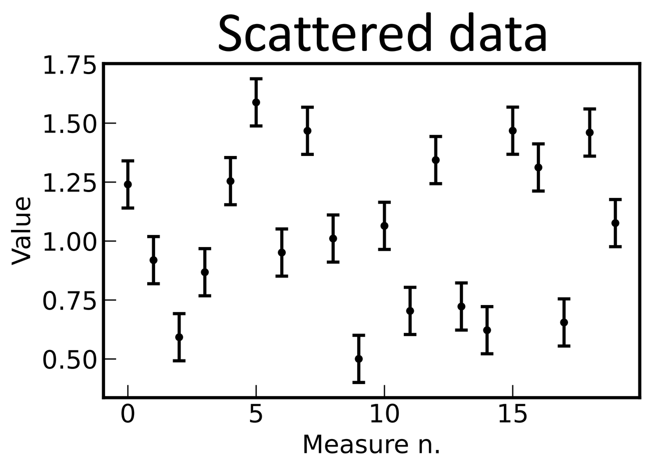
WEIGHTED AVERAGE OF SCATTERED DATA: A BAYESIAN APPROACH

M. Trassinelli, M. Maxton

Institut des NanoSciences de Paris, CNRS, Sorbonne Université, 75252 Paris, France

The problem of averages of inconsistent data

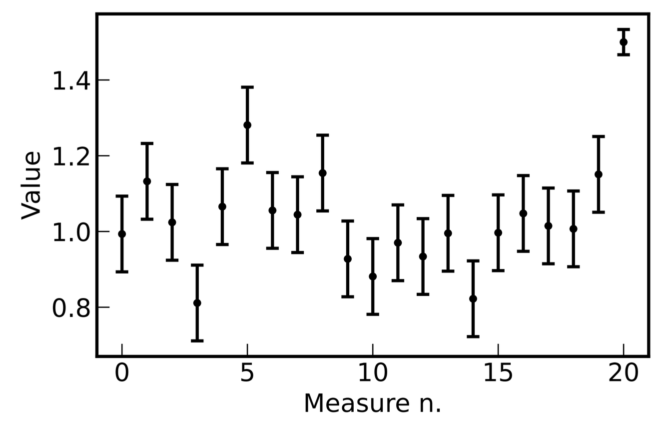
What to do with inconsistent sets of data?



Standard procedure:
inverse-invariance weighted average

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

or presence of outliers



BUT $\sigma_{\hat{\mu}}$ does not depends on the data spread, only on the uncertainties!!

with $\hat{\mu}$ maximising $p(\mu|\{x_i,\sigma_i\}) = \prod_i p(x_i|\mu,\sigma_i)$ ← Gaussian (normal) distribution assumed

and with $\sigma_{\hat{\mu}} = \frac{\partial^2}{\partial \mu^2} \log[p(\mu|\{x_i,\sigma_i\})]\Big|_{\mu=\hat{\mu}}$

Proposed alterative averages

- Increase all uncertainties to have a final $\chi^2_{\text{red}} \approx 1$ (Birge ratio [1])

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

But global adjustment, not adapted to inter-laboratory comparisons!!

- Evaluate the unknown uncertainty from the final spread [2,3]

But adjustment *a posteriori*

$$\hat{\mu} = \frac{\sum_{i=1}^m \frac{1}{\sqrt{d_i^2+\hat{\sigma}_i^2}} x_i}{\sum_{i=1}^m \frac{1}{\sqrt{d_i^2+\hat{\sigma}_i^2}}} \qquad \text{with} \quad d_i = x_i - \hat{\mu}$$

- More complex modelling of the missing systematic effects …

A Bayesian minimalistic approach

The simplest possible (and pessimistic) assumption [4]:
my uncertainty is only a lower-bound estimation of the real one

Marginalization over the unknown variable

$$\Rightarrow \quad p(x_i|\mu,\sigma_i) = \int_{\sigma_i}^{\infty} p(x_i|\mu,\sigma_i')p(\sigma_i'|\sigma_i)\mathrm{d}\sigma_i'$$

What to choose for $p(\sigma_i'|\sigma_i)$?

The "natural" choice: a non-informative (Jeffreys') prior:

Invariant over parameter transformation

$$\begin{cases} p(\sigma_i'|\sigma_i) = \frac{1}{\log(\sigma_i^{\max}/\sigma_i)}\frac{1}{\sigma'} & \text{for} \quad \sigma_i' \le \sigma_i \le \sigma_i^{\max}, \\ 0 & \text{otherwise.} \end{cases}$$

but additional parameters σ_i^{max} to add, not normalizable otherwise

Possible solution:

Sivia conservative prior [4]: $p(\sigma_i'|\sigma_i) = \frac{\sigma_i}{(\sigma_i')^2}$

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

No more a Gaussian distribution! but not a non-informative prior

Our solution:

Jeffreys' prior + marginalization

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

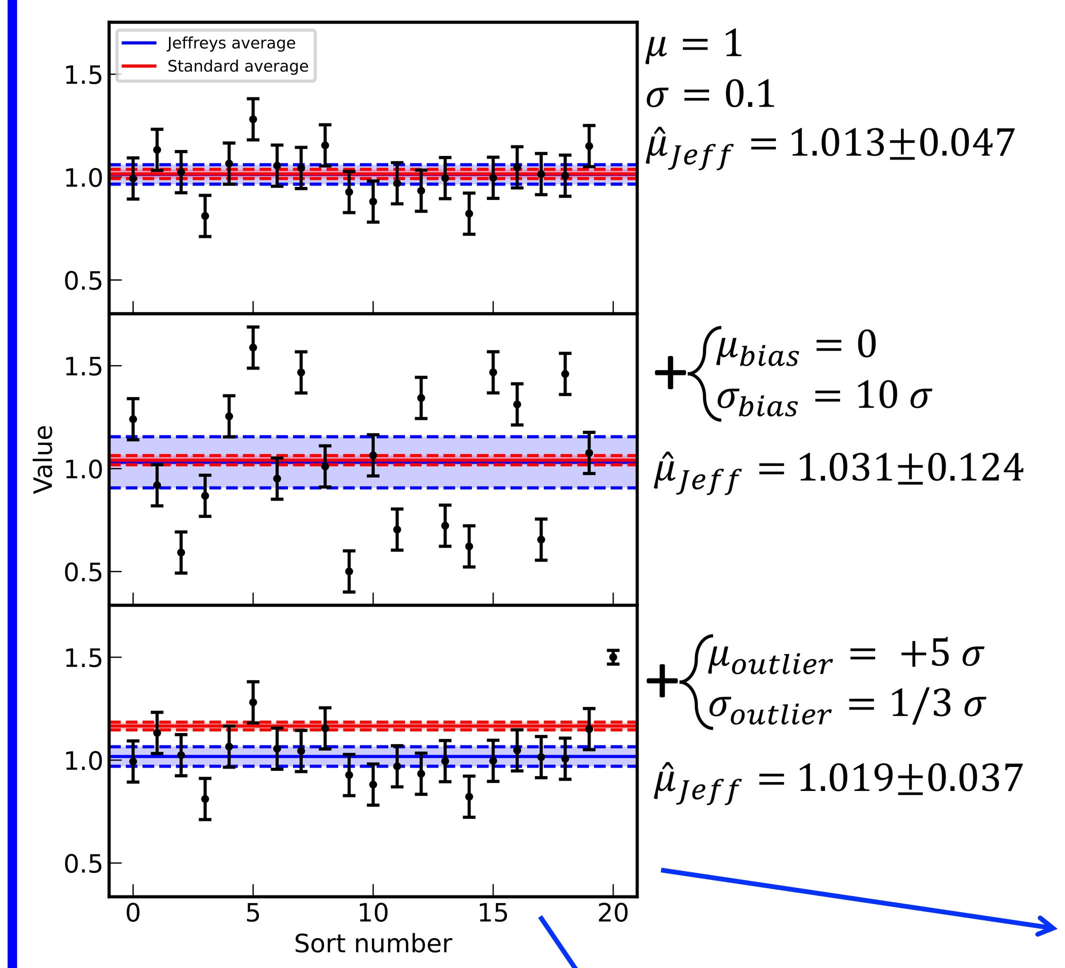
Limit case:

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

No more additional parameters, and maximum and second derivative still well defined

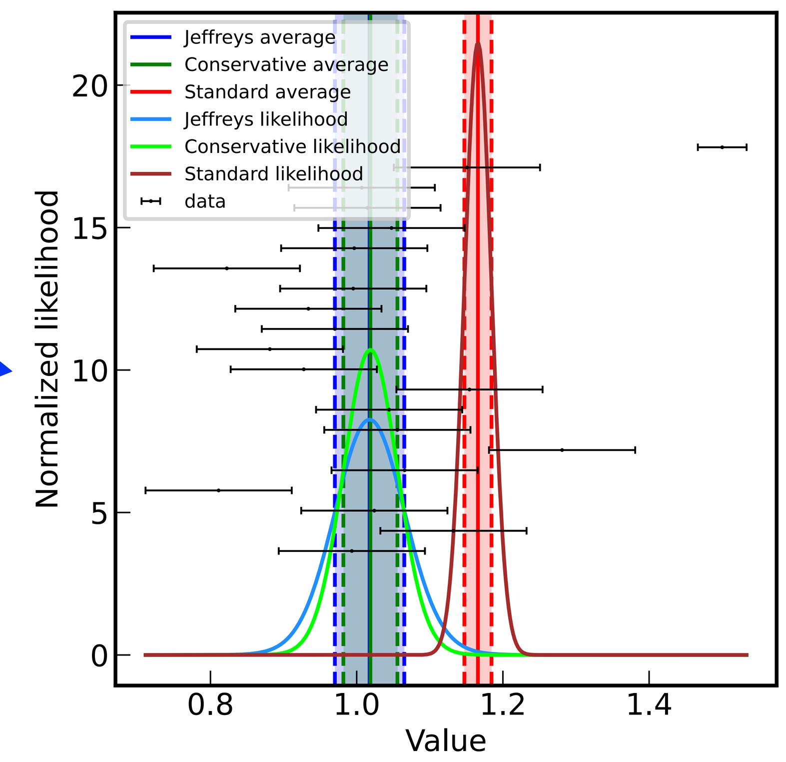
Tests and analyses

Simulations



Jeffreys' prior weighted average:

- Systematically larger uncertainty
- Reproducing the data spread
- Much more robust to outliers

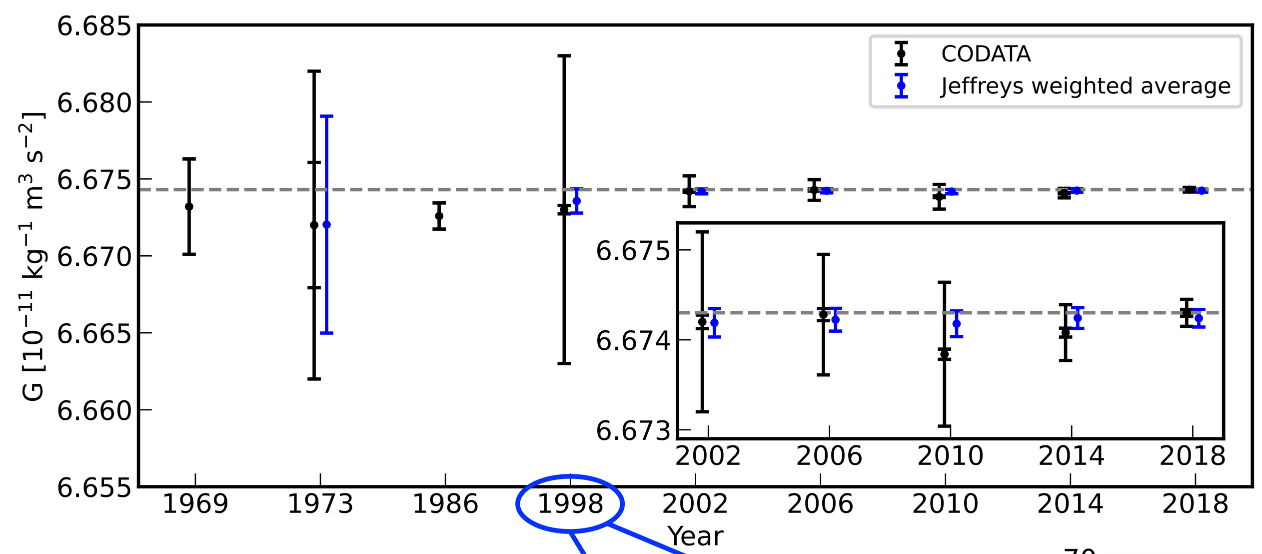


The outlier case

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]
Jeffreys' prior	1.018	0.047
Conservative	1.019	0.037

[a] Birge ratio equal to 2.9

The Newtonian constant of gravity from CODATA

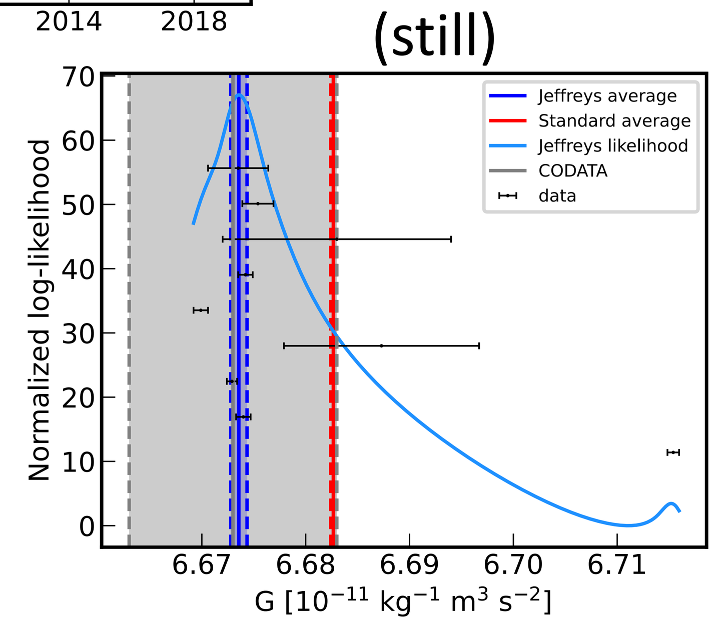


- Always consistent with the most recent value
- Good built-in estimation of the uncertainty
- Robust to outliers (still)

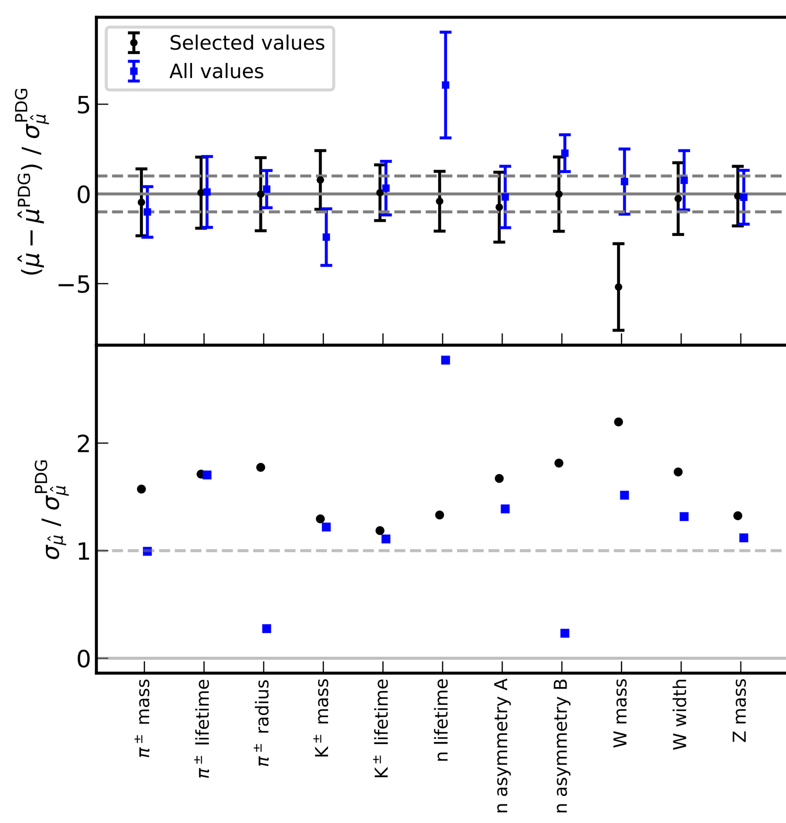
1998: a difficult case

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
Jeffreys' prior	6.6736	0.0008
CODATA 1998 [21]	6.673	0.010[b]

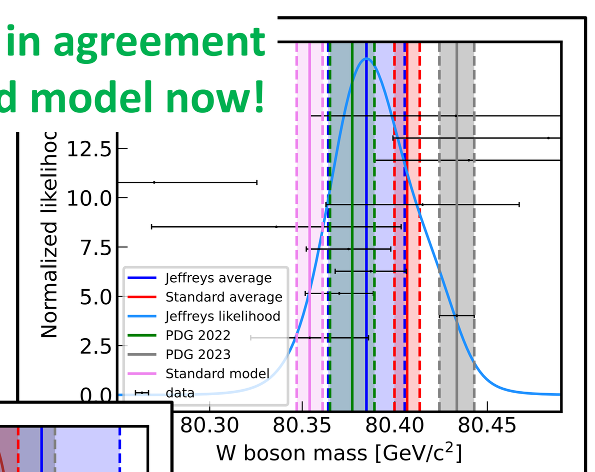
[a] Birge ratio equal to 23.6
[b] Corresponding scale factor equal to 37

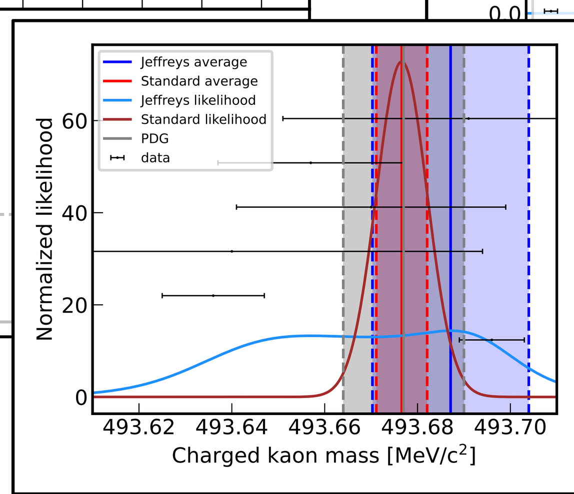


Particle properties



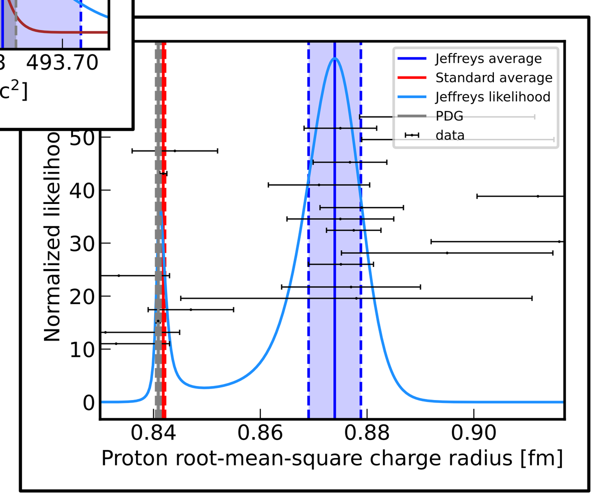
W boson mass in agreement with standard model now!





Weighted averages not recommended here

- PDG recommended values well reproduced for (almost) all selected cases
- Critical cases easily detectable



Proton puzzle is still a puzzle!!

The Python library

- Easily installable via pip: `pip install bayesian_average`
- Freely available in GitHub: https://github.com/martinit18/bayesian_average
- Different average type proposed
- Plot representation of the final probability distribution with the input data

[1] R.T. Birge, Phys. Rev. **40**, 207-227 (1932)
[2] I. Lira, Metrologia 44, 415 (2007)
[3] H. Huang, Metrologia **55**, 106 (2018)
[4] D.S. Sivia and J. Skilling., *Data analysis: a Bayesian tutorial*. Second ed. 2006: Oxford University Press

The manuscript relative to the present work has been submitted to *Metrologia* journal
ArXiv repository: arXiv:2406.08293