{N} u_i}},
\]
(19)
where \( \Lambda \) is large enough that the limit of integration can be extended to infinity and \( N > 2 \). Consequently, by combining (16a) and (19), the joint post-data distribution of \( h \) and \( \lambda \) is
\[
\Theta_{H1}(h, \lambda) = \frac{\sqrt{\chi^2 N^2 - 2}}{\sqrt{2\pi^N N^{-1} \prod_{i=1}^{N} u_i}} \exp\left(-\frac{\chi^2}{2\lambda^2}\right) \exp\left(-\frac{(h - \bar{x})^2}{2\lambda^2 \sigma^2_B}\right),
\]
(20)
where \( \Xi, \Lambda \to \infty \) and \( N > 2 \). The marginal distribution of the measurand regardless of \( \lambda \) is then
\[
\bar{\Theta}_{H1}(h) = \frac{\sqrt{\chi^2 N^2 - 2}}{\sqrt{\pi} \sigma_\pi \Gamma(N/2 - 1/2)} \exp\left(-\frac{(h - \bar{x})^2}{2\lambda^2 \sigma^2_B}\right),
\]
(21)
where the space of the \( h \) values has been extended to infinity and \( N > 2 \).

The availability of analytic results is of great help because it allows the different ways for estimating the measurand value and the confidence intervals to be compared. The first way relies on distribution (18). In this case, the \( h \) mean is \( \bar{x} \) and the 68% confidence interval is \([\bar{x} - \sigma_B, \bar{x} + \sigma_B]\). This approach assumes that the standard deviations of data are \( R_B u_i \) with certainty, but \( R_B u_i \) are only the most probable values. The second way solves this ambiguity. It relies on (21) and accounts for the lack of knowledge about the standard deviation of data. In this case, though the \( h \) mean is still \( \bar{x} \), the 68% confidence interval must be calculated numerically from the cumulative distribution function of (21).

Another and last consideration concerns the presence in evidence (19) of the ranges \( \Xi \) and \( \Lambda \) of the \( h \) and \( \lambda \) pre-data distributions. They are Occam’s factors penalizing H1 for having the free parameters \( h \) and \( \lambda \). The larger are these ranges, i.e. the model freedom to explain data, the smaller is the evidence for H1 given by data. Since we shall consider the same pre-data distributions, these Occam’s factors are of no relevance and will be omitted when comparing the H1, H2 and H3 models.

5.3. H2 Hypothesis

The H2 class corresponds to \( \sigma^2 = u_i^2 + \lambda^2 u_0^2 \), where \( 0 < \lambda < \Lambda \), \( u_0 \) is a reference \( u \) value, and the pre-data distributions of \( h \) and \( \lambda \) are (12) and (15). The parameter likelihood, model evidence and post-data probability densities are obtained from (10), (13) and (14), where \( u_i^2 + \lambda^2 u_0^2 \) substitutes for \( u_i^2 \). It is not possible to give them a concise form, but this does not prevent numerical computations.

5.4. H3 Hypothesis

We now suppose that the data uncertainties are lower bounds for the standard deviations, that is, \( u_i \leq \sigma_i \leq \max(u_i, \lambda u_0) \), where \( 0 < \lambda < \Lambda \) and \( u_0 \) is a reference \( u \) value. A variant is \( u_i \leq \sigma_i \leq \max(u_i, \lambda u_0) \). In both cases, the pre-data distributions of \( h \) and \( \lambda \) are (12) and (15). Subsequently, we must assign probabilities to the unknown \( \sigma_i \) values, for which only the \( u_i \) lower bounds are known.

In the absence of any information about the \( \sigma_i \) magnitude before considering the data, but with the uncertainties \( u_i \) being already known, we use the translation-invariant uniform distributions
\[
P(\sigma; u, \lambda) \propto \begin{cases} \frac{1}{\lambda u_0} & \text{if the model is } u \leq \sigma \leq \lambda u_0, \\ \frac{1}{u} & \text{if the model is } u < \sigma < \lambda u_0, \end{cases}
\]
(22a)
if the model is \( u < \sigma < \lambda u_0 \), and
\[
P(\sigma; u, \lambda) \propto \begin{cases} \frac{1}{\lambda} & \text{if the model is } u \leq \sigma \leq \max(u, \lambda u); \\ \frac{1}{u} & \text{if the model is } u < \sigma < \lambda u; \end{cases}
\]
(22b)
if the model is \( u < \sigma < \lambda u \); the subscript \( i \) has been dropped.

\( P(\sigma; u, \lambda) \) being stated, the sampling distribution of each datum \( x \) can be marginalized to eliminate the unknown
variance. Hence, if $u \leq \sigma \leq \max(u, \lambda u_0)$, the

$$Q(x; h, u, \lambda) = \int_u^{\lambda u_0} N(x; h, \sigma) P(\sigma; u, \lambda) \, d\sigma$$

$$= \left[ \Gamma(0, (x - h)^2/(2\lambda^2 u_0^2)) - \Gamma[0, (x - h)^2/(2u^2)] \right]$$

$$\frac{2\sqrt{2\pi} \lambda u_0 - u}{\lambda u_0 > u}$$

$$= N(x; h, u) \quad \lambda u_0 \leq u \quad (23a)$$

where $\Gamma(a, z)$ is the incomplete Euler gamma function. If $u \leq \sigma \leq \max(u, \lambda u)$, the marginalized sampling distribution is

$$Q(x; h, u, \lambda) = \int_0^\infty Q(x; h, u, \lambda) \, d\lambda$$

$$\Phi_{\lambda H_3}(h; \lambda) = \frac{\pi_b(h)\pi_s(\lambda)L_{H_3}(h; \lambda)}{Z_{H_3}(\lambda)}, \quad (24a)$$

where

$$L_{H_3}(h; \lambda) = \prod_{i=1}^N Q(x_i; h, u_i, \lambda), \quad (24b)$$

$Q$ is given in (23a) and (23b), $x_i$ and $u_i$ are the $i$th datum and its uncertainty, and the evidence for the $H_3(\lambda)$ model is

$$Z_{H_3}(\lambda) = \pi_s(\lambda) \int_{-\infty}^{\infty} \pi_b(h) L_{H_3}(h; \lambda) \, dh. \quad (24c)$$

It is not possible to obtain an analytical form of (24c), which must be evaluated numerically.

### Table 1. Values of the Planck constant. The main reference is [13]; when relevant, more specific references are listed in the table. The values determined from the ratios $h/m$ and $h/m$ depend on the same $N_A$ value.

| Method | Laboratory | $10^{14} h / s$ | Reference |
|--------|------------|-----------------|-----------|
| Measurement of the Faraday constant | NIST 1980 | 6.626 0657(88) | [15] |
| $N_A$ and $\gamma_1$ measurement | NIST 1989, IAC 2011 | 6.620 7122(73) | [16, 17] |
| $h/m_1$(c) | NIM 1995, IAC 2011 | 6.620 6866(44) | [16, 18] |
| $h/m_1$(c) | KRISS/VNIIM 1998, IAC 2011 | 6.620 7151(12) | [16, 19] |
| $\gamma_1$ measurement | NPL 1979 | 6.620 0729(67) | [20] |
| $h/e$ | NIM 1995 | 6.620 0711(11) | [18] |
| $K_f$ measurement | PTB 1991 | 6.620 0670(42) | [22] |
| Watt-balance experiments | NPL 1990 | 6.620 0682(13) | [23] |
| $h/m_1$(c) | NIST 1998 | 6.620 0689(58) | [24] |
| $h/m_1$(c) | NIST 2007 | 6.620 0691(34) | [25] |
| $h/m_1$(c) | METAS 2011 | 6.620 0691(20) | [26] |
| $h/m_1$(c) | NPL 2012 | 6.620 0712(13) | [27] |
| $h/m_1$(c) | NIM 2012 | 6.620 0706(43) | [28] |
| $N_A$ and quotient of $h$ and the neutron mass | PTB 1998, IAC 2011 | 6.620 0687(52) | [16, 29, 30] |
| $N_A$ and quotient of $h$ and the mass of the electron or an atom | CODATA 2006, IAC 2011 | 6.620 0700(20) | [13, 16] |
| $h/m_1$(c) | Stanford 2002, IAC 2011 | 6.620 0700(22) | [16, 31] |
| $h/m_1$(c) | LKB 2011, IAC 2011 | 6.620 0700(20) | [16, 32] |
| Weighted mean | 6.620 0705(21) | |
| $N_A$ and quotient of $h$ and nuclear mass defects | MIT/ILL 2005, IAC 2011 | 6.620 0764(53) | [16, 33] |
| $h/m_1$(c) | MIT/ILL 2005, IAC 2011 | 6.620 0686(34) | [16, 33] |
| Weighted mean | 6.620 0709(29) | |
The reference value is the weighted mean, $h_{\text{ref}}$, bounds to the standard deviation, according to the most probable data model. The uncertainty is mainly affected by the ratio, where $m$ is listed in Table 1 [13, 14]. The Planck constant links energy to the Planck-constant value on the basis of the measurement results. As an example of application, let us consider the choice of a prototype. The measurement of the Avogadro constant opened the way towards comparing the watt-balance measurements of $h$ with the $h/m$ quotients, where $e$ and $m$ are the electron charge and a mass.

The electric determinations are based on the measurements of $h/e$ or $h/(2e)$ by $\text{Ag} \rightarrow \text{Ag}^+ + e^-$ electrolysis, the high-field spin-flip frequency of protons in samples of $\text{H}_2\text{O}$, and the tunnelling of Cooper’s pairs in Josephson junctions. The watt-balance experiments accede directly to the $h/m_R$ ratio, where $m_R$ is the mass of the international kilogram prototype. The measurement of the Avogadro constant $N_A$ opened the way towards comparing the watt-balance measurements of $h$ with the $h/m$ quotients, where $e$ is the mass of a particle or of an atom. The $h/m(n)$ ratio was determined by time-of-flight measurements of monochromatic neutrons; low-field magnetic resonance, optical spectroscopy and atom interferometry were instrumental in determining $h/m(e), h/m(\text{Cs})$ and $h/m(\text{Rb})$. Nuclear spectroscopy of $^{29}\text{Si}$ and $^{33}\text{S}$, after capture of a thermal neutron, allowed the $N_A h$ product to be determined by comparing the mass defect of $^{29}\text{Si}$ and $^{33}\text{S}$ with the frequencies of the $\gamma$ photons emitted in the cascades from the capture state to the ground state.

Figure 1 shows the input data used in the analysis. The values calculated from the measurements of $h/m(e), h/m(\text{Cs})$ and $h/m(\text{Rb})$ have been averaged, because their uncertainty is mainly affected by the $N_A$ determination and, therefore, they are highly correlated. The values obtained from nuclear spectroscopy have been averaged because they were obtained by repetitions of the same experiment. For the sake of numerical simplicity, in the subsequent analyses, we set $m_0 = 10^{-9} h_0$ and use the reduced data $10^6/(x_i - h_0)/h_0$ and uncertainties $10^6 u_i/h_0$, where $h_0 = 6.626 070 07 \times 10^{-34}$ J s is the weighted arithmetic mean of data.

### 6. Planck-constant estimate

#### 6.1. Input data

As an example of application, let us consider the choice of a Planck-constant value on the basis of the measurement results listed in Table 1 [13, 14]. The Planck constant links energy and momentum to the standard deviation of the wavefunction. Therefore, the $h$ determinations match energy (momentum) and frequency (wavelength) measurements. According to whether an electric or mechanical system is considered, the measurement result is a value of the ratio $h/e$ or $h/m$, where $e$ and $m$ are the electron charge and a mass.

The electric determinations are based on the measurements of $h/e$ or $h/(2e)$ by $\text{Ag} \rightarrow \text{Ag}^+ + e^-$ electrolysis, the high-field spin-flip frequency of protons in samples of $\text{H}_2\text{O}$, and the tunnelling of Cooper’s pairs in Josephson junctions. The watt-balance experiments accede directly to the $h/m_R$ ratio, where $m_R$ is the mass of the international kilogram prototype. The measurement of the Avogadro constant $N_A$ opened the way towards comparing the watt-balance measurements of $h$ with the $h/m$ quotients, where $e$ and $m$ are the mass of a particle or of an atom. The $h/m(n)$ ratio was determined by time-of-flight measurements of monochromatic neutrons; low-field magnetic resonance, optical spectroscopy and atom interferometry were instrumental in determining $h/m(e), h/m(\text{Cs})$ and $h/m(\text{Rb})$. Nuclear spectroscopy of $^{29}\text{Si}$ and $^{33}\text{S}$, after capture of a thermal neutron, allowed the $N_A h$ product to be determined by comparing the mass defect of $^{29}\text{Si}$ and $^{33}\text{S}$ with the frequencies of the $\gamma$ photons emitted in the cascades from the capture state to the ground state.

Figure 1 shows the input data used in the analysis. The values calculated from the measurements of $h/m(e), h/m(\text{Cs})$ and $h/m(\text{Rb})$ have been averaged, because their uncertainty is mainly affected by the $N_A$ determination and, therefore, they are highly correlated. The values obtained from nuclear spectroscopy have been averaged because they were obtained by repetitions of the same experiment. For the sake of numerical simplicity, in the subsequent analyses, we set $m_0 = 10^{-9} h_0$ and use the reduced data $10^6/(x_i - h_0)/h_0$ and uncertainties $10^6 u_i/h_0$, where $h_0 = 6.626 070 07 \times 10^{-34}$ J s is the weighted arithmetic mean of data.

#### 6.2. $H_0$ hypothesis

When supposing the data sampled from independent Gaussian distributions, the standard deviations of which are the data uncertainties, the most probable $h$ value is the weighted arithmetic mean. The evidence for this hypothesis is $Z_{H_0} = 9.6 \times 10^{-16} \Lambda \Omega$. The awful numerical value of $Z_{H_0}$ is due to the fact that it is the normalizing factor of $L_{H_0}(h; x_i) \pi_h(h)$. Its dimensions are $(\text{J s})^{-N}$ and the coefficient $\Omega = 10^N \Lambda_0^{N-1}/(\Lambda_0 h_0^{N-1})$ originates from the normalizations of the data and uncertainties used in the calculation.

#### 6.3. $H_1$ Hypothesis

Although none of the data are totally out of scale, with $\chi^2 = 85.69$ for $v = N - 1 = 16$ degrees of freedom and Birge ratio $R_\beta = \sqrt{\chi^2/v} = 2.31$, they are nevertheless inconsistent. In order to make the data consistent, we supposed that the quoted uncertainties are wrong by a common scale factor $\lambda$. The evidence for this hypothesis is shown in Figure 2; the most probable $\lambda$ value is the Birge ratio. By multiplying the data uncertainties by $R_\beta$, the $\chi^2$ of the weighted arithmetic mean is 16 and the Birge ratio is 1; this removes the discrepancy between data spread and uncertainties.

The evidence for the whole of $H_1$ is $Z_{H_1} = 2.1 \times 10^{-6} \Omega$. Without fixing the extension $\Lambda$ of the parameter space, we cannot compare the $H_0$ and $H_1$ hypotheses, but the $H_1(\lambda = 1)$ model replicates the $H_0$ hypothesis. Hence, the comparison of the $Z_{H_1}(\lambda = 1) = 9.6 \times 10^{-16} \Omega$ evidence against $Z_{H_1}(\lambda = R_\beta) = 1.9 \times 10^{-6} \Omega$ confirms that the odds are severely against the hypothesis of Gaussian samplings having known variance.
6.4. H2 Hypothesis

This hypothesis conjectures a missing uncertainty contribution $10^{-6}\lambda h_0$ common to all the data. The evidence for this hypothesis is shown in figure 2. The most probable $\lambda$ value is $\lambda_0 = 0.17$, with $Z_{H2}(\lambda_0) = 6.1 \times 10^{-3} \Omega$. The evidence for the whole of H2 is $Z_{H2} = 8.0 \times 10^{-4} \Omega$. The comparison against the H0 hypothesis can be made by observing that $H2(\lambda = 0)$ replicates H0 and that $Z_{H2}(\lambda = 0) = 9.6 \times 10^{-16} \Omega$. As before, the odds are against the hypothesis of Gaussian samplings having known variance.

6.5. H3 Hypothesis

When the data uncertainties are supposed to be the lower bounds of unknown standard deviations, we compared two hypotheses. In the H3a case, the standard deviation’s upper bounds are proportional to the uncertainty associated with each datum, i.e. $u_i \leq \sigma_i \leq \max(u_i, \lambda u_i)$. In the H3b case, the same $10^{-6}\lambda h_0$ value bounds the greatest standard deviation of each datum, i.e. $u_i \leq \sigma_i \leq \max(u_i, 10^{-6}\lambda h_0)$. The evidence for these hypotheses is shown in figure 2. Also in these cases, H3a($\lambda = 1$) and H3b($\lambda = \min(10^6 u_i / h_0)$) replicate H0.

The H3 hypothesis could be thought pessimistic: the uncertainties $u_i$ set only lower bounds for the standard deviations of the sampling distributions, which are supposed to be very large if at all possible. However, as shown in figure 2, an infinite standard deviation is not supported by the data and finite upper bounds exist. In the H3a class, the optimal upper bound is $4.3u_i$, with $Z_{H3a}(4.3u_i) = 6.5 \times 10^{-6} \Omega$. In the H3b class, the optimal upper bound is $0.30 \times 10^{-6} h_0$, with $Z_{H3b}(0.30 \times 10^{-6} h_0) = 5.1 \times 10^{-3} \Omega$. This eliminates the worry of a very pessimistic view. The evidence for the whole of these hypothesis classes is given in table 2.

6.6. Prediction of the Planck-constant value

Predictions of the $h$ value were made from both the marginal distributions $\Theta_1(h; \lambda_0)$, where only the most probable $\lambda = \lambda_0$ value was taken into account, and $\Theta_2(h)$, where all the possible $\lambda$ values have been considered. The median $h$ value and 68% confidence intervals are given in table 2 and figure 3, together with the evidences for the different models. The weighted arithmetic mean of the data, which is also the median of the marginal distributions based on the H0 and H1 models, has been taken as the reference value.

To choose the best data model, we must compare the probabilities of each of them being true. These probabilities are proportional to the evidence given in table 2. Of those considered, the hypothesis most favoured by data is H3b. It postulates that the data uncertainties are lower bounds to the standard deviations of the sampling distributions and that a common upper bound exists for all of them. The most probable upper bound, $0.30 \times 10^{-6} h_0$, is relatively small. Only some data have smaller uncertainty; in figure 1 they are indicated by the double error bars. The evidence for hypothesis H2, which postulates that a common contribution is missing in the uncertainty budget, is almost as high as the evidence for H3b. Moreover, H2 and H3b differ only in the greater freedom allowed by H3b to the magnitude of the missed contribution.
An analysis of the measured Planck-constant values

Table 2. Medians $\bar{h}$ of the Planck-constant values calculated according to the marginal distributions $f_{\Theta}(h, \lambda_0)$ and $f_{\bar{h}}(h)$, where $\lambda_0$ is the most probable value of the model parameter. The reference $h$ value is the weighted mean, $h_0 = 6.626\,070\,07 \times 10^{-34} \text{J_s}$, of the data. The lower and upper bounds are the 16% and 84% quantiles. $Z_H$ is the evidence for the model, which is proportional (through the same proportionality factor) to the probability $\text{Prob}(H|x)$ of being true.

| Model | $\lambda_0$ | $f_{\Theta}(h, \lambda_0)$ | $f_{\bar{h}}(h)$ | $\sum H f_{\bar{h}}(h) Z_H / 10^{(N-1)}$ | $\text{Prob}(H|x)$ |
|-------|-------------|-----------------|----------------|---------------------------------------------|------------------|
| H0    | --          | --              | 0.10           | 9.6 $\times 10^{-10}$                        | --               |
| H1    | 2.3         | $0^{+42}_{-1}$  | 0.10           | 2.1 $\times 10^{-6}$                         | 0.95 $\times 10^{-4}$ |
| H2    | 0.17        | $-10^{+62}_{-2}$| -11 $^{+39}_{-59}$ | 0.8 $\times 10^{-3}$                         | 0.36             |
| H3a   | 3.4         | $-11^{+45}_{-4}$| $-12^{+48}_{-5}$ | 1.5 $\times 10^{-5}$                         | 6.82 $\times 10^{-3}$ |
| H3b   | 0.30        | $-29^{+61}_{-66}$| $-30^{+65}_{-65}$ | 1.4 $\times 10^{-3}$                         | 0.63             |
| all   | --          | --              | --             | 2.2 $\times 10^{-3}$                         | 1.00             |

H0: $\sigma = u$, H1: $\sigma = \lambda u$, H2: $\sigma^2 = u^2 + (10^{-6}\lambda h_0)^2$, H3a: $u < \sigma < \max(u, \lambda u)$, H3b: $u < \sigma < \max(u, 10^{-6} \lambda h_0)$.

Figure 3. Predictions of the Planck-constant values derived from the marginal distributions $f_{\Theta}(h, \lambda_0)$ (lower line) and $f_{\bar{h}}(h)$ (upper line). The last value (all) takes into account the ignorance of the data model by weighting the H1, H2, H3a and H3b values according to the relevant evidence. The reference $h$ value is the weighted mean, $h_0 = 6.626\,070\,07 \times 10^{-34} \text{J_s}$, of the data. The asymmetric error bars indicate 68% confidence intervals.

Birge’s hypothesis of a missing scaling factor common to all uncertainties, both in the original and extended versions H1 and H3a, though more supported than the hypothesis of Gaussian samplings having standard deviations equal to the data uncertainties, is not sustained by the data.

At the end of this analysis we have different estimates related to the different hypotheses about the uncertainties associated with data. The variability of the data model can be eliminated by marginalization. Hence,

$$\bar{h}(h) = \sum H f_{\bar{h}}(h) \text{Prob}(H|x) = \frac{\sum H f_{\bar{h}}(h) Z_H}{\sum H Z_H},$$

(25)

where (2) has been used and the same 25% probability—regardless of data—has been supposed for each hypothesis. The numerical values of the relevant evidence and the final $h$ estimate are given in table 2 and figure 3. The posterior density $f_{\Theta}(h)$, which accounts for all the data models considered here, is shown in figure 4.

Figure 4. Post-data distribution of the Planck-constant values given the measurement results listed in table 1. Marginalization has been carried out with respect to all models H1, H2, H3a and H3b.

7. Conclusions

The probability calculus offers a rigorous solution to the problem of fitting a constant to a set of measured values, no matter whether the data are consistent or not. When we suspect the presence of outliers [9] or data non-conformity [4], the post-data distribution of the measurand values depends on the model supposed for the unknown variance of the data.

Different data models can be considered and that most supported by data can be identified by calculating and comparing the probabilities of each model being true. Once the most probable model is found, a measurand value can be optimally chosen according to the measurand post-data probabilities. The ignorance about the data model can also be taken into account by weighting the measurand estimates according to the model probabilities. This corresponds to marginalization of the post-data distribution of the measurand values over the data models. Our analysis clarifies that a hypothesis cannot be accepted or rejected absolutely, but only relative to alternatives against which its probability is compared. From this viewpoint, hypothesis test and model
selection are nothing else than a Bayesian inference procedure and they do not differ from parameter estimation.

Our analysis proved also that, on the assumption that the uncertainties associated with data are in error by an unknown scale factor, the Birge ratio can be interpreted as the most probable value of such a scale factor. In addition, the analysis suggested how the Birge ratio can be generalized to the case when the data are correlated.

Acknowledgments

This work was jointly funded by the European Metrology Research Programme (EMRP) participating countries within the European Association of National Metrology Institutes (EURAMET) and the European Union.

References

[1] Luemberger D G 1969 *Optimization by Vector Space Methods* (New York: Wiley)
[2] Birge R T 1929 Probable values of the general physical constants Rev. Mod. Phys. 1 73
[3] Paule R C and Mandel J 1982 Consensus values and weighting factors J. Res. Natl Bur. Stand. 87 377–85
[4] Weise K and Wöger 2000 Removing model and data non-conformity in measurement evaluation Meas. Sci. Technol. 11 1649–58
[5] Cox M G 2002 The evaluation of key comparison data Metrologia 39 589–95
[6] Kacker R, Datla R, and Parr A 2002 Combined result and associated uncertainty from interlaboratory evaluations based on the ISO Guide Metrologia 39 279–935
[7] Willink R 2002 Statistical determination of a comparison reference value using hidden errors Metrologia 39 343–54
[8] Dose V 2007 Bayesian estimate of the Newtonian constant of gravitation Meas. Sci. Technol. 18 176–82
[9] Sivia D and Skilling J 2006 *Data Analysis: A Bayesian Tutorial* (Oxford: Oxford University Press)
[10] Jaynes E T 2003 *Probability Theory: The Logic of Science* (Cambridge: Cambridge University Press)
[11] Mc Kay D J C 2003 *Information Theory, Inference, and Learning Algorithms* (Cambridge: Cambridge University Press)
[12] Gregory P C 2005 *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge: Cambridge University Press)
[13] Mohr P J, Taylor B N and Newell D B 2008 CODATA recommended values of the fundamental physical constants: 2006 *Rev. Mod. Phys.* 80 633–730
[14] Mohr P J, Taylor B N and Newell D B 2012 CODATA recommended values of the fundamental physical constants: 2010 *Rev. Mod. Phys.* submitted; preprint http://arxiv.org/abs/1203.5425; web http://physics.nist.gov/cuu/Constants
[15] Bower V E and Davis R S 1980 The electrochemical equivalent of pure silver—a value of the Faraday *J. Res. Natl Bur. Stand.* 85 175–93
[16] Andreas B et al 2011 Counting the atoms in a 28Si crystal for a new kilogram definition *Metrologia* 48 S1–13
[17] Williams E R, Jones G R, Ye S, Liu R, Sasaki H, Olsen P T, Phillips W D and Layer H P 1989 A low field determination of the proton gyromagnetic ratio in water *IEEE Trans. Instrum. Meas.* 38 233–7
[18] Liu B, Liu H, Jin T, Lu Z, Du X, Xue S, Kong J, Yu B, Zhou X, Liu T and Zhang W 1995 A recent determination for the SI values of γp /c and 2e/h at NIM Acta Metrol. Sin. 16 161–8
[19] Park P G, Choi C H, Kim C S, Shifrin Y V and Khorev V N 1999 The proton gyromagnetic ratio by low field method measured at KRISS J. Korean Phys. Soc. 34 327–32
[20] Kibble B P and Hunt G J 1979 A measurement of the gyromagnetic ratio of the proton in a strong magnetic field *Metrologia* 15 5–30
[21] Clother W K, Sloggett G J, Bairnsfather H, Currey M F and Benjamin D J 1989 Metrologia 26 9–46
[22] Funck T and Siemenknect V 1991 Determination of the volt with the improved PTB watt balance *IEEE Trans. Instrum. Meas.* 40 158–61
[23] Kibble B P, Robinson I A and Belliss J H 1990 A realization of the SI watt by the NPL moving-coil balance *Metrologia* 27 173–92
[24] Williams E R, Steiner R L, Newell D B and Olsen P T 1998 Accurate measurement of the Planck constant *Phys. Rev. Lett.* 81 2404–7
[25] Steiner R L, Williams E R, Liu R and Newell D B 2007 Uncertainty improvements of the NIST electronic kilogram *IEEE Trans. Instrum. Meas.* 56 592–96
[26] Eichenberger A, Baumann H, Jeanneret B, Jeckelmann B, Richard P and Beer W 2011 Determination of the Planck constant with the METAS watt balance *Metrologia* 48 133–41
[27] Robinson I A 2012 Toward the redefinition of the kilogram: A measurement of the Planck constant using the NPL Mark II watt balance *Metrologia* 49 113–56
[28] Steele A G, Mejia J, Sanchez C A, Yang L, Wood B M, Sturgeon R E, Mester Z and Inglis A D 2012 Reconciling Planck constant determinations via watt balance and enriched-silicon measurements at NRC Canada *Metrologia* 49 L8–10
[29] Krueger E, Nistler W and Weirauch W 1998 Determination of the fine-structure constant by a precise measurement of h/moe: the final result *Metrologia* 35 203–9
[30] Krueger W, Nistler W and Weirauch W 1999 Re-evaluation of a precise measurement of h/moe *Metrologia* 36 147–8
[31] Wicht A, Hensley J M, Sarajlic E and Chu S 2002 A preliminary measurement of the fine structure constant based on atom interferometry *Phys. Scr.* A T102 82–8
[32] Bouchendira R, Cladé P, Guellati-Khelifa S, Nez F and Biraben F 2011 New determination of the fine structure constant and test of the quantum electrodynamics *Phys. Rev. Lett.* 106 080801
[33] Rainville S et al 2005 A direct test of E = mc² *Nature* 438 1096–7