

Using Bayes methods and mixture models in inter-laboratory studies with outliers

Garritt L. Page · Stephen B. Vardeman

Received: 11 March 2009 / Accepted: 19 February 2010 / Published online: 10 March 2010
© Springer-Verlag 2010

Abstract Inter-laboratory studies (especially so-called key comparisons) are conducted to evaluate both national and international equivalence of measurement. In these studies, a reference value of some measurand (the quantity intended to be measured) is developed and results for all laboratories are compared to this single value. How to determine the reference value is not completely obvious if there are observations and/or laboratories that could be considered outliers. Since ignoring results from one or more participating laboratories is untenable in practical terms, developing methods that are robust to the possibility that a small fraction of the laboratories produces observations unlike those from the others is critical. This paper outlines two Bayesian methods of analyzing inter-laboratory data that have been proposed in the literature and suggests three modifications of one that are more robust to outliers. A simulation study is conducted to compare the five methods.

Keywords Hierarchical models · Inter-laboratory studies · Mixtures · Outlying laboratories

Introduction

Inter-laboratory studies (especially so-called key comparisons) are conducted to evaluate both national and

international equivalence of measurement. In these studies, a reference value of some measurand (the quantity intended to be measured) is developed and results for all laboratories are compared to this single value. How to determine the reference value is not completely obvious if there are observations and/or laboratories that could be considered outliers. Since ignoring results from one or more participating laboratories is untenable in practical terms, developing methods that are robust to the possibility that a small fraction of the laboratories produces observations unlike those from the others is critical. This paper outlines two Bayesian methods of analyzing inter-laboratory data that have been proposed in the literature and suggests three modifications of one that are more robust to outliers. A simulation study is conducted to compare the five methods.

Models and Bayes procedures

This section describes two models that have been proposed to analyze inter-laboratory data and three modifications of one.

Gaussian lab model

Toman [1] proposes the following Bayesian hierarchical model for inter-laboratory data that we refer to as the Gaussian lab model (BLM) (here the “B” is used for bell-curve). Let Y_{ij} denote measurement j taken by laboratory i , with $i = 1, \dots, L$ and $j = 1, \dots, m_i$. Before displaying the BLM, we introduce some notation that will be used throughout. “ \sim ” denotes “distributed as”, “ $\overset{\text{ind}}{\sim}$ ” denotes “distributed independently as”, and “ $\overset{\text{iid}}{\sim}$ ” denotes “distributed independently and identically as”. Now, for the BLM, we suppose that

G. L. Page (✉)
Department of Statistical Science, Duke University,
122A Old Chemistry, Durham, NC 27708, USA
e-mail: page@stat.duke.edu; gpage299@iastate.edu

S. B. Vardeman
Department of Statistics, Iowa State University,
2212 Snedecor, Ames, IA 50011, USA
e-mail: vardeman@iastate.edu

$$Y_{ij} \stackrel{\text{ind}}{\sim} N(\delta_i, \sigma_i^2), \quad (1)$$

$$\delta_i \stackrel{\text{ind}}{\sim} N(\mu, \tau_i^2), \quad (2)$$

$$\mu \sim N(m, v^2), \quad (3)$$

where $N(m, s^2)$ denotes a normal distribution with mean m and variance s^2 , μ is the measurand, δ_i is the mean for the i th laboratory, σ_i^2 is the corresponding within laboratory variance, and τ_i^2 accounts for variability due to “Type B” uncertainty. This type of uncertainty, as described in the *guide to the expression of uncertainty in measurement* (GUM) [2], is evaluated by means other than the statistical analysis of series of observations. Equation 1 is the data model (sometimes referred to as the likelihood), Eq. 2 is the laboratory means model, and Eq. 3 is the prior distribution of the measurand. The parameters of the prior distribution denoted by the latin letters m and v^2 need to be specified by an analyst using this model. (The practice of using latin letters to represent parameters whose values are specified by an analyst continues throughout this article.) If there is knowledge of the most probable location and/or uncertainty about the measurand (μ), then this knowledge is incorporated in the model through m and v^2 . Toman [1] points out that because τ_i^2 represents the variability due to systematic effects, the Y_{ij} 's are not informative in evaluating it. Uncertainty due to systematic effects influences all observations in the experiment. Thus, each participating laboratory calculates and provides a value for τ_i^2 following the protocol outlined in the GUM. Finally, conditional independence within and between laboratories is assumed at both levels of the hierarchy.

Posterior distribution of μ under the BLM

Let s_i^2 be the laboratory i sample variance (an estimate of σ_i^2). Under the conditions that $v^2 \rightarrow \infty$ and σ_i^2 is much smaller than τ_i^2 (which is often the case in inter-laboratory studies) Toman [1] argues that for fixed $\tau_1^2, \tau_2^2, \dots, \tau_L^2$ the posterior distribution of μ given the Y_{ij} 's is approximately Gaussian with mean

$$\mu_p = \frac{\sum_1^L \bar{y}_i (\tau_i^2 + s_i^2/m_i)^{-1}}{\sum_1^L (\tau_i^2 + s_i^2/m_i)^{-1}} \quad (4)$$

and standard deviation

$$s_p = \frac{1}{\sqrt{\sum_1^L (\tau_i^2 + s_i^2/m_i)^{-1}}}, \quad (5)$$

where \bar{y}_i is the i th laboratories sample mean. Therefore, μ_p can be used as an estimate for μ and hence, the reference value. The above model and approximate posterior are appealing because μ_p is the reference value identified in the GUM. The condition that $v^2 \rightarrow \infty$ is crucial for obtaining the limiting Gaussian posterior distribution of μ . If we

instead used a Jeffreys prior [3] for μ , we will not have the approximate posterior distribution result. (The Jeffreys improper prior for μ is $p(\mu) \propto 1$ which is the unnormalized uniform distribution on the real line and is essentially what one would get if $v^2 = \infty$.)

t -lab model

It is known that parameter estimates in Gaussian models are typically highly influenced by outliers. This suggests that using a Gaussian distribution to model the laboratory means when one or more of the laboratories are potentially unlike the majority could prove to be problematic in estimating μ . Gelman et al. [4], among others, suggest using a t -distribution as a robust alternative to the normal. The t distributions can accommodate occasional extreme observations since they have heavier tails than normal distributions. Section 7 of Possolo et al. [5] provides an example of using a t -distribution to model laboratory means in an inter-laboratory study setting. They propose the following model which will here be referred to as the t -lab model (TLM).

$$Y_{ij} \stackrel{\text{ind}}{\sim} N(\delta_i, \sigma_i^2),$$

$$\delta_i \stackrel{\text{iid}}{\sim} t_v(\mu, \tau^2),$$

$$\sigma_i^2 \stackrel{\text{iid}}{\sim} \text{IG}(a_\sigma, b_\sigma),$$

$$\tau^2 \sim \text{IG}(a_\tau, b_\tau),$$

$$v \sim \text{DU}(a_v, b_v),$$

$$\mu \sim N(m, v^2),$$

where $\text{IG}(a, b)$ denotes an inverse gamma distribution with mean $1/(b(a - 1))$ for $a > 1$ and $b > 0$, $\text{DU}(a, b)$ denotes the discrete uniform distribution on $\{a, a + 1, \dots, b\}$ and $t_v(\mu, \tau^2)$ is a scaled (τ^2) and shifted (μ) t -distribution with v degrees of freedom. We are now modeling the laboratory means with a shifted and scaled t -distribution with v degrees of freedom.

We note here that unlike what is the case in the BLM, there is a single measure of between - laboratory variation (τ^2) that is not treated as “known,” fixed, or user-supplied but rather as an unobserved variable (that thus, has a posterior distribution). Therefore, all uncertainties in the TLM (and in subsequent models) are of type A and are estimated using statistical techniques.

For the TLM model, there is no simple description of the posterior distribution of μ . Therefore, the distribution is approximated via simulation, using Markov Chain Monte Carlo (MCMC) in some fashion. (Details of the MCMC algorithm used here will be given in the next section.) Then, using the posterior distribution, we can get an estimate of μ to use as the reference value. This estimate is usually obtained through an ergodic average from the Monte Carlo samples of μ .

Mixture *t*-lab models

Two component mixture *t*-lab model

Preliminary computations comparing interval estimates for the parameter μ common to the BLM and TLM models (based on μ_p and s_p in the first case and the posterior distribution of μ in the second case) suggested that neither method is completely satisfactory in the presence of outlying laboratories. Intervals based on μ_p and s_p fail to cover with anything close to nominal probabilities. Bayes intervals based on the TLM model can be extremely wide/uninformative. This motivates the search for a methodology that can handle outlying laboratories without sacrificing either coverage probability or interval length. To this end, consider the following modification of the TLM which we will call the two component mixture *t*-lab model (2CMTLM). For this model, we suppose

$$\begin{aligned}
 Y_{ij} &\overset{\text{iid}}{\sim} N(\delta_i, \sigma_i^2), \\
 \delta_i &\overset{\text{iid}}{\sim} (1 - \pi)t_v(\mu, \tau^2) + \pi t_v(\mu + \xi, \tau^2), \\
 \sigma_i^2 &\overset{\text{iid}}{\sim} \text{IG}(a_\sigma, b_\sigma), \\
 \pi &\sim \text{Beta}(a_\pi, b_\pi), \\
 \xi &\sim N(m_\xi, v_\xi^2), \\
 v &\sim \text{UN}(a_v, b_v), \\
 \tau^2 &\sim \text{IG}(a_\tau, b_\tau), \\
 \mu &\sim N(m, v^2).
 \end{aligned}$$

Laboratory means are modeled with a mixture of *t*-distributions where the laboratory means come from an outlier (typically “rare”) group with probability π . In this way, the mean of a laboratory from the outlier group is not straightway used to estimate a reference value, rather the mean is adjusted by ξ before being used to characterize μ . Chapter 7.2 of Frühwirth-Schnatter [6] suggests using a finite mixture of normals to model outlying observations. Here, we are using a mixture of *t*-distributions to model means of outlying labs (which are in the second level of a hierarchy). Notice that v is assumed to come from a continuous uniform distribution (UN) instead of a discrete uniform. This change from the TLM model is made for computational reasons and does not much alter inference for μ .

Three component mixture *t*-lab model

Often in large inter-laboratory studies, there is more than one laboratory that might be considered an outlier. Moreover, there could potentially be two types of outlying laboratories with one group being further from the majority of the laboratories (or on the other side of the majority)

than the other. With this in mind, we formulate a model that allows for three types of laboratories: The majority and two potentially distinct smaller groups that are different from the majority. Consider the following modification of the 2CMTLM which we will call the three component mixture *t*-lab model (3CMTLM). We suppose:

$$\begin{aligned}
 Y_{ij} &\overset{\text{iid}}{\sim} N(\delta_i, \sigma_i^2), \\
 \delta_i &\overset{\text{iid}}{\sim} \pi_1 t_v(\mu, \tau^2) + \pi_2 t_v(\mu + \xi, \tau^2) + \pi_3 t_v(\mu + \zeta, \tau^2), \\
 \sigma_i^2 &\overset{\text{iid}}{\sim} \text{IG}(a_\sigma, b_\sigma), \\
 \pi &\sim \text{Dir}(a_\pi, b_\pi, c_\pi), \\
 \zeta &\sim N(m_\zeta, v_\zeta^2), \\
 v &\sim \text{UN}(a_v, b_v), \\
 \xi &\sim N(m_\xi, v_\xi^2), \\
 \tau^2 &\sim \text{IG}(a_\tau, b_\tau), \\
 \mu &\sim N(m, v^2).
 \end{aligned}$$

We are now modeling the laboratory means with a three component mixture of *t*-distributions each with v degrees of freedom. The means of laboratories that belong to outlying groups are adjusted by ξ or ζ before being used to characterize the measurand μ . We follow common practice (Frühwirth-Schnatter [6]) and assign a Dirichlet [Dir(a, b, c)] prior distribution to π , which is a vector of the classification probabilities. The rest of the model is similar to the 2CMTLM. Once again, all Latin letters are constants whose values are assigned by the scientist. The posterior distribution of μ is approximated using MCMC.

Label switching in mixture models

A challenge in a Bayesian mixture analysis is the so-called “label switching” problem, caused by the invariance of the mixture distribution to the relabeling of components. That is,

$$\begin{aligned}
 p(\delta|\mu, \xi, \zeta, \pi, \tau^2) &= \prod_{i=1}^L \{ \pi_1 t_v(\delta_i; \mu, \tau^2) + \pi_2 t_v(\delta_i; \mu + \xi, \tau^2) \\
 &\quad + \pi_3 t_v(\delta_i; \mu + \zeta, \tau^2) \} \tag{6}
 \end{aligned}$$

is invariant to permutations of the 2-dimensional entries of $[(\pi_1, \mu), (\pi_2, \mu + \xi), (\pi_3, \mu + \zeta)]$. So using an ergodic average of the simulated draws from the posterior of μ obtained from a Gibbs sampler would not produce a useful estimate of μ . A few solutions to this problem have been offered in the literature (see Jasra et al. [7] for a review of the work). We seek a method of identifying the location of the “majority” scaled *t* distribution such that at each iteration of the Gibbs sampler, the components of (6) are correctly “relabelled” if necessary. We approach this using a “relabeling algorithm” like that of Stephens [8]. For the

t th iteration of the MCMC algorithm, we would like the simulated draw from the marginal posterior distribution of the measurand to be either $\mu^{(t)}$, $\mu^{(t)} + \zeta^{(t)}$ or $\mu^{(t)} + \zeta^{(t)}$ depending on how the mixture is labeled by the algorithm. To do this, we assess the likelihood that each $\delta_i^{(t)}$ is drawn from t -distributions centered at $\mu^{(t)}$, $\mu^{(t)} + \zeta^{(t)}$ and $\mu^{(t)} + \zeta^{(t)}$. We then choose to “relabel” the components according to which centered t -distribution most likely produces the majority of $\delta_i^{(t)}$ ’s. More precisely, at every t th iteration of the MCMC algorithm (which is described in a subsequent section) we

1. compute $P_g^{(t)} = \frac{\pi_g^{(t)} \phi(\delta_i^{(t)}; \mu_g^{(t)}, \lambda_i^{(t)} \tau^{2(t)})}{\sum_{\ell=1}^3 \pi_\ell^{(t)} \phi(\delta_i^{(t)}; \mu_\ell^{(t)}, \lambda_i^{(t)} \tau^{2(t)})}$ with $g = 1, 2, 3$ and $\mu_1 = \mu$, $\mu_2 = \mu + \zeta$, $\mu_3 = \mu + \zeta$ and where $\phi(\cdot; m, v)$ is the normal density with mean m and variance v ,
2. compute $M^{(t)} = \max_g P_g^{(t)}$ (where \max_g denotes maximum across $g = 1, 2, 3$), and
3. let μ^* denote the “re-labelled” measurand, then set $\mu^{*(t)} = \mu_g^{(t)}$ if $P_g^{(t)} = M^{(t)}$

We then use the posterior distribution of μ^* to make inferences on the measurand.

Partially specified two component mixture model

When using a finite mixture to model outlying laboratories, one must take great care in assuring that label switching is handled properly. An effective methodology that automatically avoids this difficulty would be highly desirable. To this end, we propose to model the lab means with a mixture of a t -distribution and a uniform, where the uniform is completely specified. That is:

$$\begin{aligned}
 Y_{ij} &\overset{\text{iid}}{\sim} N(\delta_i, \sigma_i^2), \\
 \delta_i &\overset{\text{iid}}{\sim} \pi t_v(\mu, \tau_i^2) + (1 - \pi) \text{UN}(a_\delta, b_\delta), \\
 \sigma_i^2 &\overset{\text{iid}}{\sim} \text{IG}(a_\sigma, b_\sigma), \\
 \pi &\sim \text{Beta}(a_\pi, b_\pi), \\
 v &\sim \text{UN}(a_v, b_v), \\
 \tau^2 &\sim \text{IG}(a_\tau, b_\tau), \\
 \mu &\sim N(m, v^2).
 \end{aligned}$$

Note that the uniform component of the mixture is fully specified since we are assigning values to a_δ and b_δ . The rest of the model is like the 2CMTLM. By fully specifying one of the components of the mixture, we avoid having to deal with the label switching problem. Choosing “good” values for a_δ and b_δ must be done with care, as making the interval (a_δ, b_δ) too narrow will result in excluding some Laboratories from the “majority” component that belong there, and making (a_δ, b_δ) too wide will result in including

some Laboratories in the “majority” component that do not belong there.

Hyper prior parameter selection

As our goal is to compare the performance of the BLM and the TLM to that of 2CMTLM, 3CMTLM, and partially specified two component mixture model (PSMTLM) when possible we use prior values like those outlined in Possolo et al. [5]. Having said this, it should be noted that when the number of laboratories is small, the posterior distributions of the τ_i^2 are highly influenced by the priors and in practice these priors should be chosen with great care. We use the same priors for those parameters that are common to all models. Following Possolo et al. [5], we assign $a_v = 2$ and $b_v = 140$. Also, assigning $a_\tau = a_\sigma = 2.0001$ and $b_\tau = b_\sigma = 1.0001$ gives flat inverse gamma distributions that have means of 1 and coefficient of variations of about 100 for $\tau^2, \sigma_1^2, \dots, \sigma_J^2$ (that provide a prior specification like that in Possolo et al. [5]). We assign $m = 0$ and $v^2 = 10^6$ that is a common “non-informative” prior specification for means of normal distributions. Similarly, we use a diffuse normal for ξ and ζ . Finally for the component weights (π) we follow Frühwirth-Schnatter [6] and make the prior probability of being an outlier of any type small (0.1 for 2CMTLM and 0.375 for 3CMTLM). Values for a_π, b_π , and c_π were chosen to produce the prior probability of not being an outlier greater than 0.6. The posterior distribution of μ was not sensitive to values selected for these prior parameters. Table 1 contains a summary of the hyper prior parameters used in what follows.

MCMC algorithm

What follows is a description of a MCMC algorithm corresponding to 3CMTLM. MCMC algorithms for the remaining procedures are similar.

One could use the Metropolis-Hastings(M-H) algorithm (see Metropolis et al. [9] and Hastings [10]) to simulate draws from the joint posterior distribution corresponding to 3CMTLM by using the t -density directly. But because the corresponding Gibbs sampling algorithm (Geman et al. [11]) is often easier to implement, we use the scaled

Table 1 Hyper prior parameter values

Parameter	m	v^2	Parameter	a	b	c
μ	0	10^6	τ^2	2.0001	1.0001	–
ξ	0	10^5	σ_i^2	2.0001	1.0001	–
ζ	0	10^5	v	2	140	–
			$\pi_{2\text{CMTLM}}$	9	1	–
			$\pi_{3\text{CMTLM}}$	10	5	1

mixture of normals representation of the t -distribution. That is, we restate the lab mean portion of the 3CMTLM as

$$\delta_i | \mu, \tau, \lambda_i \stackrel{\text{ind}}{\sim} \pi_1 N(\mu, \lambda_i \tau^2) + \pi_2 N(\mu + \xi, \lambda_i \tau^2) + \pi_3 N(\mu + \zeta, \lambda_i \tau^2),$$

$$\lambda_i | v \stackrel{\text{iid}}{\sim} \text{IG}(v/2, 2/v),$$

$$v | a_v, b_v \sim \text{UN}(a_v, b_v).$$

Now, because

$$t_v(\delta_i; \mu, \tau^2) = \int_0^\infty N(\delta_i; \mu, \lambda_i \tau^2) \text{IG}(\lambda_i; v/2, 2/v) d\lambda_i,$$

inferences on μ are unchanged by introducing the auxiliary variables $\{\lambda_i\}_{i=1}^L$.

The conditional distribution of δ_i is a mixture of t distributions. In order to utilize a Gibbs sampling algorithm, it is common practice to interpret the mixture as a missing data problem (see Gelman et al. [4]) by introducing latent/auxilliary variables

$$\gamma_{ig} = \begin{cases} 1 & \text{if the } i\text{th, lab is a in the } g\text{th component} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, each $\gamma_i = [\gamma_{1i}, \gamma_{2i}, \gamma_{3i}]$ is a vector that consists of a one and two zeros. To complete the model augmentation, we assign a Multinomial-Dirichlet hierarchical structure to the γ_i 's. Using the auxiliary variables just described the 3CMTLM becomes:

$$Y_{ij} | \delta_i, \sigma_i^2 \stackrel{\text{ind}}{\sim} N(\delta_i, \sigma_i^2),$$

$$\delta_i | \mu, \xi, \zeta, \tau^2, \pi, v, \gamma_i \stackrel{\text{ind}}{\sim} [N(\mu, \lambda_i \tau^2)]^{\gamma_{1i}} [N(\mu + \xi, \lambda_i \tau^2)]^{\gamma_{2i}} [N(\mu + \zeta, \lambda_i \tau^2)]^{\gamma_{3i}},$$

$$\sigma_i^2 | a_\sigma, b_\sigma \stackrel{\text{iid}}{\sim} \text{IG}(a_\sigma, b_\sigma),$$

$$\gamma_i | \pi \stackrel{\text{iid}}{\sim} \text{Multinomial}(1; \pi_1, \pi_2, \pi_3) \text{ with } \sum_{g=1}^3 \pi_g = 1,$$

$$\pi | a_\pi, b_\pi, c_\pi \sim \text{Dir}(a_\pi, b_\pi, c_\pi),$$

$$\lambda_i | v \stackrel{\text{iid}}{\sim} \text{IG}(v/2, 2/v),$$

$$v | a_v, b_v \sim \text{UN}(a_v, b_v),$$

$$\xi | m_\xi, v_\xi^2 \sim N(m_\xi, v_\xi^2),$$

$$\zeta | m_\zeta, v_\zeta^2 \sim N(m_\zeta, v_\zeta^2),$$

$$\tau^2 | a_\tau, b_\tau \sim \text{IG}(a_\tau, b_\tau),$$

$$\mu | m_\mu, v_\mu^2 \sim N(m_\mu, v_\mu^2).$$

This more complicated model facilitates the use of a Gibbs sampler. v is the only parameter whose conditional is not of recognizable form, which leads to using an M-H step for updating it. It is straightforward to simulate from $p(\mu, \xi, \zeta, \tau^2, \sigma^2, v, \pi, \delta, \lambda, \{\gamma_i\}_{i=1}^L | y)$ by cycling through the complete posterior conditionals on an individual basis using the Gibbs sampler and a M-H step.

Simulation study

To assess the performance of Bayesian procedures corresponding to the models presented here, we performed a simulation study. The study consisted of generating a data set that is representative of a inter-laboratory study, then from the generated data estimating μ by computing credible intervals from the posterior distributions obtained from the five models outlined previously, then repeating the whole process. The Bayesian procedures were compared in terms of credible interval widths and empirical coverage relative frequency (which here is the fraction of computed credible intervals containing the value that generated lab means out of the total number of computed credible intervals).

It should be noted that we used frequentist metrics to compare the performance of Bayesian procedures. This is done solely for the purpose of comparison, with the hope that we might conclude that the procedures that produce intervals that contain the “truth” more often are more “accurate” than those that do not and procedures that have smaller interval widths (on average) are more “precise” than those that produce larger interval widths.

Generation of data sets

In this study, we generated data under the assumptions of the data model and laboratory means model of the BLM and the TLM. First, we considered the data model combined with the lab means model of the BLM as a data generating mechanism. That is, after specifying values for μ, τ_i^2 , and σ_i^2 we randomly drew δ_i 's from $N(\mu, \tau_i^2)$. Then for each δ_i , m values were randomly drawn from $N(\delta_i, \sigma_i^2)$. This produced a data set with L laboratories and m observations for each laboratory. The same procedure outlined for the BLM was followed using the TLM except v was fixed and δ_i was drawn from $t_v(\mu, \tau^2)$. To include an outlying laboratory in a data set, we randomly selected a δ_i and set its value to $\mu \pm k\tau$ for some constant k . To include an “extreme” outlier, we randomly selected a δ_i and set its value to $\mu \pm 2k\tau$ for the same k as before.

We assigned the same value to σ_i^2 and the same value to τ_i^2 for all i . A real inter-laboratory data set was used to get a rough idea of a realistic value for τ^2/σ^2 . The values of $\sigma^2 = 1/4$ and $\tau^2 = 25/16$ were used. The same data set indicated that $\mu = 30$ is reasonable for the variance ratio we used. Also, the same data set indicated that $k = 6$ provides a reasonable representation of outlying laboratory means. We arbitrarily fixed v to be 10. Finally, when computing μ_p and s_p (the center and variance of the approximate posterior distribution under the BLM), we used the same value of τ^2 that is used to generate data. (This gives the Bayes procedures under the BLM a

potential advantage when comparing them to the Bayesian procedures under the other models.) See Fig. 1 for a representation of data generated under two different scenarios.

Versions of procedures compared

We did a four-factor simulation study and compare the credible intervals obtained from the posterior distributions of μ for the BLM, TLM, 2CMTLM, 3CMTLM, and PSMTLM. The four factors with their levels are:

1. The basic model assumptions that generate the data (the data model and lab means model under the BLM and the TLM).
2. The number of laboratories in the study (5, 10 and 20).
3. The number of observations per laboratory (3, 5, and 10).
4. The number of outlying laboratories present (none, one, and three with one being “extreme”).

Under each scenario, 1000 data sets were generated, and for each one five posterior distributions of μ corresponding to the five models outlined were obtained. From these five posterior distributions, five credible intervals were computed and were compared in terms of interval width and coverage relative frequency. The BLM posterior distribution of μ is obtained after computing m_p and s_p . The TLM, 2CMTLM, 3CMTLM, and PSMTLM posterior distributions for μ were approximated with 20000 simulated MCMC draws after a burn-in period of 30000 and thinning of 5. Convergence of the chains was assessed using the

gibbsit function in the statistical software R [12] from Raftery et al. [13]. For a few data sets under each scenario, the MCMC algorithms were run convergence was confirmed. For the remaining MCMC chains, convergence was assumed, although this was not explicitly checked.

Results

The results from the simulation study (1000 simulated data sets) are presented in tables below. In these tables, the header “ m ” represents the number of observations per laboratory. The column “cp” is the absolute fraction of data sets that produce 95% credible intervals that contain the “true” μ (which in this study is 30). The columns “ciwid” provide the median credible interval width across the 1000 data sets. The column “Data generating process” indicates under which data model and lab mean model the data sets were generated, and the column “#Labs” indicates the number of participating laboratories. The “Bayes Procedure” column indicates which procedure was used to estimate μ .

The next three sections detail the results of the simulation study. It is worth noting that increasing the number of observations per laboratory did not change results and we do not discuss this aspect of the study further.

No outliers

First, we compare the performance of the Bayesian procedures based on the five models when no laboratories are

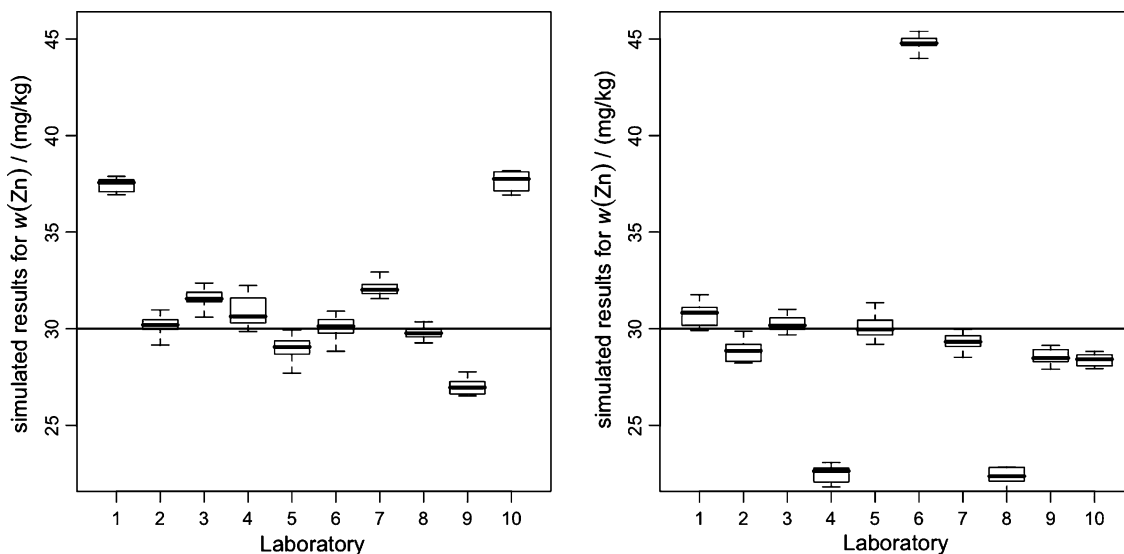


Fig. 1 Example of two data sets containing 10 laboratories each with 10 observations that were generated in the simulation study. Here, w denotes mass fraction. The solid black line highlights the value (30) which was used to generate laboratory means. The left hand panel displays a data set with no outlying laboratories. The right hand panel displays a data set with two outliers and one “extreme” outlier. The

same scale was used on both plots to facilitate comparison. Each boxplot summarizes the observations from one laboratory: in each of these, the rectangular region comprises the middle 75% of the data, the horizontal line drawn through the rectangle represents the median of the observations, and the vertical lines extend to the extrema

outliers. Table 2 provides a summary of the results. For the most part, the five methods give comparable answers. As expected, about 95% of the credible intervals computed using the Δ BLM contained the “truth” when the lab means were drawn from a normal distribution, but not when lab means come from a t -distribution. For all five procedures, the median credible interval length decreased as the number of laboratories increased with Δ 3CMTLM having the slowest rate of decrease. Overall, when an inter-laboratory study produces data with no outlying laboratories, the answers one gets by using Bayes procedures based on more complicated models (here the mixture models) differ little from those obtained from Bayes procedures for simpler models (the BLM and the TLM).

One outlier

The results from the simulation study with one outlying laboratory are summarized in Table 3. Here, we see a shortcoming of Δ BLM, since the only about 20% of the credible intervals computed using this procedure contained the “truth” while approximately 95% of the credible intervals associated with the Δ TLM, Δ 2CMTLM, Δ 3CMTLM, Δ PSMTLM were “correct.” When only considering procedures based on the models with t -distributions, Δ PSMTLM produces the shortest median length among the four, but has the lowest coverage relative frequency. As expected, the outlying laboratory’s effect diminishes as the number of laboratories increases.

Table 2 Results from the simulation study when none of the laboratories are outliers

Data generating process	#Labs	Bayes procedure	$m = 3$		$m = 5$		$m = 10$	
			cp	ciwid	cp	ciwid	cp	ciwid
BLM	5	Δ BLM	0.9440	2.2447	0.9610	2.2251	0.9360	2.2086
		Δ TLM	0.9010	2.0126	0.9150	2.0104	0.9010	2.0154
		Δ 2CMTLM	0.8980	2.0082	0.9160	2.0077	0.9020	2.0090
		Δ 3CMTLM	0.9030	2.0570	0.9220	2.0436	0.9060	2.0494
		Δ PSMTLM	0.8990	2.0361	0.9170	2.0346	0.9020	2.0384
	10	Δ BLM	0.9420	1.5873	0.9380	1.5734	0.9500	1.5617
		Δ TLM	0.9080	1.4625	0.9130	1.4738	0.9070	1.4734
		Δ 2CMTLM	0.9100	1.4673	0.9170	1.4831	0.9110	1.4825
		Δ 3CMTLM	0.9250	1.6740	0.9320	1.6839	0.9340	1.6624
		Δ PSMTLM	0.9100	1.4722	0.9170	1.4755	0.9100	1.4841
	20	Δ BLM	0.9470	1.1225	0.9510	1.1126	0.9510	1.1043
		Δ TLM	0.9190	1.0919	0.9150	1.0783	0.9340	1.0891
		Δ 2CMTLM	0.9250	1.1103	0.9200	1.0969	0.9400	1.1098
		Δ 3CMTLM	0.9730	1.7444	0.9770	1.6894	0.9730	1.6963
		Δ PSMTLM	0.9200	1.0923	0.9160	1.0783	0.9360	1.0943
TLM	5	Δ BLM	0.9290	2.2444	0.9180	2.2250	0.9270	2.2084
		Δ TLM	0.9200	2.1070	0.9010	2.0619	0.9080	2.0971
		Δ 2CMTLM	0.9210	2.1081	0.8960	2.0641	0.9130	2.0969
		Δ 3CMTLM	0.9350	2.1591	0.9020	2.1067	0.9180	2.1582
		Δ PSMTLM	0.9250	2.1402	0.9000	2.0777	0.9140	2.1261
	10	Δ BLM	0.9150	1.5870	0.9160	1.5733	0.9250	1.5617
		Δ TLM	0.9140	1.5934	0.9120	1.5833	0.9060	1.5938
		Δ 2CMTLM	0.9190	1.5996	0.9140	1.5933	0.9070	1.5970
		Δ 3CMTLM	0.9380	1.8418	0.9370	1.8025	0.9340	1.8027
		Δ PSMTLM	0.9160	1.5978	0.9160	1.5841	0.9080	1.5844
	20	Δ BLM	0.9130	1.1225	0.9270	1.1127	0.9210	1.1043
		Δ TLM	0.9180	1.1789	0.9370	1.1639	0.9270	1.1771
		Δ 2CMTLM	0.9220	1.2006	0.9420	1.1885	0.9340	1.1978
		Δ 3CMTLM	0.9780	1.8341	0.9720	1.8055	0.9770	1.7913
		Δ PSMTLM	0.9150	1.1733	0.9350	1.1618	0.9290	1.1717

Under the header “cp”, we report the fraction of the 1000 credible intervals containing 30. Under the header “ciwid”, we report the median credible interval width calculated across the 1000 credible intervals

Table 3 Results from the simulation study where one laboratory was randomly selected to be an outlier

Data generating process	#Labs	Bayes procedure	$m = 3$		$m = 5$		$m = 10$	
			cp	ciwid	cp	ciwid	cp	ciwid
BLM	5	Δ BLM	0.2150	2.2437	0.2240	2.2242	0.2390	2.2082
		Δ TLM	0.9580	4.7758	0.9550	4.7474	0.9630	4.7583
		Δ 2CMTLM	0.9550	4.2252	0.9450	4.2423	0.9480	4.2858
		Δ 3CMTLM	0.9500	3.4541	0.9450	3.4073	0.9390	3.4145
		Δ PSMTLM	0.9350	2.7779	0.9340	2.7324	0.9190	2.7580
	10	Δ BLM	0.5150	1.5879	0.5270	1.5732	0.5370	1.5616
		Δ TLM	0.9610	2.4695	0.9670	2.4346	0.9670	2.4509
		Δ 2CMTLM	0.9530	2.1790	0.9510	2.0948	0.9470	2.1804
		Δ 3CMTLM	0.9560	1.8680	0.9450	1.8155	0.9440	1.8690
		Δ PSMTLM	0.9370	1.6590	0.9220	1.6012	0.9310	1.6443
	20	Δ BLM	0.7430	1.1227	0.7160	1.1127	0.7400	1.1043
		Δ TLM	0.9460	1.3354	0.9460	1.3488	0.9510	1.3184
		Δ 2CMTLM	0.9420	1.2231	0.9440	1.2286	0.9580	1.2245
		Δ 3CMTLM	0.9640	1.4639	0.9660	1.5004	0.9800	1.4587
		Δ PSMTLM	0.9280	1.1188	0.9330	1.1183	0.9470	1.1144
TLM	5	Δ BLM	0.2430	2.2425	0.2400	2.2243	0.2300	2.2083
		Δ TLM	0.9550	4.8222	0.9470	4.8222	0.9510	4.8253
		Δ 2CMTLM	0.9510	4.2746	0.9460	4.3509	0.9420	4.4027
		Δ 3CMTLM	0.9470	3.4593	0.9420	3.5441	0.9400	3.6358
		Δ PSMTLM	0.9390	2.8136	0.9250	2.9096	0.9210	2.9816
	10	Δ BLM	0.5610	1.5875	0.5430	1.5736	0.5390	1.5618
		Δ TLM	0.9740	2.5936	0.9530	2.5887	0.9620	2.5467
		Δ 2CMTLM	0.9580	2.3911	0.9330	2.3851	0.9400	2.3575
		Δ 3CMTLM	0.9540	2.0777	0.9340	2.0692	0.9370	2.0466
		Δ PSMTLM	0.9340	1.7579	0.9160	1.7623	0.9130	1.7504
	20	Δ BLM	0.7470	1.1226	0.7190	1.1128	0.7220	1.1043
		Δ TLM	0.9550	1.4455	0.9410	1.4439	0.9500	1.4403
		Δ 2CMTLM	0.9550	1.3735	0.9430	1.3955	0.9400	1.3837
		Δ 3CMTLM	0.9720	1.6218	0.9610	1.6299	0.9730	1.6075
		Δ PSMTLM	0.9450	1.2137	0.9310	1.2282	0.9300	1.2278

Under the header “cp”, we report the fraction of the 1000 credible intervals containing 30. Under the header “ciwid” we report the median credible interval width calculated across the 1000 credible intervals

Overall, when an inter-laboratory study produces data with an outlying laboratories, Δ 3CMTLM and Δ PSMTLM perform quite well in balancing coverage relative frequency and interval length.

Three outliers

Table 4 provides a summary of the simulation results when two laboratories were outliers and a third was an “extreme” outlier. Note that the situation when there are five laboratories is extreme, as more than half of the laboratories are outliers. Hence, the credible intervals are very uninformative. Here, Δ 3CMTLM and Δ PSMTLM have comparable

median credible interval lengths and coverage relative frequencies in all scenarios. The same holds true for Δ TLM and Δ 2CMTLM. The former duo has shorter credible intervals compared to the latter regardless of the number of laboratories. In contrast, the latter duo has higher coverage relative frequencies compared to the former for all scenarios except that when the number of laboratories is 5. Generally speaking, it appears that using Δ PSMTLM or Δ 3CMTLM to analyze data from inter-laboratory studies produces coverage relative frequencies and credible interval widths at least as favorable as those from the Δ BLM, Δ TLM, and Δ 2CMTLM. But this advantage diminishes as the number of laboratories increases.

Table 4 Results from the simulation study where two laboratories were randomly selected to be outliers with one additional laboratory being an “extreme” outlier

Data generating process	#Labs	Bayes procedure	$m = 3$		$m = 5$		$m = 10$	
			cp	ciwid	cp	ciwid	cp	ciwid
BLM	5	Δ BLM	0.5060	2.2453	0.4960	2.2245	0.4900	2.2087
		Δ TLM	0.5370	11.9303	0.5240	11.3983	0.5110	10.7906
		Δ 2CMTLM	0.5250	10.8552	0.5210	10.8928	0.5020	10.8439
		Δ 3CMTLM	0.7720	10.4665	0.7790	10.4085	0.7610	10.4068
		Δ PSMTLM	0.8610	10.9107	0.8990	10.7670	0.9020	10.7075
	10	Δ BLM	0.5010	1.5880	0.5000	1.5737	0.5080	1.5616
		Δ TLM	0.9750	5.4523	0.9710	5.4569	0.9600	5.3987
		Δ 2CMTLM	0.9860	5.5723	0.9760	5.6016	0.9700	5.5620
		Δ 3CMTLM	0.9400	2.2056	0.9350	2.4512	0.9310	2.2394
		Δ PSMTLM	0.9460	2.1701	0.9360	2.2732	0.9320	2.2471
	20	Δ BLM	0.4920	1.1231	0.4630	1.1126	0.4940	1.1043
		Δ TLM	0.9540	1.4915	0.9560	1.4746	0.9620	1.4546
		Δ 2CMTLM	0.9600	1.6315	0.9600	1.5974	0.9690	1.5761
		Δ 3CMTLM	0.9470	1.2011	0.9310	1.2083	0.9490	1.1978
		Δ PSMTLM	0.9450	1.2006	0.9310	1.2010	0.9380	1.1955
TLM	5	Δ BLM	0.4950	2.2434	0.4820	2.2245	0.4810	2.2082
		Δ TLM	0.5390	11.8010	0.5150	10.8360	0.5120	10.8453
		Δ 2CMTLM	0.5280	10.8229	0.4990	10.7027	0.5000	10.7564
		Δ 3CMTLM	0.7550	10.6090	0.7420	10.5143	0.7170	10.4669
		Δ PSMTLM	0.8530	10.9938	0.8550	10.7682	0.8590	10.6688
	10	Δ BLM	0.4840	1.5873	0.4950	1.5739	0.4890	1.5617
		Δ TLM	0.9620	5.6568	0.9680	5.6010	0.9590	5.6226
		Δ 2CMTLM	0.9720	5.7769	0.9730	5.7295	0.9700	5.7487
		Δ 3CMTLM	0.9490	2.8456	0.9380	2.8147	0.9470	2.901
		Δ PSMTLM	0.9460	2.5406	0.9330	2.5548	0.9300	2.6721
	20	Δ BLM	0.4610	1.1226	0.4730	1.1126	0.4510	1.1043
		Δ TLM	0.9680	1.6307	0.9680	1.5831	0.9670	1.5703
		Δ 2CMTLM	0.9720	1.8314	0.9730	1.7669	0.9690	1.7477
		Δ 3CMTLM	0.9570	1.3617	0.9400	1.3409	0.9480	1.3391
		Δ PSMTLM	0.9450	1.3126	0.9310	1.3024	0.9470	1.2929

Under the header “cp”, we report the fraction of the 1000 credible intervals containing 30. Under the header “ciwid”, we report the median credible interval width calculated across the 1000 credible intervals

General remarks

It is interesting that under all scenarios, as the number of laboratories increases, the coverage relative frequency for Δ 3CMTLM increases (this is particularly true when there are not outliers present). This could indicate that this procedure is over fitting the data (particularly when there are no outliers present). Also, for all procedures, introducing an outlying laboratory improves the coverage relative frequency. This might be due to the fact that an outlying laboratory creates wider credible intervals making them more conservative.

Analyses of data from a NIST/NOAA inter-laboratory study

Here, we apply the five procedures outlined in previous sections to data coming from a NIST/NOAA inter-laboratory study conducted to measure trace elements in marine mammals. We provide a brief description of the inter-laboratory protocol and refer interested readers to details in Christopher et al. [14]. In this study, 33 laboratories measured concentrations of several analytes on marine mammal tissue [white-sided dolphin liver homogenate (QC04LH4)] with varying replicates per laboratory. We

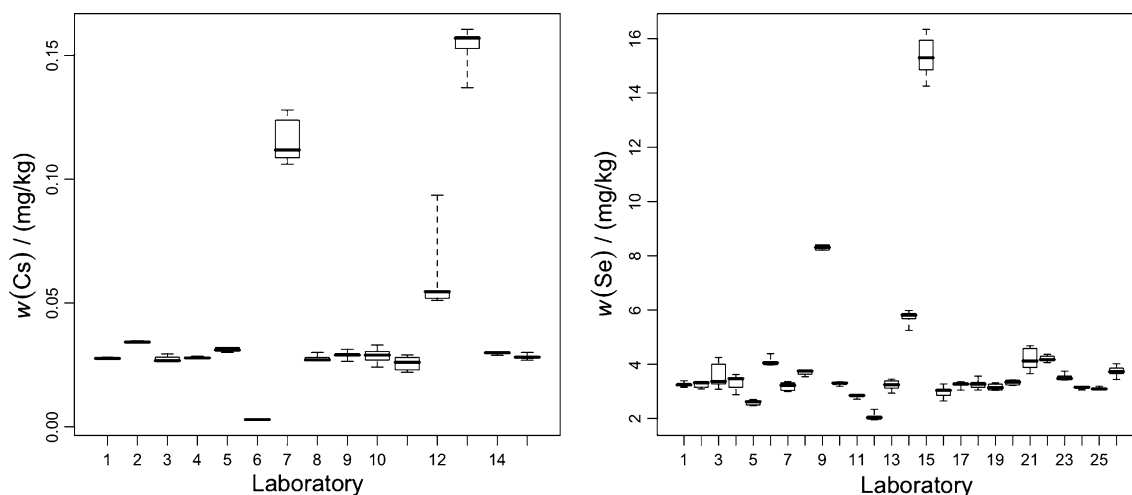


Fig. 2 Side-by-side boxplots depicting results from the NIST/NOAA sponsored inter-laboratory study. The material used was white-sided dolphin (*Lagenorhynchus acutus*) liver homogenate (QC04LH4). Here, w denotes mass fraction. The figure on the left displays the mass fraction results for Cs (left) and figure on the right Se. Each boxplot

summarizes the observations from one laboratory: in each of these, the rectangular region comprises the middle 75% of the data, the horizontal line drawn through the rectangle represents the median of the observations, and the vertical lines extend to the extrema

Table 5 Means and 95% intervals from the posterior distributions of μ obtained using the five procedures outlined in the article and results from the NIST/NOAA sponsored inter-laboratory study of the mass fractions of Cs and Se (all results in mg/kg)

	Posterior mean (Cs)	95% Credible interval	Posterior mean (Se)	95% Credible interval
BLM	0.043	(0.043, 0.043)	4.052	(3.551, 4.553)
TLM	0.034	(0.019, 0.052)	3.300	(3.008, 3.590)
2CMTLM	0.030	(0.017, 0.046)	3.336	(3.020, 3.654)
3CMTLM	0.029	(0.018, 0.039)	3.277	(2.961, 3.603)
PSMTLM	0.028	(0.019, 0.038)	3.273	(3.010, 3.552)

consider only selenium (Se) and cesium (Cs) here because Se was measured by a large number of laboratories (26) while Cs was measured by a relatively small number of laboratories (15). Figure 2 provides side-by-side box-plots of the measurement results for both analytes. We use the five procedures to estimate a reference value and its uncertainty for both trace elements. For the analysis of Cs, the values of a_σ , b_σ , a_τ , and b_τ used in the simulation study were not appropriate. Their use artificially inflated the estimates of the σ_i^2 's and τ^2 . Because of this, we used $a_\sigma = a_\tau = 2.0001$ and $b_\sigma = b_\tau = 1000$. This resulted in prior distributions for the σ_i^2 's and τ^2 that had means approximately equal to 0.001 and standard deviations approximately equal to $\sqrt{10}$. Also, we set $a_\delta = 0$, $b_\delta = 0.5$. Results are provided in Table 5 and Fig. 3.

For Se, where the number of laboratories that provided measurements was fairly large (26), there is little difference in reference value estimates and intervals between the four procedures designed to be robust alternatives to that for the Gaussian lab model. However, for Cs, the credible interval coming from PSMTLM was the shortest with that coming from 3CTMLM being slightly larger. The TLM

and 2CMTLM procedures provided similar estimates and credible intervals. The results of the analysis echo a finding of the simulation study. For both trace elements, the estimates of the measurands under the BLM (which is the weighted average recommended by the GUM) were pulled toward the outlying laboratories. Hence, the BLM estimates were influenced by outlying laboratories more than the other four procedures.

Conclusions

In an inter-laboratory study setting, completely ignoring results from one or more participating laboratories is usually untenable in practical terms. Because of this, we have attempted to develop a methodology that even in cases involving outlying laboratories produces credible intervals that maintain a roughly 95% coverage relative frequencies while remaining informative. We proposed the 2CMTLM, 3CMTLM, and PSMTLM as three alternatives to using t distributions as the basis for a Bayes analyses. Then, we conducted a simulation study that

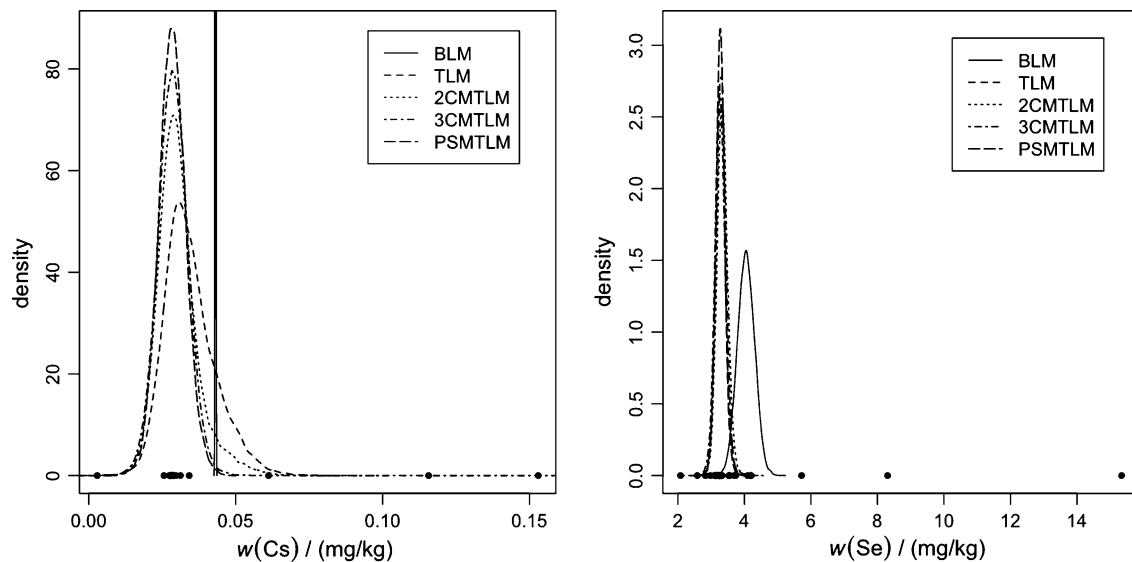


Fig. 3 Posterior distributions for μ under the five procedures. Here, w denotes mass fraction. The *solid dots* are the raw lab means from the NIST/NOAA marine mammal inter-laboratory data set.

compared the Bayesian procedures based on five models (the BLM, TLM, 2CMTLM, 3CTMLM, and PSMTLM). We found that the Bayes procedures under the 3CMTLM and the PSMTLM produce results comparable to those under the BLM and the TLM when no outliers are present and usually have better coverage relative frequencies with smaller interval widths when outliers were present. But advantages that accompany Δ 3CMTLM and Δ PSMTLM diminish as the size of the inter-laboratory study increases. The analyst should carefully weigh the advantages and disadvantages of the methods before making an analysis decision.

Acknowledgments The authors are grateful to Antonio Possolo (chief of the Statistical Engineering Division at the NIST) with help in gaining access to the marine mammal data set, to the study's sponsor, NOAA/National Marine Fisheries Service, and the Inorganic Chemical Metrology Group (Dr. Gregory Turk, Group Leader) in the Analytical Chemistry Division (Dr. Stephen Wise, Chief) of the Chemical Science and Technology Laboratory (Dr. Willie May, Director). Supported by NSF grant DMS #0502347 EMSW21-RTG awarded to the Department of Statistics, Iowa State University.

References

1. Toman B (2007) Bayesian approaches to calculating a reference value in key comparison experiments. *Technometrics* 29:81–87
2. ISO (1993) Guide to the expression of uncertainty in measurement. <http://www.bipm.org/en/publications/guides/gum>

White-sided dolphin (*Lagenorhynchus acutus*) liver homogenate (QC04LH4) was the material used in the inter-laboratory study

3. Jeffreys H. (1946) An invariant form for the prior probability in estimation problems. *Proc R Soc Lond A Math Phys Sci* 186:453–461
4. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. Chapman and Hall, New York
5. Possolo A, Toman B (2007) Assessment of measurement uncertainty via observation equations. *Metrologia* 44:464–475
6. Frühwirth-Schnatter S (2006) Finite mixture and markov switching models. Springer, New York
7. Jasra A, Holmes CC, Stephens DA (2005) Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat Sci* 20:50–67
8. Stephens M (2000) Dealing with label switching in mixture models. *J Royal Stat Soc: Ser B* 62:795–809
9. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. *J Chemical Phys* 21:1087–1091
10. Hastings WK (1970) Monte Carlo methods using markov chains and their applications. *Biometrika* 57:97–109
11. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 6:721–741
12. R Development Core Team (2009) R: a language and environment for statistical computing. R foundation for statistical computing, ISBN 3-900051-07-0
13. Raftery AE, Lewis S (1992) How many iterations in the gibbs sampler?
14. Christopher SJ, Pugh RS, Ellisor MB, Mackey EA, Spatz RO, Porter BJ, Bealer KJ, Kucklick JR, Rowles TK, Becker PR (2007). Description and results of the NIST/NOAA 2005 inter-laboratory comparison exercise for trace elements in marine mammals. *Accred Qual Assur* 12:175–187