Evaluation of Uncertainty in the Adjustment of Fundamental Constants

Olha Bodnar[1], Clemens Elster[1], Joachim Fischer[1], Antonio Possolo[2], Blaza Toman[2]

[1] *Physikalisch-Technische Bundesanstalt, Abbestrasse 2-12, 10587 Berlin, Germany*

[2] *National Institute of Standards and Technology, US Department of Commerce, Gaithersburg, MD, USA*

Motivation

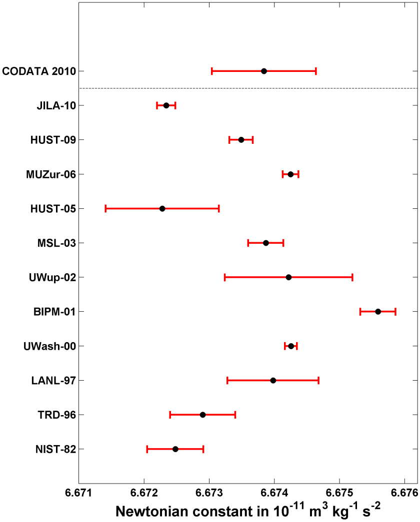
- Task: combine information from multiple studies on same quantity
- Applications
  - Interlaboratory studies and key comparisons
  - Adjustment of fundamental constants
  - Meta-analysis of clinical studies
- Challenges:
  - inconsistent data
  - presence of correlations, ...

Motivation

- Task: combine information from multiple studies on same quantity
- Applications
  - Interlaboratory studies and key comparisons
  - Adjustment of fundamental constants
  - Meta-analysis of clinical studies
- Challenges:
  - inconsistent data
  - presence of correlations, ...

> *Combining information from disparate sources is a fundamental activity in both scientific research and policy decision making*
>
> *The process of learning is one of combining information: we are constantly called upon to update our beliefs in the light of new evidence, which may come in various forms*
>
> *How does this process work in practice?*
>
> — National Research Council (1992)
> *Combining information:*
> *Statistical issues and opportunities for research*

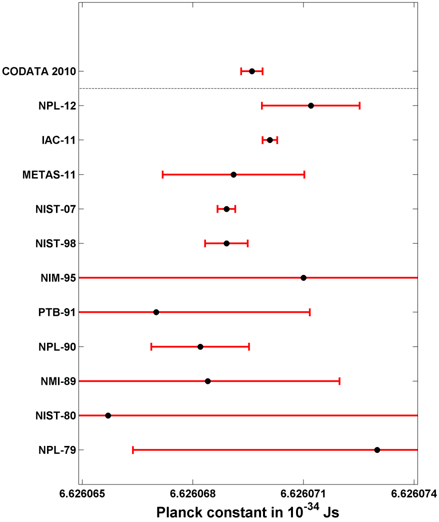
Newtonian Constant of Gravitation G



- Inconsistency of data illustrates current metrological challenges
- Ideally, resolve discrepancies by identifying metrological reasons
- Need for consensus value: statistical (adjustment) problem

P. J. Mohr, B. N. Taylor and D. B. Newell (2012) CODATA Recommended Values of the Fundamental Physical Constants: 2010 *Reviews of Modern Physics* **84**, 1527-1605.

Planck Constant h



- Inconsistency of data illustrates current metrological challenges
- Ideally, resolve discrepancies by identifying metrological reasons
- Need for consensus value: statistical (adjustment) problem

P. J. Mohr, B. N. Taylor and D. B. Newell (2012) CODATA Recommended Values of the Fundamental Physical Constants: 2010 *Reviews of Modern Physics* **84**, 1527-1605.

Problem

- Values measured by different laboratories are often more dispersed than should be expected in light of their associated uncertainties

Problem

- Values measured by different laboratories are often more dispersed than should be expected in light of their associated uncertainties
- Over-dispersion suggests that there may be uncertain components that are yet unrecognized in the corresponding, laboratory-specific uncertainty budgets

Dark uncertainty — Thompson & Ellison (2011)

M. Thompson and S. L. R. Ellison (2011). Dark uncertainty. *Accreditation and Quality Assurance* **16**, 483-487.

Solutions

- Downweigh or discard measurement results that are markedly discordant with bulk of the others

M. G. Cox (2002) The evaluation of key comparison data *Metrologia* **39**, 589-595.

M. G. Cox (2007) The evaluation of key comparison data: determining the largest consistent subset *Metrologia* **44**, 187-200.

G. Ratel (2006) Median and weighted median as estimators for the key comparison reference value (KCRV) *Metrologia* **43**, S244-S248.

A. L. Rukhin and A. Possolo (2011) Laplace random effects models for interlaboratory studies *Computational Statistics and Data Analysis* **55**, 1815-1827.

Solutions

- Downweigh or discard measurement results that are markedly discordant with bulk of the others
- Expand stated uncertainties by common multiplicative factor sufficiently to achieve mutual consistency (Birge Ratio)

P. J. Mohr, B. N. Taylor and D. B. Newell (2012) CODATA Recommended Values of the Fundamental Physical Constants: 2010 *Reviews of Modern Physics* **84**, 1527-1605.

Solutions

- Downweigh or discard measurement results that are markedly discordant with bulk of the others
- Expand stated uncertainties by common multiplicative factor sufficiently to achieve mutual consistency (Birge Ratio)
- Explicitly acknowledge presence of yet unidentified, laboratory-specific uncertainty components, and take them into account when estimating measurand and evaluating associated uncertainty

B. Toman and A. Possolo (2009) Laboratory effects models for interlaboratory comparisons *Accreditation and Quality Assurance* **14**, 553-563.

Solutions

- Downweigh or discard measurement results that are markedly discordant with bulk of the others
- Expand stated uncertainties by common multiplicative factor sufficiently to achieve mutual consistency (Birge Ratio)
- Explicitly acknowledge presence of yet unidentified, laboratory-specific uncertainty components, and take them into account when estimating measurand and evaluating associated uncertainty
- Bayesian model averaging has been suggested for the estimation of particular laboratory effects models

C. Elster and B. Toman (2010) Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using Bayesian model averaging *Metrologia* **47**, 113-119.

Conventional Method

Measurement results from n laboratories: $(x_1, u_1), \ldots, (x_n, u_n)$

- Weighted Mean

$$\boxed{\widehat{\mu}_{\mathrm{W}} = \frac{\sum_{i=1}^{n} \frac{x_i}{u_i^2}}{\sum_{i=1}^{n} \frac{1}{u_i^2}}, \quad u(\widehat{\mu}_{\mathrm{W}}) = \left(\sum_{i=1}^{n} \frac{1}{u_i^2}\right)^{-\frac{1}{2}}}$$

Conventional Method

Measurement results from n laboratories: $(x_1, u_1), \ldots, (x_n, u_n)$

- Weighted Mean

$$\boxed{\widehat{\mu}_{\mathrm{W}} = \frac{\sum_{i=1}^{n} \frac{x_i}{u_i^2}}{\sum_{i=1}^{n} \frac{1}{u_i^2}}, \quad u(\widehat{\mu}_{\mathrm{W}}) = \left(\sum_{i=1}^{n} \frac{1}{u_i^2}\right)^{-\frac{1}{2}}}$$

- Birge Ratio $R_{\mathrm{B}} = \sqrt{\frac{\chi^2_{\mathrm{obs}}}{n-1}}$, $\chi^2_{\mathrm{obs}} = \sum_{i=1}^{n} \frac{(x_i - \widehat{\mu}_{\mathrm{W}})^2}{u_i^2}$

$$\boxed{\widehat{\mu}_{\mathrm{B}} = \widehat{\mu}_{\mathrm{W}}, \quad u(\widehat{\mu}_{\mathrm{B}}) = R_{\mathrm{B}}\, u(\widehat{\mu}_{\mathrm{W}})}$$

  - Fails to recognize uncertainty associated with R_{B}
  - Assumes that same multiplicative inflation factor should be applied to all of $u_1, \ldots, u_n$

Statistical Models

Location-Scale Model (LSM)

$$X_i = \mu + \sigma_{\mathrm{B}} \epsilon_i, \quad \text{with} \quad \epsilon_i \sim N(0, u_i^2) \quad \text{for} \quad i = 1, \ldots, n$$

Statistical Models

Location-Scale Model (LSM)

$$X_i = \mu + \sigma_{\mathrm{B}}\epsilon_i, \quad \text{with} \quad \epsilon_i \sim N(0, u_i^2) \quad \text{for} \quad i = 1, \ldots, n$$

Random Effects Model (REM)

$$X_i = \mu + \lambda_i + \epsilon_i, \quad \text{with} \quad \lambda_i \sim N(0, \sigma_\lambda^2), \quad \epsilon_i \sim N(0, u_i^2) \quad \text{for} \quad i = 1, \ldots, n$$

- $\lambda_1, \ldots \lambda_n$ denote random effects independent of $\epsilon_1, \ldots, \epsilon_n$
- There may be correlations between $\lambda_1, \ldots, \lambda_n$ or between $\epsilon_1, \ldots, \epsilon_n$
- σ_{B} and σ_λ determine *Dark Uncertainty*

Statistical Models

Location-Scale Model (LSM)

$$X_i = \mu + \sigma_{\mathrm{B}}\epsilon_i, \quad \text{with} \quad \epsilon_i \sim N(0, u_i^2) \quad \text{for} \quad i = 1, \ldots, n$$

Random Effects Model (REM)

$$X_i = \mu + \lambda_i + \epsilon_i, \quad \text{with} \quad \lambda_i \sim N(0, \sigma_\lambda^2), \quad \epsilon_i \sim N(0, u_i^2) \quad \text{for} \quad i = 1, \ldots, n$$

- $\lambda_1, \ldots \lambda_n$ denote random effects independent of $\epsilon_1, \ldots, \epsilon_n$
- There may be correlations between $\lambda_1, \ldots, \lambda_n$ or between $\epsilon_1, \ldots, \epsilon_n$
- σ_{B} and σ_λ determine *Dark Uncertainty*

LSM amplifies lab-specific uncertainties by common factor $\sigma_{\mathrm{B}} > 1$

REM introduces additive laboratory effects, with common variance σ_λ^2

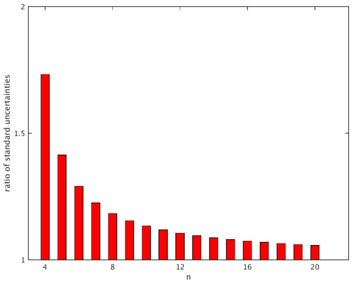
Model Fitting

Both LSM and REM may be fitted to data using anyone of several different methods for estimating μ and σ_{B} (for LSM), and μ and σ_λ (for REM)

Location-scale model

- Weighted mean for μ and method of moments for σ_{B}
  - R. T. Birge 1932 The calculation of errors by the method of the least squares. *Physical Reviev* **40**, 207-227.
- Bayesian estimates for μ and σ_{B}
  - O. Bodnar and C. Elster (2014) On the adjustment of inconsistent data using the Birge ratio. *Metrologia* **51**, 516-521.
  - B. Toman, J. Fischer and C. Elster (2012) Alternative analyses of measurements of the Planck constant. *Metrologia* **49**, 567-571.

Bayesian Improvement of Birge Ratio Procedure



- Ratio of standard uncertainties obtained by the modified Birge adjustment and the Birge adjustment as a function of the number n of adjusted values
- O. Bodnar and C. Elster (2014) On the adjustment of inconsistent data using the Birge ratio. *Metrologia* **51**, 516-521.

  B. Toman, J. Fischer and C. Elster (2012) Alternative analyses of measurements of the Planck constant. *Metrologia* **49**, 567-571.

Random effects model

- First estimate σ_λ, then estimate μ by weighted mean with weights proportional to $\{1/(\widehat{\sigma}_\lambda^2 + u_i^2)\}$

Viechtbauer (2010) Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36 (3)**, 1-48

A. L. Rukhin (2013) Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society Series B* **75**, 451-469.

- (Restricted) Maximum Likelihood

A. L. Rukhin and M. G. Vangel (1998) Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association* **93**, 303-308.

S. R Searle, G. Casella and C. E. McCulloch (2006) *Variance Components*. John Wiley & Sons, New Jersey.

B. Toman and A. Possolo (2009) Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance* **14**, 553-563.

- Bayesian methods

A. Gelman (2006) Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis* **1**, 515-534.

O. Bodnar, A. Link and C. Elster (2015) Objective Bayesian inference for a generalized marginal random effects model. To appear in *Bayesian Analysis*.

Model Fitting

BAYESIAN REFERENCE ANALYSIS for RANDOM EFFECTS MODEL

- The reference prior is most uninformative in a certain sense

J. Berger and J. M. Bernardo (1992a) Ordered group reference priors with application to the multinomial problem. *Biometrika* **79**, 25-37.

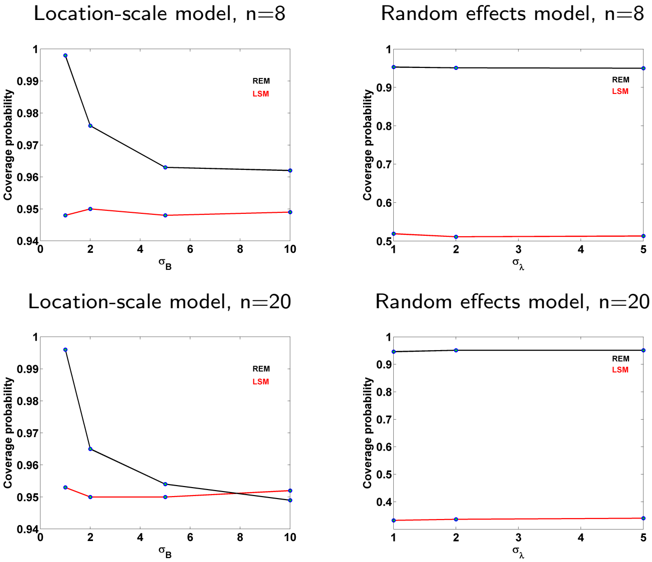
O. Bodnar and C. Elster (2014) Analytical derivation of the reference prior by sequential maximization of Shannon's mutual information in the multi-group parameter case. *Journal of Statistical Planning and Inference* **147**, 106-116.

Model Fitting

BAYESIAN REFERENCE ANALYSIS for RANDOM EFFECTS MODEL

- The reference prior is most uninformative in a certain sense
  - J. Berger and J. M. Bernardo (1992a) Ordered group reference priors with application to the multinomial problem. *Biometrika* **79**, 25-37.
  - O. Bodnar and C. Elster (2014) Analytical derivation of the reference prior by sequential maximization of Shannon's mutual information in the multi-group parameter case. *Journal of Statistical Planning and Inference* **147**, 106-116.
- Analytic expression for the reference prior available
- Calculation of posterior distribution by simple 1-d quadrature, no MCMC required
  - O. Bodnar, A. Link and C. Elster (2015) Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Analysis, online available*.

Robustness, True Model *vs.* Fitted Model



Model Robustness

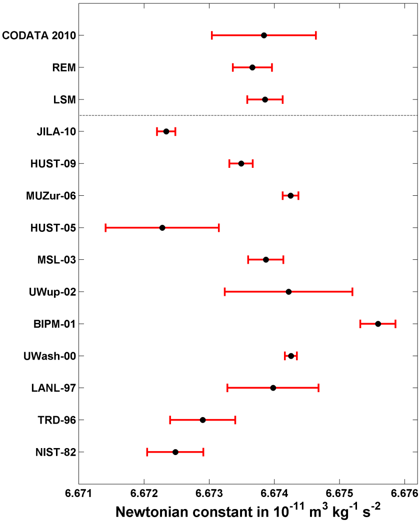
- Accuracy of coverage probability largely maintained when REM is fitted to data consistent with LSM, but seriously degraded when LSM is fitted to data consistent with REM

Model Robustness

- Accuracy of coverage probability largely maintained when REM is fitted to data consistent with LSM, but seriously degraded when LSM is fitted to data consistent with REM

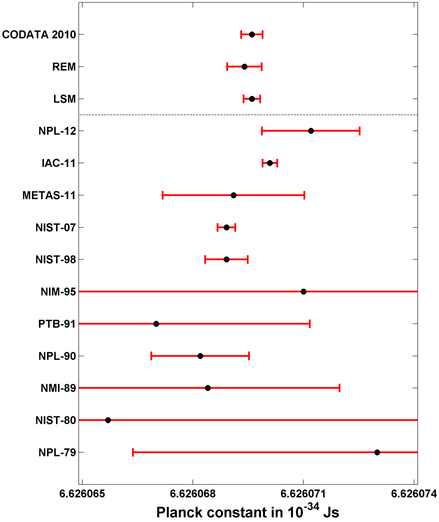
- REM achieves greater accuracy of coverage probability than LSM when the $\{\epsilon_i\}$ or the $\{\lambda_i\}$ are correlated, and when both the $\{\epsilon_i\}$ and the $\{\lambda_i\}$ have rescaled Student's t_5 distributions

Newtonian Constant of Gravitation G



- LSM — Conventional estimate and uncertainty evaluation based on Birge Ratio
- REM — Bayesian estimate and uncertainty evaluation based on reference prior

Planck Constant h



- LSM — Conventional estimate and uncertainty evaluation based on Birge Ratio
- REM — Bayesian estimate and uncertainty evaluation based on reference prior

Summary & Conclusions (1/2)

- Unrecognized uncertainty components often present in interlaboratory or multiple method studies (*Dark Uncertainty*)

Summary & Conclusions (1/2)

- Unrecognized uncertainty components often present in interlaboratory or multiple method studies (*Dark Uncertainty*)

- The standard uncertainty, based on the Birge Ratio, associated with the conventional weighted mean estimate of the consensus value, tends to be too small

Summary & Conclusions (1/2)

- Unrecognized uncertainty components often present in interlaboratory or multiple method studies (*Dark Uncertainty*)
- The standard uncertainty, based on the Birge Ratio, associated with the conventional weighted mean estimate of the consensus value, tends to be too small
- Bayesian reference analysis leads to an improvement of the Birge Ratio Method

Summary & Conclusions (1/2)

- Unrecognized uncertainty components often present in interlaboratory or multiple method studies (*Dark Uncertainty*)
- The standard uncertainty, based on the Birge Ratio, associated with the conventional weighted mean estimate of the consensus value, tends to be too small
- Bayesian reference analysis leads to an improvement of the Birge Ratio Method
- Random effects model (REM) is more robust than location-scale model (LSM)

Summary & Conclusions (1/2)

- Unrecognized uncertainty components often present in interlaboratory or multiple method studies (*Dark Uncertainty*)
- The standard uncertainty, based on the Birge Ratio, associated with the conventional weighted mean estimate of the consensus value, tends to be too small
- Bayesian reference analysis leads to an improvement of the Birge Ratio Method
- Random effects model (REM) is more robust than location-scale model (LSM)
- Fitting procedure influential: Bayesian reference analysis appears to work well

Summary & Conclusions (1/2)

- Unrecognized uncertainty components often present in interlaboratory or multiple method studies (*Dark Uncertainty*)
- The standard uncertainty, based on the Birge Ratio, associated with the conventional weighted mean estimate of the consensus value, tends to be too small
- Bayesian reference analysis leads to an improvement of the Birge Ratio Method
- Random effects model (REM) is more robust than location-scale model (LSM)
- Fitting procedure influential: Bayesian reference analysis appears to work well
- Bayesian reference analysis in conjunction with REM assigns uncertainties to consensus values for G and h that are substantially different from CODATA's

Summary & Conclusions (2/2)

- We recommend the use of *random effects statistical models* to derive consensus values from measurement results obtained by different methods or laboratories

Summary & Conclusions (2/2)

- We recommend the use of *random effects statistical models* to derive consensus values from measurement results obtained by different methods or laboratories

- Ideally, application of statistical techniques and interpretation of corresponding results should be done in *collaboration between metrologists and statisticians*