A minimalistic and general weighted averaging method for inconsistent data

Martino Trassinelli, Marleen Maxton

A minimalistic and general weighted averaging method for inconsistent data

M. Trassinelli [1,*] and M. Maxton [1]

[1] *Institut des NanoSciences de Paris, CNRS, Sorbonne Université, F-75005 Paris, France*

(Dated: December 21, 2024)

The weighted average of inconsistent data is a common and tedious problem that many scientists have encountered. The standard weighted average is not recommended for these cases, and different alternative methods are proposed in the literature. Here, we discuss a method first proposed by Sivia in 1996 that is based on Bayesian statistics and keeps the number of assumptions to a minimum. We propose this approach as a new standard for calculating weighted averages. The uncertainty associated with each input value is considered to be just a lower bound of the true unknown uncertainty. The resulting likelihood function is no longer Gaussian, but has smoothly decreasing wings, which allows for a better treatment of scattered data and outliers. The proposed method is tested on a series of data sets: simulations, CODATA recommended value of the Newtonian gravitational constant, and some particle properties from the Particle Data Group, including the proton charge radius. A freely available Python library is also provided for a simple implementation of the proposed averaging method.

I. INTRODUCTION

The standard method for combining different independent evaluations x_i of the same quantity is to use the weighted average

$$\hat{\mu} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}, \tag{1}$$

that employs the inverse of the square of the associated uncertainties σ_i as weights. The corresponding uncertainty is given by

$$\sigma_{\hat{\mu}} = \sqrt{\frac{1}{\sum_i 1/\sigma_i^2}}. \tag{2}$$

The big advantage of such a procedure is the analytical and simple formula that anyone can easily apply to any data set. In addition, it is statistically well justified with a very small number of simple assumptions. More importantly, the method is sufficiently universal to be considered as a standard procedure in the scientific community and can be found in any basic data analysis lecture.

However, the inverse-invariance method, referred to in the following pages simply as *standard*, has a drawback. As we can see from Eq. (2), the final uncertainty depends only on the data uncertainties σ_i, but not on the data spread, which could be larger than the values of σ_i (see, e.g., [1] for a more detailed discussion). This is, however, a common scenario in science, possibly caused by an uncontrolled systematic effect in the measurement procedure or by different biases in measurements conducted in different laboratories and/or with different methods. Common questions that arise are how to take into account such information on the data dispersion in the calculation of a weighted average and how to treat outliers.

To answer such questions, several approaches have been proposed in the literature. In the case of inconsistent data sets, a very common and basic method is to use the standard weighted average while artificially increasing its associated uncertainty. But how should one choose objectively the uncertainty expansion factor? The most common method has been proposed by Birge [2] almost one hundred years ago. It is based on the χ^2 value obtained by the difference between the standard weighted average and the single input values. The uncertainty expansion factor R_{Birge}, the *Birge ratio*, is applied to the single uncertainties $\tilde{\sigma}_i = R_{\text{Birge}}\sigma_i$, with

$$R_{\text{Birge}} = \sqrt{\frac{1}{n-1}\sum_i \frac{(x_i-\hat{\mu})^2}{\sigma_i^2}} = \sqrt{\frac{\chi^2}{n-1}}, \tag{3}$$

where n is the number of x_i data points. In this way, the final value of the reduced χ^2 is adjusted to be close to unity, as expected for consistent data sets. The use of the Birge ratio is indeed one of the pillars of the statistical treatment employed by the Task Group on Fundamental Constants of the Committee on Data of the International Science Council (here simply abbreviated by CODATA) and the Particle Data Group (PDG). A modified version of the Birge ratio has been proposed in past works based on Bayesian statistics. The scaling factor R between the estimated uncertainty σ_i and the real uncertainty σ_i' is considered unknown but common to all data points [1, 3–5]. Assuming a non-informative Jeffreys' prior probability $p(R) \propto 1/R$ and marginalising over the possible values of R, a final probability distribution is obtained with an average value (the mode of the final probability distribution) equal to the standard weighted average. The uncertainty corresponds to an expansion of the standard weighted average uncertainty by a factor equal to $R_{\text{Bayes}} = \sqrt{(n-1)/(n-3)}R_{\text{Birge}}$. Variations of this approach are discussed in Refs. [6, 7].

In principle, because of the common scaling factor for each measurement result, the Birge ratio and its modified versions discussed above are not well suited for inter-

* martino.trassinelli@insp.jussieu.fr

laboratory averages, where very different systematic effects can occur. To compensate partially for such an issue, past works [3, 7] proposed to assign a random bias β_i to each measurement, with a common mean value and standard deviation $\sigma_{\rm bias}$ for the entire ensemble of measurements. Here, a double marginalisation over the β_i values and their shared uncertainty $\sigma_{\rm bias}$ is required, for which a prior probability has to be chosen, typically a non-informative Jeffreys' prior. An evolution of such an approach has been proposed [8]. It consists of organising the input values into clusters, each with a different $\sigma_{\rm bias}$ value, and subsequently performing a Bayesian model average, in which each model corresponds to a different clustering choice.

An alternative approach, which like the Birge ratio avoids formulating any hypothesis on the nature of the missing systematic uncertainty, consists of estimating the uncertainty directly from the data dispersion with no particular assumption about the associated probability distribution [3, 9]. Datum-by-datum, the associated uncertainty $\tilde{\sigma}_i$ is obtained by a quadratic sum $\tilde{\sigma}_i = \sqrt{\sigma_i^2 + d_i^2}$ of the known uncertainty σ_i and the estimated missing uncertainty obtained from the difference $d_i = x_i - \hat{\mu}$ between the input value x_i and the standard weighted average itself. Since $\hat{\mu}$ is thus present in both the left and right expressions of Eq. (1), the final average is obtained recursively. Like the original formulation of the Birge ratio, this method lacks statistical foundations but has the benefit of being very simply formulated.

All the previously described methods share the implicit assumption that the uncertainty σ_i is a lower bound of the real uncertainty σ_i'. This simple and clear statement has been translated into formulas by Sivia and Skilling in 2004 [10], avoiding common scaling factors (like $R_{\rm Birge}$) or random bias dispersion (like $\beta_{\rm bias}$), but considering for each point a modification of the Gaussian distribution by the marginalisation over σ_i'. For this approach, a prior probability $p(\sigma_i')$ for σ_i' has to be chosen. The natural choice would be the non-informative Jeffreys' prior $p(\sigma_i') \propto 1/\sigma_i'$. If not constraint by an upper bound, this choice causes divergence because of the non-integrability of the resulting final probability distribution associated to each datum. To avoid this problem, a modified prior $p(\sigma_i') \propto 1/(\sigma_i')^2$ has been proposed and discussed in Refs. [6, 10, 11]. Alternatively, the inverse-gamma distribution, which is integrable but introduces additional parameters, is proposed as a prior distribution [1, 12]. Other more complex approaches with no lower bounds for σ_i' can be found in Refs. [1, 10].

Another approach involves the assumption of a likelihood that is not Gaussian. For example, Dose [13] proposes using Laplace and hyperbolic cosine likelihoods in addition to the Gaussian likelihood and averaging the three distributions by applying Bayesian model selection.

Here, we consider an approach with very few assumptions, similar to the work of Sivia and Skilling [10], but adopt the Jeffreys' non-informative prior, like in Sivia's earlier work [14]. Our intent is not to propose an automatic approach to replace the work of data selection, but to offer an easily implementable method as a common reference for averaging, while avoiding the issues associated with the standard weighted average. For these reasons, like the standard weighted average, we keep the basic assumptions simple and very general, with the risk of being too conservative in the estimation of the final uncertainty in certain cases. To facilitate the application of the proposed method in daily data analysis, a Python code is provided.

Details on the derivation of the method are presented in Sec. II. In Section III, the proposed method is compared to the other methods in a series of cases using simulated and real data. In section IV, the Python library based on the introduced method is presented. The final section is devoted to the discussion and conclusion.

II. DERIVATION OF THE WEIGHTED AVERAGE FOR INCONSISTENT DATA

A. General considerations

The standard weighted average of independent measurements x_i with uncertainties σ_i is obtained by maximising the total probability of the mean value μ, which results from the product of the single probabilities for each (assumed independent) measurement, with

$$p(\mu|\{x_i,\sigma_i\}) = \prod_i p(x_i|\mu,\sigma_i)\ p(\mu). \tag{4}$$

When a Gaussian distribution is considered for each x_i together with a flat prior probability for μ (a Jeffreys' prior), the most probable value $\hat{\mu}$ is given by the standard weighted average, i.e., Eq. (1). The associated uncertainty $\sigma_{\hat{\mu}}$ given in Eq. (2) is simply derived by the uncertainty propagation in $\hat{\mu} = f(x_1, x_2, \ldots)$ of the uncertainty of the single x_i values (see e.g. [15, 16]).

An alternative derivation of Eq. (2) can be obtained from the second derivative of the logarithm of $p(\mu|\{x_i,\sigma_i\})$ by supposing that the final probability distribution can be well approximated by a Gaussian distribution, where

$$\sigma_{\hat{\mu}} = \left(-\frac{\partial^2}{\partial \mu^2} \log[p(\mu|\{x_i,\sigma_i\})] \Big|_{\mu=\hat{\mu}} \right)^{-\frac{1}{2}}. \tag{5}$$

In line with standard methods, we consider the average as the value $\hat{\mu}$ that maximises Eq. (4) (assuming the Jeffreys' prior $p(\mu) = \text{const.}$), with its uncertainty given by Eq. (5). The single probabilities $p(x_i|\mu,\sigma_i)$ are no longer Gaussian, but are instead obtained by assuming certain hypotheses on the priors and performing marginalisations.

As wrote above, our final goal is to provide the scientific community with a new accessible tool to obtain a robust weighted average that can be easily understood.

Two prerogatives are thus essential: to propose something very general and to consider a minimal number of assumptions. The generality is particularly important to treat very common but different scenarios of inconsistent data averaging: i) from measurements obtained with the same apparatus (with a common uncontrolled systematic effect) or ii) from different types of measurements in different laboratories (with possible uncorrelated biases). For these purposes, we adopt a pessimistic framework where the uncertainty σ_i is regarded as a lower bound of the real uncertainty σ'_i, without any additional assumptions on the possible biases and relations influencing the available data set (similarly to Refs. [10, 17]). Any systematic error is considered to be included in the unknown uncertainty σ'_i.

Because of the unknown value of σ'_i of each measurement x_i, the associated probability distribution is obtained by marginalising over σ'_i:

$$p(x_i|\mu,\sigma_i) = \int_{\sigma_i}^{\infty} p(x_i|\mu,\sigma'_i)p(\sigma'_i|\sigma_i)\mathrm{d}\sigma'_i. \tag{6}$$

If only pairs (x_i,σ_i) of measured values and associated uncertainties are available, following the maximum entropy principle, a Gaussian distribution $p(x_i|\mu,\sigma_i)$ can be assumed for each datum. A choice for the prior probability distribution $p(\sigma'_i|\sigma_i)$ for σ'_i has to be made. The natural choice is a Jeffreys' prior, which is a non-informative prior that is invariant under reparametrisation, to avoid introducing other possible biases, with

$$\begin{cases} p(\sigma'_i|\sigma_i) = \dfrac{1}{\log(\sigma_i^{\max}/\sigma_i)}\dfrac{1}{\sigma'_i} & \text{for } \ \sigma'_i \leq \sigma_i \leq \sigma_i^{\max}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The problem of such a prior is the introduction of an additional parameter $\sigma_i^{\max}$ for each data point. Indeed, if one tries to eliminate such additional parameters by considering the limit $\sigma_i^{\max} \to \infty$, $p(\sigma'_i|\sigma_i)$ is no longer a proper probability distribution, i.e. normalised to unity, because $\int_{\sigma_i}^{\infty} 1/\sigma'_i \mathrm{d}\sigma'_i = \infty$. Two alternative solutions to this issue are presented in the next paragraphs.

B. Sivia and Skilling's conservative weighted average

To avoid the introduction of additional parameters $\sigma_i^{\max}$ required for the normalisation of Jeffreys' prior, a *conservative formulation* has been proposed by Sivia and Skilling for general regression problems [10]. It consists of the modified version of the Jeffreys' prior

$$p(\sigma'_i|\sigma_i) = \frac{\sigma_i}{(\sigma'_i)^2}. \tag{8}$$

Keeping the assumption of a Gaussian distribution for $p(x_i|\mu,\sigma'_i)$ and combining Eqs. (6) and (8), we obtain

$$p(x_i|\mu,\sigma_i) = \frac{\sigma_i}{\sqrt{2\pi}}\left[\frac{1-e^{\frac{(x_i-\mu)^2}{2\sigma_i^2}}}{(x_i-\mu)^2}\right]. \tag{9}$$

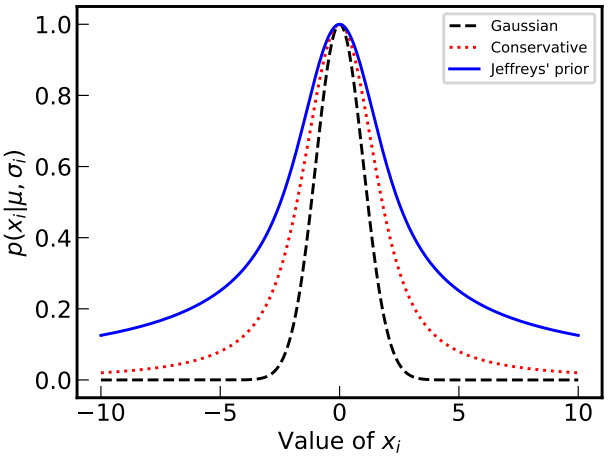


FIG. 1. Comparison between the different assumed probability distribution for each datum for $\mu = 0, \sigma_i = 1$.

Compared with a Gaussian distribution, as visible in Fig. 1, the above expression is characterised by a significantly larger spread, with tails proportional to $1/(x_i)^2$. Once plugged into Eq. (4), such slowly descending tails supply sufficient flexibility to be tolerant of inconsistent data. Unlike the standard weighted average, the maximising value $\hat{\mu}$ and its associated uncertainty $\sigma_{\hat{\mu}}$ have no analytical form, but can be easily determined with numerical methods. Now, both $\hat{\mu}$ and $\sigma_{\hat{\mu}}$ depend on the data spread. Note that, because of the presence of the tails, even for consistent data sets, the final uncertainty of the weighted average is generally greater than the one obtained by the standard method.

C. Limit solution with Jeffreys' prior

A criticism that could be directed at Eq. (9) is that the choice of the prior probability distribution of σ'_i in Eq. (8) does not respect the non-informative criterion of Jeffreys' prior. A possible solution to keep Jeffreys' prior without introducing any new parameters is given by Sivia in Ref. [14]. It considers the limit case $\sigma_i^{\max} \to \infty$. The divergence of the resulting probability distribution is circumvented by considering only its maximum and the second-order log-derivative for the estimation of the limit value of the weighted average and the associated uncertainty, respectively.

When the Jeffreys' prior from Eq. (7) is adopted, Eq. (6) becomes

$$p(x_i|\mu,\sigma_i) = \frac{1}{\log(\sigma_i^{\max}/\sigma_i)}\frac{\mathrm{erf}\left(\frac{x_i-\mu}{\sqrt{2}\sigma_i}\right) - \mathrm{erf}\left(\frac{x_i-\mu}{\sqrt{2}\sigma_i^{\max}}\right)}{2(x_i-\mu)}. \tag{10}$$

Compared to Eq. (9), we can note that the distribution

tails decrease even more smoothly than in the conservative approach, with a dependency of $1/x_i$ instead of $1/x_i^2$, which can tolerate the presence of outliers even better. These more pronounced tails are well visible in Fig. 1. The logarithmic form of the total probability $p(\mu|\{x_i, \sigma_i, \sigma_i^{\max}\})$ is thus given by

$$\log[p(\mu|\{x_i, \sigma_i, \sigma_i^{\max}\})] = \sum_i \log\left[\frac{\mathrm{erf}\left(\frac{x_i-\mu}{\sqrt{2}\sigma_i}\right) - \mathrm{erf}\left(\frac{x_i-\mu}{\sqrt{2}\sigma_i^{\max}}\right)}{2(x_i-\mu)}\right] - C, \quad (11)$$

where there is now a dependency on $\{\sigma_i^{\max}\}$, and

$$C = \sum_i \log\left[\log(\sigma_i^{\max}/\sigma_i)\right]. \quad (12)$$

This constant term does not play a role in the search for $\hat{\mu}$, as C depends only on the boundary values σ_i and $\sigma_i^{\max}$.

The limit $\sigma_i^{\max} \to \infty$ of the above equation is

$$\log[p(\mu|\{x_i, \sigma_i\})] = \lim_{\sigma_i^{\max}\to\infty} \log[p(\mu|\{x_i, \sigma_i, \sigma_i^{\max}\})] = \sum_i \log\left[\frac{\mathrm{erf}\left(\frac{x_i-\mu}{\sqrt{2}\sigma_i}\right)}{2(x_i-\mu)}\right] - C^{\infty}, \quad (13)$$

where the constant $C^{\infty} = \lim_{\sigma_i^{\max}\to\infty} C = \infty$ is indeed divergent, but for more than one data point the distribution is integrable, and the variance is finite for more than two data points. In particular, the position of the maximum and the value of the second derivative and thus the weighted average and its uncertainty are well defined. As in the case of the conservative weighted average, no analytical solution is available for $\hat{\mu}$ and $\sigma_{\hat{\mu}}$, so the solution must be found numerically.

We consider the formulas presented in this section as a new reference method for the weighted average. In fact, the conservative approach of Sec. II B still involves some arbitrariness in the choice of the prior for the average. This is not the case for the above formulas, where the universal choice of the Jeffreys' prior is adopted. In the following section, we will refer to this approach as the *Jeffreys' weighted average*.

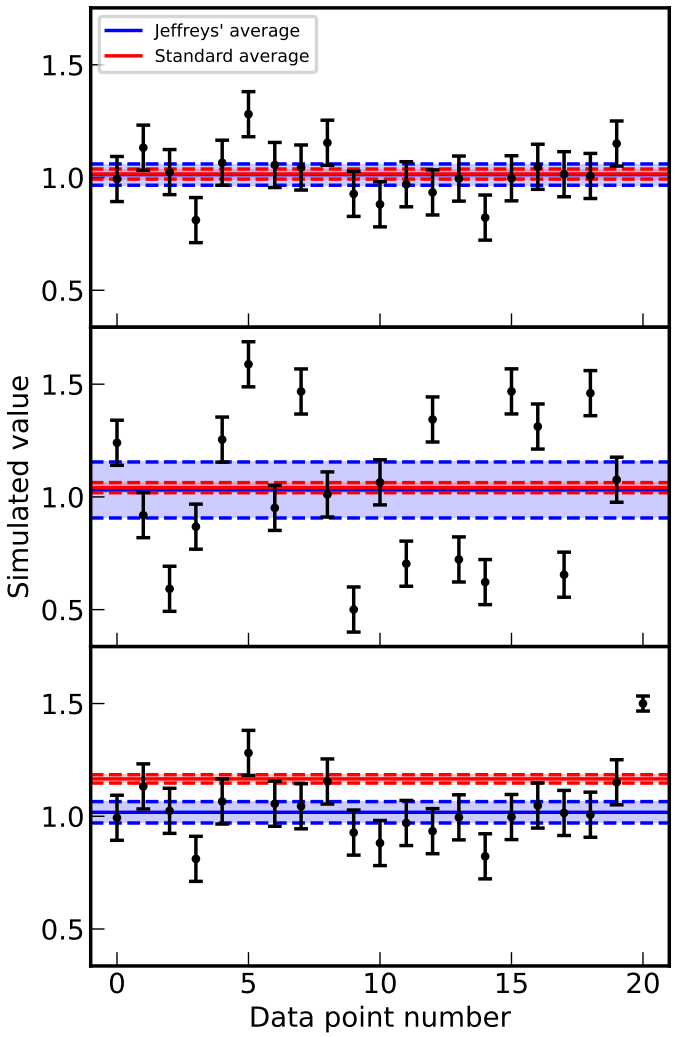


FIG. 2. Standard and Jeffreys' weighted averages of different simulated data sets: data randomly sampled from a normal distribution (top), and with the addition of a random bias (middle) or an outlier (bottom).

III. SOME APPLICATIONS

In this section, we present a series of applications for common data analysis cases. In the first subsection, we will study simulated data with known theoretical values of mean and standard deviation. In the second subsection, an analysis of the different values of the Newtonian gravitational constant from past CODATA compilations is proposed. The third subsection is dedicated to fundamental particle properties, including the controversial data set of the proton charge radius. Additional calculations of the Jeffreys' weighted average for inconsistent data measured using the same experimental apparatus can be found in Ref. [18].

A. Synthetic tests

Different simulated data sets are considered for comparing the different averaging methods:

1. The first set of values x_i is obtained by a normal distribution with a mean value of $\mu = 1$ and a standard deviation equal to $\sigma = 0.1$. For each data point, the uncertainty $\sigma_i = \sigma$ is considered.

2. The second set simulates inconsistent data. It is derived from set 1 by adding a random bias, with

TABLE I. Weighted average values with corresponding uncertainties for the synthetic data set 1 (data shown in Fig. 2, top).

Weighted average type	Average $\hat{\mu}$	Unc. $\sigma_{\hat{\mu}}$
Inverse-variance (standard)	1.015	0.022
Inverse-variance with Birge ratio	1.015	0.025[a]
Conservative	1.013	0.037
Jeffreys' prior	1.013	0.047

[a]Birge ratio equal to 1.1

TABLE II. Weighted average values with corresponding uncertainties for the synthetic data set 2 (data shown in Fig. 2, middle).

Weighted average type	Average $\hat{\mu}$	Unc. $\sigma_{\hat{\mu}}$
Inverse-variance (standard)	1.041	0.022
Inverse-variance with Birge ratio	1.041	0.076[a]
Conservative	1.04	0.094
Jeffreys' prior	1.031	0.124

[a]Birge ratio equal to 3.4

a standard deviation of $\sigma_{\text{bias}} = 10\,\sigma$ and a mean of $\mu_{\text{bias}} = 0$, to each data point.

3. The third set is the same as the first, but with the addition of an outlier at $\mu + 5\,\sigma$ that has an uncertainty of $\sigma_{\text{out}} = \sigma/3$.

For the three sets, the standard weighted average, with or without Birge ratio correction, is compared to the conservative and Jeffreys' weighted averages. The results are presented in Tables I–III and in Fig. 2 (for the standard and Jeffreys' weighted averages only) and Fig. 3. As we can see, for normal and inconsistent data without outliers, the mean value is well reproduced by all methods. Because of the pessimistic priors on σ_i', the Jeffreys' and conservative final uncertainties are generally larger than the standard uncertainty. For consistent data (set 1), they are larger by a factor of about two. As expected, for the inconsistent data (set 2), the uncertainty associated with the standard weighted average is significantly smaller than the others, with the Jeffreys' uncertainty being the largest, followed by the conservative and the Birge uncertainties. Jeffreys' weighted average uncertainty is almost six time larger than the standard weighted average uncertainty and double that of the Birge-ratio-corrected uncertainty.

TABLE III. Weighted average values with corresponding uncertainties for the synthetic data set 3 (data shown in Fig. 2, bottom, detailed analysis in Fig. 3).

Weighted average type	Average $\hat{\mu}$	Unc. $\sigma_{\hat{\mu}}$
Inverse-variance (standard)	1.166	0.019
Inverse-variance with Birge ratio	1.166	0.054[a]
Conservative	1.019	0.037
Jeffreys' prior	1.018	0.047

[a]Birge ratio equal to 2.9

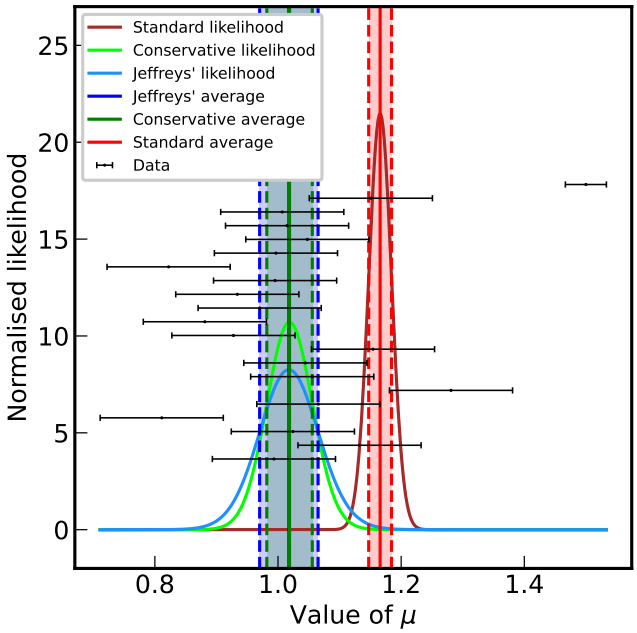


FIG. 3. Final likelihood distribution for the data set 3 together with the input values.

When an outlier is present, Jeffreys' weighted average values are quite different from the standard weighted average. The effect of the presence of an outlier (set 3) is clearly visible in Fig. 3, where the final likelihoods are plotted together with the data, and in the results presented in in Table III. As we can see, the effect on the standard likelihood, obtained by the product of Gaussian distributions, is drastic, with a shift in the direction of the outlier. If the data uncertainties are regarded as lower bounds only, the effect is greatly mitigated, resulting in just an asymmetry of the tails for the Jeffreys' and conservative priors, which have very similar final probability distributions. This behaviour looks quite similar to other methods that utilize the deviation of the data point from the calculated value $\hat{\mu}$ to estimate the possible missing uncertainty contributions [3, 9]. However, unlike these past works, here the distributions are derived from the initial assumptions on the uncertainties, for which we consider *a priori* that the provided values σ_i are only lower bounds of the real uncertainty.

B. The Newtonian constant of gravity

A significant example of an average between independent and possibly inconsistent measurements is the determination of the Newtonian constant of gravity, which, due to the difficulties associated with its measurement, has long been the fundamental constant with the high-

TABLE IV. Weighted average values with corresponding uncertainties for the data set of the CODATA 1998 compilation for the Newtonian gravitational constant and corresponding recommended value.

Weighted average type	Average $\hat{\mu}$	Unc. $\sigma_{\hat{\mu}}$
Inverse-variance (standard)	6.6827	0.0003
Inverse-variance with Birge ratio	6.6827	0.0063[a]
Conservative	6.6735	0.0006
Dose model average [13]	6.6746	0.0035
Jeffreys' prior	6.6736	0.0008
CODATA 1998 [24]	6.673	0.010[b]

[a] Birge ratio equal to 23.6
[b] Corresponding scale factor equal to 37

est relative uncertainty. Such difficulties are mainly due to the challenging experimental conditions, where very small forces must be isolated from a noisy environment [19, 20]. The official value is provided by CODATA with a standard inverse-variance weighted average, and the associated uncertainty is eventually multiplied by an expansion factor to maintain consistency between the final result and the considered measurements.

Here, we apply the proposed averaging method to all data sets included in the different editions of the CODATA compilation [21–30]. The results are presented together with the official values in Fig. 4. For each reported CODATA value, the large error bar corresponds to the recommended value of the uncertainty, and the small one to the uncertainty calculated by the standard weighted average. As we can see, the standard weighted average is, for some years, several standard deviations away from the most recent CODATA value from 2018 (the horizontal dashed line in the figure, which is the same as the more recent CODATA 2022 value [30]), considered here as the reference value. Contrary to standard procedures, one can see that the values obtained by the Jeffreys' weighted average are consistently in good agreement, being less than one standard deviation away from the most recent CODATA value, and are characterised by a more plausible uncertainty.

The 1998 case is particularly difficult due to the inconsistency within the data set, arising from one very precise measurement [31] that differs significantly from the average of the other measurements. The value was later found to be affected by a large systematic error [32]. Details of the analysis of this specific case are presented in Table IV and Fig. 5. The CODATA recommended value is obtained by the standard weighted average of all values, excluding the suspicious measurement. The corresponding uncertainty is obtained by applying an expansion factor of 37 to the standard weighted average uncertainty to reflect the presence of the outlier. More precisely, the final uncertainty has been chosen to ensure that the difference between the recommended value and the outlier is four times larger than the final uncertainty. Like in set 3 of the previous section, the Jeffreys' weighted average is only slightly affected by the outlier in this challenging case. Looking at Table IV, one can note that the Jeffreys' weighted average has a significantly smaller uncertainty than the method tested on the same data set by Dose [13]. There, a Bayesian model average was applied on different likelihood functions.

C. Particle properties

Another field that has to deal with very different measurements to compile reference data is particle physics. The Particle Data Group (PDG) [33], who is providing the official reference values, implements a very well documented procedure, mainly based on the Birge ratio and data selection. The goal of this selection is to minimise possible correlations between considered data and to exclude evident outliers. The muon magnetic anomalous moment is an exception. No measurements are selected except the one considered as unique reference from the most recent experiment [34] (which is now in agreement with the Standard Model predictions [35]).

In this section, we evaluate the Jeffreys' weighted average of some particle properties, and compare the results with the PDG recommended values. Moreover, to test the robustness of the Jeffreys' weighted average, we evaluate two sets:

A: the measurements selected by the PDG for the determination of the recommended value.

B: the whole set of values listed by PDG.

For both sets, no correlations between the data have been considered. This assumption is, however, not well adapted to set B, where strongly correlated measurements are present in some cases.

The obtained average values are compared with the PDG values in Fig. 6. Because of the very different quantities, we normalise the difference of the Jeffreys' weighted average to the PDG value by the uncertainty provided by the PDG. It comes as no surprise that, when the selected data are considered, the Jeffreys' weighted average is in good agreement with the PDG values. Similar to the case of the simulated data sets, the associated final uncertainty is generally larger than the PDG uncertainty by at most a factor of 2.2. For set B, smaller values of the final uncertainty can also be found because of the larger set of considered data. For both sets, deviations of less than two standard deviations are observed, indicating the good robustness of the Bayesian method even in difficult cases. The exception of the deviation of $K^{\pm}$ meson mass for set B is caused by very strong correlations between measurements. Moreover, the deviation of the neutron asymmetry parameter B is caused by a single additional value in set B, which is excluded by the PDG (set A). Like the case of the Newtonian constant in CODATA 1998, this very precise additional value has a strong influence due to the lack of precise measurements.

However, even if in good agreement with PDG values, some results must be treated with caution. In contrast to

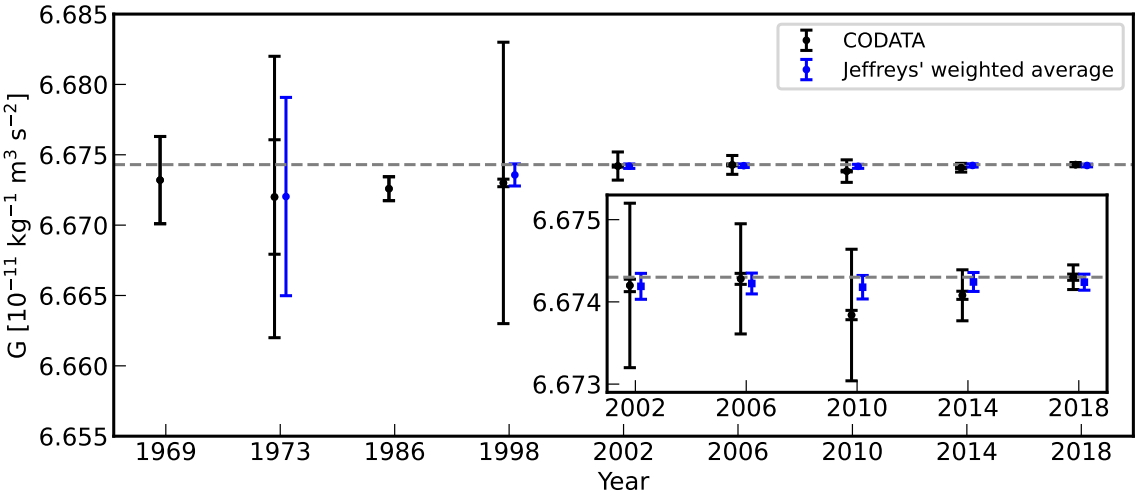


FIG. 4. Comparison between the official CODATA values [21–29] of the Newtonian constant and the values obtained by the Bayesian weighted average using Jeffreys' prior. CODATA values obtained from single measurements are presented alone, as no weighted average could be performed. The small error bar of the CODATA values indicates the uncertainty calculated by the standard weighted average, and the large one indicates the recommended uncertainty. The horizontal dashed line corresponds to the latest CODATA value (2022 edition [30], equal to the 2018 edition value).

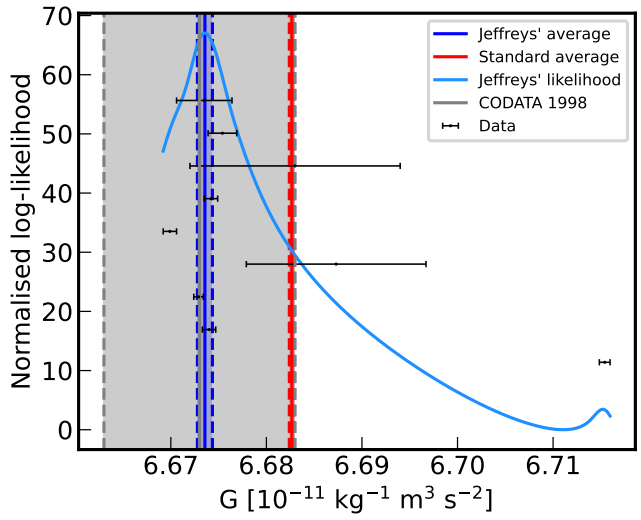


FIG. 5. Final likelihood distribution (in log-scale) of the measurements of the Newtonian constant included in the CODATA 1998 compilation [24]. The CODATA 1998 recommended value is also reported (in gray), which differs from the standard weighted average for the considered measurements (in red).

the standard weighted average, typically associated with a sharp (Gaussian) probability distribution, the probability distribution for the Jeffreys' prior may more readily exhibit strong asymmetry and multimodality. As an example, the deviation in the neutron lifetime in set B is caused by an asymmetry in the final probability distribution. Similar cases are found for the $K^{\pm}$ meson mass (set A only), the neutron lifetime (set A only), neutron asymmetry parameter A, muon mass, and the e^- magnetic moment anomaly. For the latter in particular, no deviation from the PDG value is visible. Two typical examples are shown in Fig. 7, presenting the neutron lifetime and the charged kaon mass in detail. As we can see, the use of any kind of weighted average is not appropriate because it does not reflect the final probability distribution, which should be considered for further inferences instead. These considerations are in complete agreement with the PDG recommendations that for these non-trivial cases point out possible issues with these sets and provide an ideogram corresponding to the combination of the measurement results (assumed to be Gaussian with a weight proportional to $1/\sigma_i$, and not to $1/\sigma_i^2$ like for the standard weighted average) to underline the importance of the single values in the average. Unlike the standard and PDG methods, such a conclusion can be directly deduced by looking at the final probability distribution for the Jeffreys' prior.

An extreme example of asymmetry and multimodality in the final probability distribution is the case of the proton rms charge radius. For this quantity, very different results are obtained from the Jeffreys' weighted average with respect to the recommended PDG value and the standard weighted average (so different that they are not reported in Fig. 6). As we can see from the plot of the probability distribution (Fig. 8), a pronounced bi-modal distribution appears when evaluating the entire ensemble of available data. In this situation, any weighted average, for which we always assume unimodality of the associated

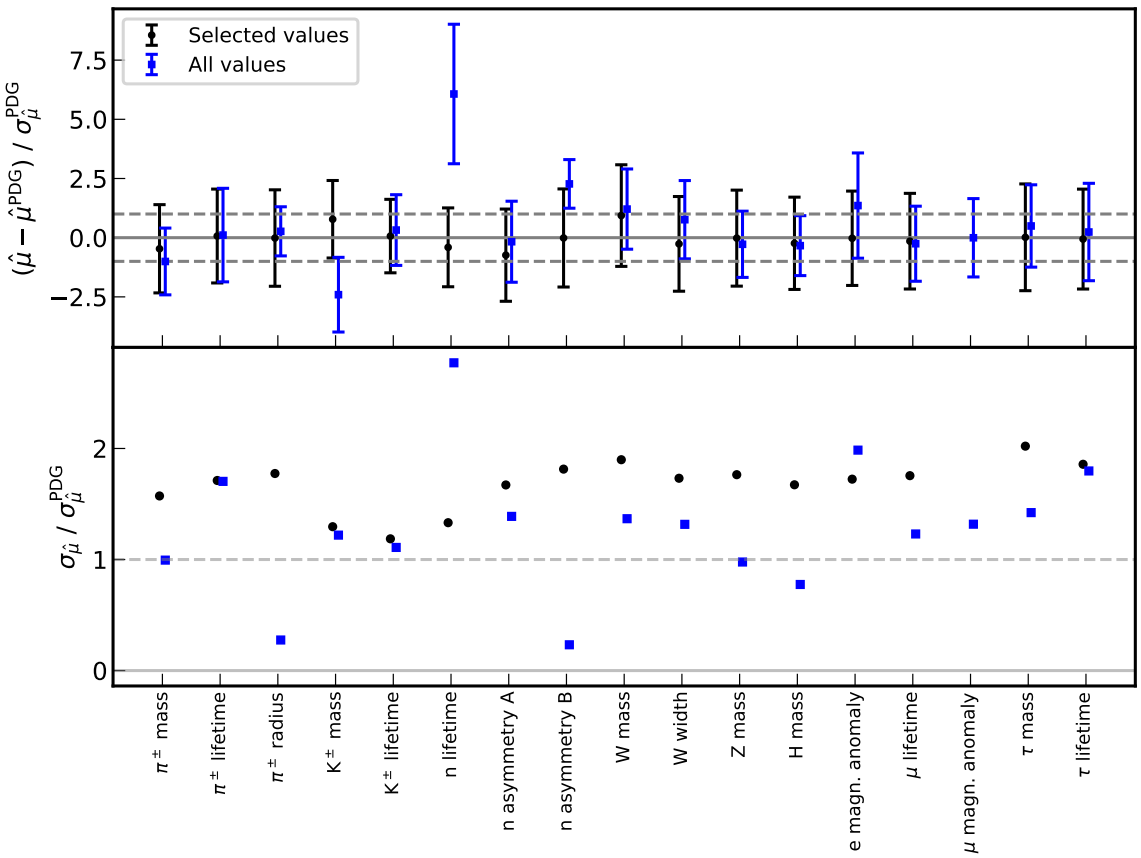
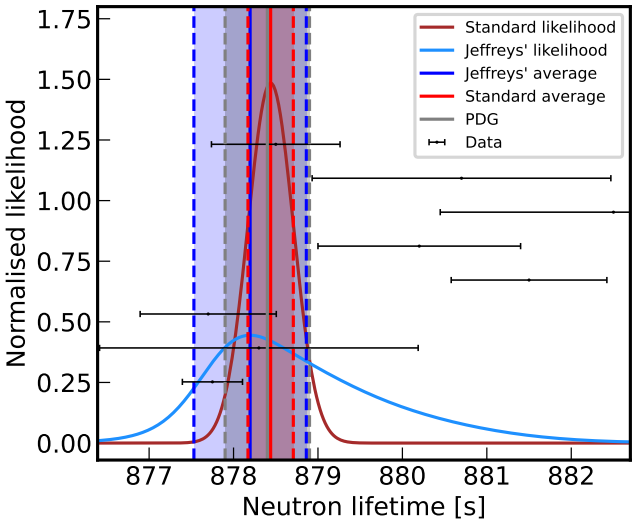


FIG. 6. Top: Comparison between the values obtained by the Bayesian weighted average using Jeffreys' prior and the official PDG values of several particle properties [33]. All values are normalised to the PDG uncertainties. The two dashed horizontal lines correspond to $\hat{\mu} - \hat{\mu}^{\rm PDG} = \pm\sigma_{\hat{\mu}}^{\rm PDG}$, i.e. to ± 1 in the relative scale. Bottom: Ratio between the uncertainties obtained for the Jeffreys' weighted average and the official PDG uncertainties.

probability distribution, is not suitable. This situation is similar to the case of a data set composed of only two data points with different values but the same uncertainties. Regardless of the chosen method, a weighted average will propose the midpoint between the data points due to the symmetry of the problem. In such cases, as written above, the whole probability distribution should be taken into account.

IV. THE ASSOCIATED CODE

Despite the well-known problems of the standard weighted average based on the inverse of the variance, its still widespread use is undoubtedly largely due to the simplicity of its formula, which can be easily employed by anyone. The method presented here, as well as others proposed in the literature, has the disadvantage of requiring numerical methods for determining the weighted average and its uncertainty due to the complexity of the associated analytical formulas. The need for improved averaging methods is, however, prevalent, but due to the implementation difficulties, such alternative methods are generally quickly abandoned, sometimes in favour of the simpler Birge ratio. To close this gap, we provide a numerical tool for the proposed weighted averaging method. More precisely, we propose the Python library BAYESIAN_AVERAGE, which can be easily installed in any Python environment using the `pip` command and is also freely available via the GitHub repository [36]. In addition to providing the weighted average based on the Jeffreys' and conservative method priors, the standard inverse-variance method and its modified version with the Birge ratio are included for comparison. A graphical tool is available to plot the final weighted averages and the associated final likelihood probability distributions together with the input data. Figures 3, 5, 7 and 8 are typical outputs from BAYESIAN_AVERAGE (with minimal changes to the label of the axes).



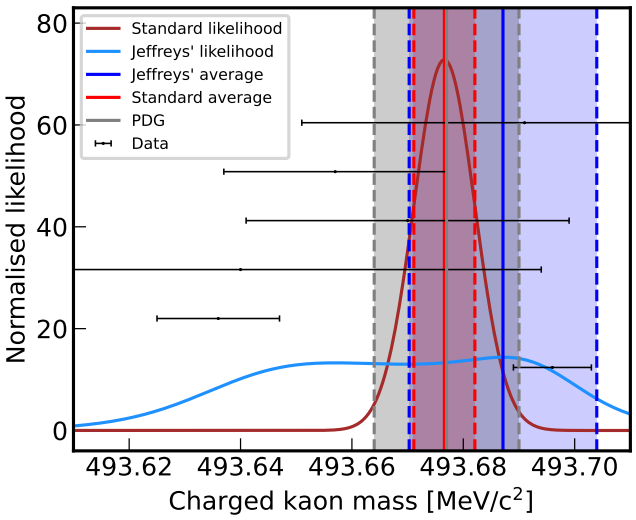


FIG. 7. Final likelihood distribution (in linear scale) corresponding to the PDG measurement selection for the neutron lifetime (top) and the charged kaon mass (bottom). In these cases, the final probability distribution generated by the Bayesian method is highly asymmetric and cannot be approximated by a Gaussian distribution. Therefore, the whole probability distribution should be considered for further inference analyses instead of any kind of weighted averages.

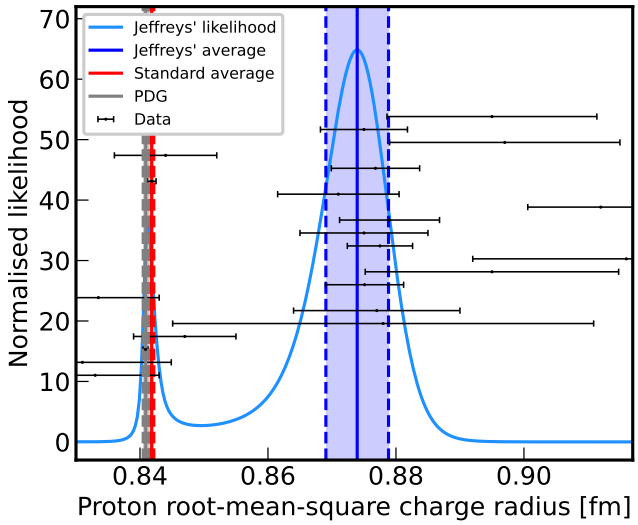


FIG. 8. Final likelihood distribution (in linear scale) corresponding to the PDG measurement selection for the proton radius. The PDG official value is also reported, which is very close to the standard weighted average. Due to the bimodality of the final probability distribution, using a weighted average is completely inappropriate.

V. DISCUSSION AND CONCLUSIONS

We present a robust method for averaging inconsistent data, originally proposed by Sivia [14]. We suggest to use such a method as an alternative to the standard weighted average based on the inverse of the variance. In addition, we propose a Python package to easily calculate such a weighted average.

Compared with other similar methods previously proposed in the literature, the number of working hypotheses is kept to the minimal requirements. The weighted average based on Bayesian statistics is proposed to avoid formulating complex hypotheses on the nature and behaviour of the unknown component of the true uncertainties. For each data point, a Gaussian distribution is considered, but the provided uncertainty σ_i is regarded as a lower bound of the true uncertainty value. Using Bayes' theorem and assuming a non-informative Jeffreys' prior for σ'_i, a new probability distribution is obtained by marginalising over $\sigma'_i \in [\sigma_i, \sigma_i^{\max}]$. The new arbitrary parameters $\sigma_i^{\max}$ are eliminated by looking at the asymptotic solution of the resulting weighted average $\hat{\mu}$. Contrary to the final associated probability distribution that diverges, the limit value of $\hat{\mu}$ and the associated uncertainty $\sigma_{\hat{\mu}}$ are well defined.

The proposed method based on the Jeffreys' prior is applied to a series of cases that show its reliability and robustness. For this purpose, simulated data, CODATA values of the Newtonian constant, as well as a series of particle property values from PDG are considered. In particular, for the CODATA 1998 case, where an outlier was causing issues, the method has proven to be a very robust tool. In the case of particle properties, different scenarios are encountered. For the largest part of the cases, the Jeffreys' weighted average reproduces very well the PDG recommended values, but with a slightly larger uncertainty. In some cases, however, we show that a weighted average procedure should be taken with caution and the entire probability distribution should be considered instead, which is in agreement with the PDG recom-

mendations. This is particularly true for the case of the proton radius, which shows a pronounced multimodality in the corresponding final probability. With the proposed method, these difficult cases are easily identified by looking at the final probability distribution.

The focus of future developments will be on incorporating correlations between the input data for the calculation of the weighted average.

Author contributions

The original idea was formulated by M.T. The analysis was mainly performed by M.M. The manuscript was written by both authors.

Data availability

The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

Code availability

The weighted averages were calculated using the publicly available code BAYESIAN_AVERAGE from the author, which is accessible in the repository `https://github.com/martinit18/bayesian_average`.

ACKNOWLEDGMENTS

We would like to thank François Nez for the very useful discussions on CODATA values, Louis Duval for providing the motivation to start this work with his analysis of "inconsistent" data, Mark Plimmer for the careful reading of the manuscript, and Andrew Charman for introducing us to Sivia's earlier work that confirmed our independent calculations.

[1] von der Linden W, Dose V and von Toussaint U 2014 *Bayesian Probability Theory: Applications in the Physical Sciences* (Cambridge University Press) ISBN 9781139952460
[2] Birge R T 1932 *Phys. Rev.* **40** 207–227
[3] Lira I 2007 *Metrologia* **44** 415
[4] Toman B, Fischer J and Elster C 2012 *Metrologia* **49** 567
[5] Bodnar O and Elster C 2014 *Metrologia* **51** 516
[6] Mana G, Massa E and Predescu M 2012 *Metrologia* **49** 492
[7] Bodnar O, Elster C, Fischer J, Possolo A and Toman B 2016 *Metrologia* **53** S46
[8] Rukhin A L 2019 *Metrologia* **56** 035002
[9] Huang H 2018 *Metrologia* **55** 106
[10] Sivia D S and Skilling J 2006 *Data analysis: a Bayesian tutorial* 2nd ed (Oxford University Press)
[11] Shirono K, Tanaka H and Ehara K 2010 *Metrologia* **47** 444
[12] Dose V 2003 *Rep. Prog. Phys.* **66** 1421
[13] Dose V 2007 *Meas. Sci. Technol.* **18** 176
[14] Sivia D 1996 Dealing with duff data *Proceedings of Maximum Entropy and Bayesian Methods Conference* (Port Elizabeth, South Africa: NMB printers) pp 131–137
[15] Bevington P and Robinson D 2003 *Data reduction and error analysis for the physical sciences* (McGraw-Hill) ISBN 9780072472271
[16] Taylor J 1997 *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (University Science Books) ISBN 9780935702750
[17] Mana G 2021 *Metrologia* **58** 055002
[18] Duval L, Lamour E, Macé S, Machado J, Maxton M, Paul N, Prigent C, Trassinelli M and Indelicato P 2024 *Eur. Phys. J. D* **78** 116
[19] Rothleitner C and Schlamminger S 2017 *Rev. Sci. Instrum.* **88**
[20] Merkatas C, Toman B, Possolo A and Schlamminger S 2019 *Metrologia* **56** 054001
[21] Taylor B N, Parker W H and Langenberg D N 1969 *Rev. Mod. Phys.* **41** 375–496
[22] Cohen E R and Taylor B N 1973 *J. Phys. Chem. Ref. Data* **2** 663–734
[23] Cohen E R and Taylor B N 1987 *Rev. Mod. Phys.* **59** 1121–1148
[24] Mohr P J and Taylor B N 2000 *Rev. Mod. Phys.* **72** 351–495
[25] Mohr P J and Taylor B N 2005 *Rev. Mod. Phys.* **77** 1–107
[26] Mohr P J, Taylor B N and Newell D B 2008 *Rev. Mod. Phys.* **80** 633–98
[27] Mohr P J, Taylor B N and Newell D B 2012 *Rev. Mod. Phys.* **84** 1527–1605 URL `http://link.aps.org/doi/10.1103/RevModPhys.84.1527`
[28] Mohr P J, Newell D B and Taylor B N 2016 *Rev. Mod. Phys.* **88** 035009
[29] Tiesinga E, Mohr P J, Newell D B and Taylor B N 2021 *J. Phys. Chem. Ref. Data* **50** 033105
[30] CODATA Task Group on Fundamental Constants 2024 Codata recommended values of the fundamental physical constants: 2022 `https://pml.nist.gov/cuu/Constants/`
[31] Michaelis W, Haars H and Augustin R 1995 *Metrologia* **32** 267
[32] Winfried M, Jürgen M and Holger H 2004 *Metrologia* **41** L29
[33] Navas S, Amsler C, Gutsche T and *et al* (Particle Data Group Collaboration) 2024 *Phys. Rev. D* **110**(3) 030001 URL `https://link.aps.org/doi/10.`

1103/PhysRevD.110.030001
[34] Aguillard D P, Albahri T, Allspach D and *et al* (The Muon $g-2$ Collaboration) 2023 *Phys. Rev. Lett.* **131**(16) 161802 URL https://link.aps.org/doi/10.1103/PhysRevLett.131.161802
[35] Ignatov F V, Akhmetshin R R, Amirkhanov A N and *et al* (CMD-3 Collaboration) 2024 *Phys. Rev. Lett.* **132**(23) 231903 URL https://link.aps.org/doi/10.1103/PhysRevLett.132.231903
[36] https://github.com/martinit18/bayesian_average