Reply to the comment on 'Bayesian estimate of the Newtonian constant of gravitation'

REPLY

Reply to the comment on 'Bayesian estimate of the Newtonian constant of gravitation'

Volker Dose

Max-Planck-Institut für Plasmaphysik, EURATOM Association, Boltzmannstrasse 2,
D-85748 Garching bei München, Germany

Abstract
The author responds to the criticisms in the preceding comment.

I choose to address first Willink's alternative approach to estimating mean and variance from a set of inconsistent data. The assumption that a constant offset σ^2 should be added to the quoted variances of the individual measurements can be traced back to the assumption that all data d_i suffer from a common unrecognized systematic offset $\Delta d_i = \sigma$. The rigorous Bayesian solution to this problem has been presented by Fröhner [1] with application to the robust estimate of the fission cross section of the reaction ^{239}Pu(n, f) at 14.7 MeV neutron energy. The result is formally identical to that given by Willink though he starts from a different concept and his derivation is ad hoc.

My concern is about the use of this result. Willink uses subsets of the available data to obtain an estimate of G which means censoring of the data according to their influence on the estimate of the mean. I oppose vigorously against such a procedure, which brings physical data analysis down to a beauty contest. In fact, the initial motivation for my work [2] derives from a similar discussion of the evaluation of G in the CODATA report [3]. In addition to deleting selected values from the analysis, the authors also consider the possibility of multiplying all quoted uncertainties by a common factor α such that the Birge ratio becomes equal to one. A Bayesian analysis resting on the same assumption introduces this factor α into the calculation and marginalizes it subsequently using an appropriate distribution. Since α is a scale variable the transformation invariance requirement results in Jeffreys' prior $p(\alpha) = 1/\alpha$. This is criticized by Willink. The nature of this prior is more obvious if we look at $p(\ln \alpha)$ which turns out to be flat in $-\infty < \alpha < \infty$. The argument that $p(\alpha)$ puts infinite weight on α near zero is therefore misleading. Moreover, the

likelihood kills the $\alpha \to 0$ behaviour exponentially, while the prior enforces the decay of the likelihood for $\alpha \to \infty$ by a factor of $1/\alpha$.

$p(\alpha)$ is not normalizable and is called improper. The use of improper priors in a Bayesian analysis is quite acceptable if they are considered as the limit of a sequence of proper priors and if the transition to the limit does not affect the posterior distribution. Both requirements have been considered in [2] and shown to hold.

Let me return to marginalization. This important ingredient of Bayesian probability theory, which has no counterpart in the traditional frequentist statistics, is quite often slighted by critics of the Bayesian approach. Simply multiplying all uncertainties by a suitably chosen factor will bring down the Birge ratio to unity but the structure and degeneracy of the data analysis problem leave unaltered the traditional weighted mean albeit with modified uncertainties. It is the marginalization step in the Bayesian analysis which leads to a result with an uncertainty of the mean incorporating both the quoted uncertainties of the input data and the scatter of the data themselves. Does not this meet with common sense? It also removes the problem that mean and variance from a given set of data calculated in the traditional way can be obtained from an infinity of data sets with different scatter and unchanged uncertainties. The Bayesian marginalization removes this degeneracy and the result of the Bayesian treatment is therefore 'characteristic' (is 'unique' a better word?) for a particular set.

Of course, the assumption of a common scaling of uncertainties in a Bayesian treatment is not the only way to remove the degeneracy. It was employed in [2] in order to

stay as close as possible to the considerations in [3]. A much more reasonable assumption for the analysis of the G-data affording this was already indicated in the last paragraph of [2], which does not imply in the least that I 'leave the Bayesian interpretation and return to the classical understanding'.

Willink also criticizes the choice of $p(\ln \alpha) = \text{const}$ on the grounds that it is uninformative. By far the most frequent argument which I have to deal with in talks on Bayesian data analysis is that the choice of the prior distribution anticipates posterior inference. The answer to this objection in a nut shell is that the prior influences the posterior depending on its information content. The use of uninformative priors is therefore a conservative approach in case the formulation of an informative prior is beyond the field of competence of the analyst. I dare not imagine the number of comments if I had chosen a structured prior instead. The same applies for the flat prior in model comparison of the three different likelihoods. By the way, the equal prior weights used at this stage result from employing the principle of maximum entropy, subject to the side condition of normalization only.

Willink also contests the use of the likelihoods employed in [2]. Though the introduction of a scaling variable α and its subsequent marginalization has cured the deficiency of the traditional uncertainty estimate, the mean has remained unchanged for the Gaussian likelihood. However, if the data are inconsistent with the assumption that they obey a Gaussian likelihood, then this assumption must be abandoned. There is no unique method to do so. One possibility is to assume a distribution of the uncertainties instead of regarding them as being the exact true ones [4]. Another possibility is to assume that the data stem with a certain probability β from a Gaussian distribution with quoted uncertainties σ_i and with a probability $(1 - \beta)$ from a Gaussian with uncertainty $\alpha \cdot \sigma_i$ [5]. This leads to a mixture model likelihood with at least two hyper parameters which must be marginalized. The mixture model was also applied in a robustness study by Sivia [6]. My rationale for choosing the Laplace distribution in [2] was the widespread experience that the $L1$ norm is much less sensitive to outliers than the $L2$ norm. But is it preferable to the analysis of the G-data? Bayesian probability theory offers a tool to access the quality of a model compared to at least one alternative quantitatively. This is the evidence of a data set. It is the marginal likelihood of the data given their uncertainties and the model. From table 1 in [2] it is seen that this measure is in favour of the Laplace distribution compared to the Gaussian by a factor of 209 for a very discordant data set. The $\cosh^{-1}$ likelihood has Gaussian character for small arguments and the heavier tails of the Laplace distribution for outliers. It resembles in a way the assumptions of mixture modelling without introducing new parameters. Willink raises the question of why not using instead of the $L1$ norm an arbitrary power γ of it (including $\gamma = 2$ for the $L2$ norm). The figure of merit of any such choice compared to any other would of course again be the evidence. A truly Bayesian treatment of this suggestion would include a final marginalization over γ. This might be a problem worth tackling.

I need finally to comment on a misinterpretation of figure 2 in [2] which displays the results of a sensitivity analysis of the three likelihood functions employed. Willink notices correctly that when the fictitious measurement has a value near 'the middle' of the data the sensitivity of the estimates is greater than with the Gaussian likelihood. The explanation is that the fictitious data become a 'good' data point in this region and lower consequently the variance of the estimated mean. In the Gaussian case, the influence of the fictitious measurement on the estimate of the mean is independent of the location of the fictitious data point. This is a manifestation of robustness and the conclusion of Willink on his correct observation is wrong.

References

[1] Froehner F H 2000 *Evaluation and Analysis of Nuclear Resonance Data* (Paris: OECD Nuclear Energy Agency)
[2] Dose V 2007 *Meas. Sci. Technol.* **18** 176
[3] Mohr P J and Taylor B N 2005 *Rev. Mod. Phys.* **77** 1
[4] Dose V, Preuss R and Roth J 2001 *J. Nucl. Mater.* **288** 153
[5] Press W H 1997 Understanding data better with Bayesian and global statistical methods *Unsolved Problems in Astrophysics* ed J N Bahcall and J P Ostriker (Princeton: Princeton University Press)
[6] Sivia D S 1996 *Data Analysis: A Bayesian Tutorial* (Oxford: Clarendon)