# Bayesian Local Contamination Models for Multivariate Outliers

Garritt L. Page
Department of Statistical Science
Duke University
page@stat.duke.edu

David B. Dunson
Department of Statistical Science
Duke University
dunson@stat.duke.edu

**Abstract**

In studies where data are generated from multiple locations or sources it is common for there to exist observations that are quite unlike the majority. Motivated by the application of establishing a reference value in an inter-laboratory setting when outlying labs are present, we propose a local contamination model that is able to accommodate unusual multivariate realizations in a flexible way. The proposed method models the process level of a hierarchical model using a mixture with a parametric component and a possibly non-parametric contamination. Much of the flexibility in the methodology is achieved by allowing varying random subsets of the elements in the lab-specific mean vectors to be allocated to the contamination component. Computational methods are developed and the methodology is compared to three other possible approaches using a simulation study. We apply the proposed method to a NIST/NOAA sponsored inter-laboratory study which motivated the methodological development.

**Keywords**: Bayesian robustness; Component-wise classification; Local contamination; Mixtures; Multivariate outliers; Inter-laboratory Studies

## 1 Introduction

Random effects or multi-level models are commonly used to model data that have a hierarchical or nested structure. These types of models are appealing because of their wide applicability and the availability of estimation and inference for subject-specific and global parameters. As with many statistical procedures, estimation and inference available from

these models are sensitive to the presence of observations that are unlike the majority. If the data have a multivariate structure, then handling outliers can be even more complicated as observation vectors can be composed of a combination of elements, some of which are similar in magnitude to the majority of other observations and some of which are outlying.

As an example, consider a NIST/NOAA sponsored inter-laboratory study conducted to assess the quality of trace element measurements in the marine mammal population. The 33 participating laboratories were instructed to produce 5 replicate measurements of the concentration of 15 trace elements in a sample of marine mammal tissue. Figure 1 provides boxplots of measured concentrations of arsenic (As) and selenium (Se) where each boxplot corresponds to a lab. Notice that lab 18 recorded concentrations for both trace elements that are much larger than the other labs. Conversely, Lab 17 reported concentrations for Se that are larger than the majority of the labs, but the concentrations measured for As seem quite reasonable relative to the majority of the other labs. In light of this, there is a need to develop methods that are able to handle multivariate outliers in a flexible way.

There is a large literature devoted to the development of statistical methodologies that are robust to the presence of outliers. Rocke (1983), Mandel (1995) and Bednarski and Zontek (1996) handle outliers by proposing robust estimators, such as M-estimators or estimators resulting from Frèchet differentiable functionals. Muller and Uhlig (2001) and Lischer (1996) propose estimators based on the differences between observations. Another approach is to discard all measurements that are classified as outliers via a detection method such as those proposed in Peña and Prieto (2001) and Penny and Jolliffe (2001). However, the uncertainty associated with outlier detection is not considered. Song et al. (2007) provide an example of accommodating outliers by using heavy-tailed distributions. Although the fatter tails of the $t$-distribution more readily accommodate outliers compared to a Gaussian distribution,

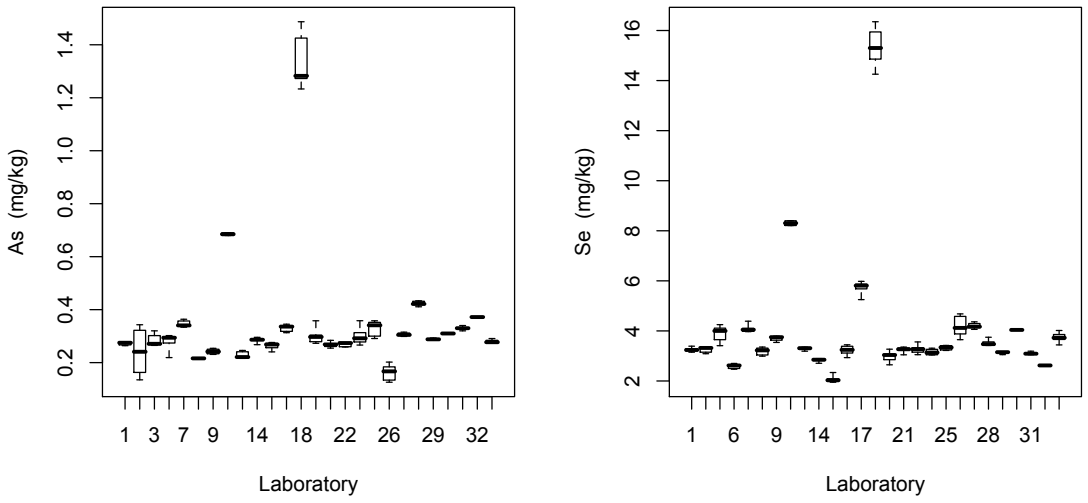


Figure 1: Boxplots for concentrations of As and Se measured in the NIST/NOAA sponsored inter-laboratory study and recorded as mass fraction mg per gram

it still is fairly restrictive in that its shape is symmetric and uni-modal.

Introducing some distributional perturbation by way of a finite mixture is another approach. Box and Tiao (1968) first proposed this method and used a two-component mixture to accommodate outliers. Approaches that use mixtures of some kind can be quite flexible but typically classify an entire multivariate object as being an outlier or not.

In this article we develop a Bayesian methodology that allows the different elements of a multivariate vector to vary in their outlier status, while accommodating uncertainty in outlier classification in estimation and inferences. The general idea is to introduce some distributional contamination (Bayarri and Morales (2003)) to accommodate multivariate outliers. This is done by constructing a hierarchical model where the process is modeled with a mixture of mixtures. A majority (non-outlying) component is modeled by a multivariate normal,

while the other component corresponds to unusual observations and is modeled with a finite mixture. Much of the flexibility in the methodology is achieved by allowing allocation to the non-Gaussian component to vary for the different vector elements. One might think of this as a type of local clustering of the location/source mean vectors. The notion of local clustering or clustering multivariate objects on a subset of attributes has been introduced in the literature. Specifically, Dunson (2009) uses the local clustering idea to choose a prior for an unknown random effects distribution within a hierarchical model and Hoff (2006) develops a model-based clustering approach that clusters a multivariate vector of attributes using a subset of the attributes.

The methods developed here are applicable to any study in which multivariate measurements arising from different studies are to be combined. This could include, for example, meta-analysis and interlaboratory studies. For clarity and ease of exposition we motivate ideas from an inter-laboratory perspective. Because of this we provide a very brief introduction to inter-laboratory studies here. These studies are conducted to ensure measurement capability for commerce, evaluate national and international equivalence of measure, and validate measurement devices and methods or standard materials. Typically, the overarching goal in the analysis of data produced by inter-laboratory studies is to establish a reference value, which can sometimes be thought of as an estimate of a measurand (quantity intended to be measured), and estimate its uncertainty. To determine a "degree of equivalence" each laboratory's measurements are compared to the reference value. This analysis is usually carried out under the guidelines set forth in *The Guide to the Expression of Uncertainty in Measurement* (GUM) created by the International Organization of Standardization (ISO). Since its inception criticisms and alternative approaches to estimating a reference value have been proposed in the literature (Gleser (1998), Rukhin and Vangel (1998), Rukhin (2007) and Toman (2007)).

In what follows, Section 2 provides a detailed description of the model and Section 3 develops computational methods. Section 4 describes a simulation study conducted to compare the performance of the proposed method to three reasonable alternatives. Section 5 provides an example using a NIST/NOAA marine mammal inter-laboratory data set. In Section 6 we make conclusions.

# 2 Description of the Local Contamination Model

Though we focus on an inter-laboratory application for clarity, the methodology can be implemented in a more general setting. The proposed model will be referred to as a local contamination (LC) model. We assume that the replicates for each laboratory vary according to a normal distribution and that if a laboratory makes an unusual measurement for a particular element (compared to the rest of the labs) it will continue to do so for that element. However, within laboratory outliers can be handled in a fairly straightforward manner through the use of a heavy tailed residual density such as a multivariate $t$ distribution (e.g., Section 5.4). For sake of simplicity and ease of exposition we describe the LC model with a Gaussian residual distribution.

Let $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijp})'$ be the $j$th vector of $p$ measurements taken by the $i$th laboratory with $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. Assume that

$$\boldsymbol{y}_{ij} \overset{ind}{\sim} \mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

with $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip})'$ a lab-specific mean and $\boldsymbol{\Sigma}_i$ a lab-specific covariance which we model with an inverse-Wishart (i.e. $\boldsymbol{\Sigma_i} \sim IW(v, \boldsymbol{A})$). For studies where measurements are

physically constrained to be nonnegative one can simply log-transform the measurements prior to analysis.

A novel contribution of the LC model is how the $\boldsymbol{\mu}_i$'s are modeled. Ultimately, we model the $\boldsymbol{\mu}_i$ with a multivariate normal whose mean and variance allow each $\boldsymbol{\mu}_i$ to potentially be composed of elements that are members of a majority (non-outlying) component and others that are members of a non majority (outlying) component which we refer to as a contamination. This latent allocation structure is modeled by introducing classification variables that classify the $p$ elements of each lab mean vector $\boldsymbol{\mu}_i$ as being part of the majority or not. Let $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{ip})'$ be a $p$ dimensional vector of 0's and 1's such that

$$\gamma_{ik} = \begin{cases} 1 & \text{if } \mu_{ik} \in \mathcal{M}_i \\ 0 & \text{if } \mu_{ik} \notin \mathcal{M}_i \end{cases}$$

where $\mathcal{M}_i$ is a collection of elements that make up the majority component for the $i$th lab, $d_i = \sum_{k=1}^{p} \gamma_{ik}$ is the number of elements in the majority component for the $i$th lab and $\gamma_{ik} \overset{ind}{\sim} \text{Ber}(\pi_i)$ with $\pi_i \overset{iid}{\sim} \text{Beta}(a_\pi, b_\pi)$ ($a_\pi$ and $b_\pi$ are user supplied). $\pi_i$ is the prior probability that lab $i$ falls within the majority component for a randomly-selected element. By specifying a hyperprior for $\pi_i$, we allow the data to inform about the proportion of outlying elements. Typically, one would elicit $a_\pi$ and $b_\pi$ based on prior knowledge of the proportion of outlying elements with $a_\pi \gg b_\pi$ so that $E(\pi_i) \gg 0.5$, with ideally $E(\pi_i) \approx 1 - \epsilon$ for small $\epsilon$. This leads to a local $\epsilon$-contaminated model.

We now introduce atoms $\{(\boldsymbol{\mu}_h^*, \boldsymbol{\Sigma}_h^*); h = 0, 1, \ldots, L\}$ that are used to construct the mean and covariance matrix corresponding to $\boldsymbol{\mu}_i$. The pair $(\boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0^*)$ are the mean vector and covariance matrix that correspond to the majority component (which makes $\boldsymbol{\mu}_0^*$ of principle interest in an inter-laboratory study analysis), while $(\boldsymbol{\mu}_h^*, \boldsymbol{\Sigma}_h^*)$ for $h = 1, \ldots, L$ correspond to

the $L$ clusters that make up the contamination. $\boldsymbol{\mu}_0^*$ is sampled *a priori* from $N_p(\boldsymbol{m}, \boldsymbol{S})$ while $\boldsymbol{\mu}_h^*$'s for $h = 1, \ldots, L$ are sampled independently from $t_p(\boldsymbol{m}, \boldsymbol{S}, \nu)$ *a priori*. (Here $t_p(\boldsymbol{m}, \boldsymbol{S}, \nu)$ denotes a scaled ($\boldsymbol{S}$) and shifted ($\boldsymbol{m}$) $p$-dimensional $t$-distribution with $\nu$ degrees of freedom.) Multivariate $t$-distributions are used to accommodate the possibility of contamination cluster locations being highly variable. For computational purposes (detailed in Section 3) the $\boldsymbol{\Sigma}_h^*$'s are diagonal matrices with diagonal entries $\sigma_{hk}^{*2},\ k = 1, \ldots, p$. To preserve conjugacy we use an inverse-Gamma prior $\sigma_{hk}^{*2} \overset{ind}{\sim} IG(a_\sigma, b_\sigma)$ for $k = 1, \ldots, p$ (values for $a_\sigma, b_\sigma$ are supplied by the user). Thus, depending on $\boldsymbol{\gamma}_i$, the mean and covariance matrix corresponding to $\boldsymbol{\mu}_i$ is made up of elements from $\boldsymbol{\mu}_0^*$ and $\boldsymbol{\Sigma}_0^*$ and/or elements from $\boldsymbol{\mu}_\ell^*$ and $\boldsymbol{\Sigma}_\ell^*$ for some $\ell = 1, \ldots, L$.

The value of $L$ actually represents an upper bound on the number of clusters that make up the contamination. The form of the contamination is motivated by a finite Dirichlet approximation to the Dirichlet process as proposed by Ishwaran and Zarepour (2002). As this upper bound increases, there is convergence to a nonparameteric limit, but the finite approximation is somewhat simpler to implement. Allowing the contamination to potentially consist of more than one cluster provides more flexibility in handling outliers compared to a one cluster contamination. The necessity of this flexibility can be seen in Figure 1. It is fairly obvious that the As and Se entries of Lab 18's mean vector will occupy a cluster of the contamination. Also, it is possible that the As entry of Lab 26's mean vector will be allocated to the contamination. If this turns out to be the case, then at least two contamination clusters will be necessary to accommodate the two element means.

To identify the $\ell$th contamination cluster used to construct the mean vector and covariance matrix associated with $\boldsymbol{\mu}_i$ we introduce an $S_i$ for each of the $i = 1, \ldots, m$ labs. The $S_i$'s take on values $\ell = 1, \ldots, L$ with $Pr(S_i = \ell) = \nu_\ell$. For simplicity, all elements of the mean and covariance matrix of $\boldsymbol{\mu}_i$ that are allocated to the contamination come from the same

contamination cluster. The vector of probability weights on the $L$ clusters that make up the contamination is modeled as

$$\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_L)' \sim \text{Dir}(\alpha/L, \ldots, \alpha/L)$$

with $\alpha = 1$ being fixed in the applications that we consider to favor high weights on few clusters.

Finally, we model the $\boldsymbol{\mu}_i$'s independently for $i = 1, \ldots, m$ with

$$\boldsymbol{\mu}_i \overset{ind}{\sim} \text{N}_p(\boldsymbol{\delta}_{\gamma_i}, \boldsymbol{T}_{\gamma_i})$$

The entries of the vector $\boldsymbol{\delta}_{\gamma_i}$ depend on $\gamma_{ik}$ in the following way. If $\gamma_{ik} = 1$, then $\delta_{\gamma_{ik}} = \mu^*_{0k}$, otherwise, $\delta_{\gamma_{ik}} = \mu^*_{S_i k}$ for all $k = 1, \ldots, p$. This structure can be succinctly written as

$$\boldsymbol{\delta}_{\gamma_i} = \boldsymbol{\gamma}_i \otimes \boldsymbol{\mu}^*_0 + (1 - \boldsymbol{\gamma}_i) \otimes \boldsymbol{\mu}^*_{S_i}$$

where $\otimes$ denotes the Hadamard or element-wise multiplication.

$\boldsymbol{T}_{\gamma_i}$ is constructed in a similar fashion. Since $\boldsymbol{\Sigma}^*_h$, for $h = 0, \ldots, L$ are diagonal matrices $\boldsymbol{T}_{\gamma_i}$ is diagonal as well. The $k$th entry on the diagonal of $\boldsymbol{T}_{\gamma_i}$ is the $k$th diagonal value of $\boldsymbol{\Sigma}^*_0$ (or $\sigma^{*2}_{0k}$) if $\gamma_{ik} = 1$ otherwise, it is the $k$th diagonal entry of $\boldsymbol{\Sigma}^*_{S_i}$ (or $\sigma^{*2}_{S_i k}$). Thus we can construct $\boldsymbol{T}_{\gamma_i}$ as

$$\boldsymbol{T}_{\gamma_i} = \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i' \otimes \boldsymbol{\Sigma}^*_0 + (1 - \boldsymbol{\gamma}_i)(1 - \boldsymbol{\gamma}_i)' \otimes \boldsymbol{\Sigma}^*_{S_i}.$$

Figure 2 provides a graphical representation of how the mean and covariance of $\boldsymbol{\mu}_i$ is con-

structured.

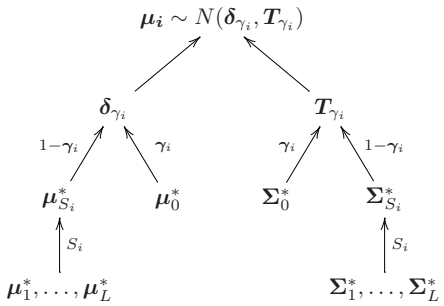


Figure 2: Graphical display of how the mean and covariance matrix of $\boldsymbol{\mu}_i$ is constructed. Recall that $S_i = \ell$ with probability $\nu_\ell$ for $\ell = 1, \ldots, L$ and $(\delta_{\gamma_i k} = \mu^*_{0k}, t^2_{\gamma_i k} = \sigma^{*2}_{0k})$ with probability $\pi_i$ for $i = 1, \ldots, m$

Values for all parameters denoted by Roman letters need to be specified by the user. To automate prior specification and avoid problems inherent to dealing with multivariate data on vastily different scales, we recommend standardizing the $p$ variables prior to analysis. If this is done, the prior specification of $\boldsymbol{m} = \boldsymbol{0}$ and $\boldsymbol{S} = \boldsymbol{I}_p$ is natural. because the $p$ variables are standardized using all data (including any outliers), values for $(a_\sigma, b_\sigma)$ should be chosen so the majority of prior distribution mass associated with the $\sigma^2_{hk}$'s is less than 1. Informal investigation indicated that inferences from the LC procedure are fairly insensitive to the prior specification for $\sigma^2_{hk}$. In this paper we use an inverse-gamma distribution with mean 0.25 and standard deviation 0.5 giving $Pr(\sigma^2_{hk} > 1) \approx 0.025$. Following suggestions made by Verdinelli and Wasserman (1991), to make probability of being an outlier small ($\approx 0.05$), we set $a_\pi = 9.5$ and $b_\pi = 0.5$.

# 3 Computation

The joint posterior distribution of the parameters in the LC model is analytically intractable. We use a Gibbs sampler to obtain correlated draws from the posterior distribution. The full conditional distributions which can be used to construct a Markov Chain whose stationary distribution is the joint posterior are described. In what follows we use $[\theta|-]$ to denote the distribution of $\theta$ conditioned on all other parameters. As an example, $[\mu_i|-]$ is shorthand for $[\mu_i|\{\boldsymbol{\mu}_j\}_{j\neq i}, \{\boldsymbol{\Sigma}_i\}_{i=1}^m \{\boldsymbol{\gamma}_i\}_{i=1}^m, \{\boldsymbol{\mu}_h^*\}_{h=0}^L, \{\boldsymbol{\Sigma}_h^*\}_{h=0}^L, \{\pi_i\}_{i=1}^m, \{\nu\}_{\ell=1}^L, \{S_i\}_{i=1}^m, \boldsymbol{y}]$. The following full conditionals are fairly common and can be derived using routine algebra.

$$[\boldsymbol{\mu}_i|-] \sim \mathrm{N}_p\left(\left[n_i\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{T}_{\boldsymbol{\gamma}_i}^{-1}\right]^{-1}\left[n_i\boldsymbol{\Sigma}_i^{-1}\bar{\boldsymbol{y}}_{ij} + \boldsymbol{T}_{\boldsymbol{\gamma}_i}^{-1}\boldsymbol{\delta}_{\boldsymbol{\gamma}_i}\right], \left[n_i\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{T}_{\boldsymbol{\gamma}_i}^{-1}\right]^{-1}\right), \tag{2}$$

$$[\boldsymbol{\Sigma}_i|-] \sim \mathrm{IW}\left(n_i + v, \sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i)(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i)' + \boldsymbol{A}\right), \tag{3}$$

$$[\gamma_{ik}|-] \sim \mathrm{Ber}(p^*) \text{ with } p^* = \frac{\phi(\mu_{ik}; \mu_{0k}^*, \sigma_{0k}^{2*})\pi_i}{\phi(\mu_{ik}; \mu_{0k}^*, \sigma_{0k}^{2*})\pi_i + \phi(\mu_{ik}; \mu_{S_ik}^*, \sigma_{S_ik}^{2*})(1-\pi_i)}, \tag{4}$$

$$[\pi_i|-] \sim \mathrm{Beta}\left(\sum_{k=1}^p \gamma_{ik} + a, p - \sum_{k=1}^p \gamma_{ik} + b\right), \tag{5}$$

$$[\boldsymbol{\nu}|-] \sim \mathrm{Dir}(\alpha_1^*, \ldots, \alpha_L^*) \text{ with } \alpha_\ell^* = \sum_{i=1}^m 1[S_i = \ell] + \frac{\alpha}{L}, \quad \ell = 1, \ldots, L, \tag{6}$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes a normal density with mean $\mu$ and variance $\sigma^2$.

Full conditionals are also needed for the component-specific parameters. We first provide the full conditionals for $\mu_0^*$ and $\boldsymbol{\Sigma}_0^*$. Here, we introduce some useful notation. Let $\boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}$ denote the matrix that results from setting the entries of the rows and columns of the matrix $\boldsymbol{T}_{\boldsymbol{\gamma}_i}$ that are associated with $\gamma_{ik} = 0$ for $k = 1, \ldots, p$ to zero. Similarly $\boldsymbol{T}_{0:\boldsymbol{\gamma}_i=1}$ denotes the matrix whose row and column entries associated with $\gamma_{ik} = 1$ for $k = 1, \ldots, p$ are set to zero. The full conditional for $\mu_0^*$ is

$$[\boldsymbol{\mu}_0^*|-] \sim \mathrm{N}_p \left( \left[ \sum_{i=1}^{m} \boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}^{-1} + \boldsymbol{S}_0^{-1} \right]^{-1} \left[ \sum_{i=1}^{m} \boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}^{-1} \boldsymbol{\mu}_i + \boldsymbol{S}_0^{-1} \boldsymbol{m}_0 \right], \left[ \sum_{i=1}^{m} \boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}^{-1} + \boldsymbol{S}_0^{-1} \right]^{-1} \right). \quad (7)$$

This result is a consequence of the assumption that $\boldsymbol{\Sigma}_0^*$ is diagonal.

Since $\boldsymbol{\Sigma}_0^*$ is a diagonal we consider the $\sigma_{0k}^{2^*}$'s individually for each $k$. The complete conditional for the $k$th diagonal element of $\boldsymbol{\Sigma}_0^*$ is

$$[\sigma_{0k}^{2^*}|-] \sim \mathrm{IG} \left( a_\sigma + \frac{1}{2} \sum_{i=1}^{m} \gamma_{ik}, \left[ \frac{1}{b_\sigma} + \frac{1}{2} \sum_{i=1}^{m} \gamma_{ik} (\mu_{ik} - \mu_{0k}^*)^2 \right]^{-1} \right). \quad (8)$$

Derivations of (7) and (8) are provided in the Appendix.

To facilitate computation we use a scale mixture of normal representation of the $t$ distributions that correspond to the contamination cluster locations. This requires introducing an auxiliary variable ($\omega_\ell \sim IG(w/2, 2/w)$) for each of the $L$ contamination clusters where $w$ is the degrees of freedom (which we set to 4). The complete conditionals for $\{(\boldsymbol{\mu}_h^*, \boldsymbol{\Sigma}_h^*); h = 1, \ldots, L\}$ are then

$$[\boldsymbol{\mu}_h^*|-] \sim \mathrm{N}_p \left( \left[ \sum_{i:S_i=h} (\omega_h \boldsymbol{T})_{0:\boldsymbol{\gamma}_i=1}^{-1} + \boldsymbol{S}_h^{-1} \right]^{-1} \left[ \sum_{i:S_i=h} (\omega_h \boldsymbol{T})_{0:\boldsymbol{\gamma}_i=1}^{-1} \boldsymbol{\mu}_i + \boldsymbol{S}_h^{-1} \boldsymbol{m}_h \right], \left[ \sum_{i:S_i=h} (\omega_h \boldsymbol{T})_{0:\boldsymbol{\gamma}_i=1}^{-1} + \boldsymbol{S}_h^{-1} \right]^{-1} \right), \quad (9)$$

$$[\sigma_{hk}^{2^*}|-] \sim IG \left( a_\sigma + \frac{1}{2} \sum_{i:S_i=h} (1 - \gamma_{ik}), \left[ \frac{1}{b_\sigma} + \frac{1}{2} \sum_{i:S_i=h} (1 - \gamma_{ik}) (\mu_{ik} - \mu_{hk}^*)^2 \right]^{-1} \right). \quad (10)$$

The full conditional of the auxilliary variable $\omega$ is

$$[\omega_h|-] \sim \text{IG}\left(0.5(w+p), (0.5[(\boldsymbol{\mu}_h - \boldsymbol{m})'\boldsymbol{S}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{m}) + w])^{-1}\right). \tag{11}$$

The full conditional of $S_i$ is discrete with probability mass function

$$Pr[S_i = \ell|-] = \frac{\boldsymbol{\phi}(\boldsymbol{\mu}_i; \boldsymbol{\delta}_{\ell\gamma_i}, \boldsymbol{T}_{\ell\gamma_i})\nu_\ell}{\sum_{h=1}^{L} \boldsymbol{\phi}(\boldsymbol{\mu}_i; \boldsymbol{\delta}_{h\gamma_i}, \boldsymbol{T}_{h\gamma_i})\nu_h} \quad \text{for} \quad \ell = 1, \ldots, L. \tag{12}$$

Where $\boldsymbol{\phi}(;\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal density, $\boldsymbol{\delta}_{\ell\gamma_i} = \boldsymbol{\gamma}_i \otimes \boldsymbol{\mu}_0^* + (1 - \boldsymbol{\gamma}_i) \otimes \mu_{S_i=\ell}^*$, and $\boldsymbol{T}_{\ell\gamma_i} = \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i' \otimes \boldsymbol{\Sigma}_0^* + (1 - \boldsymbol{\gamma}_i)(1 - \boldsymbol{\gamma}_i)' \otimes \boldsymbol{\Sigma}_{S_i=\ell}^*$.

A Markov chain associated with the joint distribution of interest can be had by iteratively cycling through the complete conditional distributions on an individual basis.

When using auxiliary variables for classifying in a mixture setting it is possible for chains coming from an MCMC algorithm to mix poorly. To improve mixing, we use an adaptive-type of Gibbs sampler (Roberts and Rosenthal (2007)). This was done by replacing $p^*$ in (4) with

$$p^* + \exp\{-0.01(t-1)\}(0.5 - p^*)$$

where $t = 1, \ldots, M$ denotes the $t^{\text{th}}$ MCMC iterate. This initially pulls $p^*$ to 0.5 but then converges to (4) exponentially fast as $t \to \infty$. Since the adaptation vanishes at an exponential rate, the necessary regularity conditions are satisfied and the algorithm converges to the correct distribution.

# 4 Simulation Study

We compare the performance of the LC model in estimating a reference vector to three other reasonable alternatives by way of a simulation study. The simulation study consists of generating several of multivariate data sets (with and without outlying elements) and for each, estimating a reference vector using three competing methods and the LC model. The four procedures were compared using frequentist metrics such as bias and mean square error along with credible region area and coverage. In this section, we briefly describe the competing methods, detail how data sets were generated, and provide the simulation study results.

## 4.1 Description of Competing Methods

We compare the performance of the LC procedure in estimating $\boldsymbol{\mu}_0^*$ to three reasonable alternatives via a simluation study. The description of the three competitors follows.

1. The first competing method is a random effects model with random effects assumed to originate from a MVN distribution. Specifically, $\boldsymbol{y}_{ij} \overset{ind}{\sim} N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $\boldsymbol{\mu_i} \overset{iid}{\sim} N_p(\boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0)$, and $\boldsymbol{\mu}_0^* \sim N_p(\boldsymbol{m}, \boldsymbol{S})$. $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma_0}$ are drawn from inverse-Wishart distributions. We refer to this model as the MVN model.

2. The second competing method is a random effects model with random effects assumed to originate from a multivariate $t$-distribution. More specifically, $\boldsymbol{y}_{ij} \overset{ind}{\sim} N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $\boldsymbol{\mu_i} \overset{iid}{\sim} t_p(\boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0, \nu)$, and $\boldsymbol{\mu}_0^* \sim N_p(\boldsymbol{m}, \boldsymbol{S})$. A uniform $(0, 100)$ prior is used for $\nu$. The other parameters were assigned the same priors as in the MVN model. The $t$-distribution is often used as a robust alternative to the MVN when outliers are present and we refer to this model as the MVT model.

3. The third competing method is a random effects model with the random effects being

modeled with an unknown density. A Dirichlet Process (DP) is used as a prior for the unknown density. That is, $\boldsymbol{y}_{ij} \overset{ind}{\sim} N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}_i \sim \mathcal{P}$, and $\mathcal{P} \sim DP(\alpha P_0)$. Here $P_0$ follows a $N_p(\boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{\mu}_0^* \sim N_p(\boldsymbol{m}, \boldsymbol{S})$. Similar to the LC model we fixed $\alpha = 1$. Other parameters were assigned the same priors as in the MVN and MVT models. This model is very flexible in accommodating unusual observations and we refer to it as the DP model.

## 4.2 Creation of Synthetic Data Sets

We consider the MVN and MVT models as a data generating mechanisms. We generate $\boldsymbol{\mu}_i$ vectors from a MVN or MVT distribution after fixing values for $\boldsymbol{\mu}_0^*$ and $\boldsymbol{\Sigma}_0$. Then after fixing $\boldsymbol{\Sigma}$, lab-specific observation vectors are generated by using a MVN with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}$ and then log transformed. We used the marine mammal data set as a guide to picking appropriate values for $\boldsymbol{\mu}_0^*$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}$. The marine mammal data set was also used to choose the number of laboratories (30), the number of observations per laboratory (5), and the dimension of the observation vectors (15).

We fix the $k$th entry of $\boldsymbol{\mu}_0^*$ at the median of the $k$th element computed across all laboratories from the marine mammal data set. $\boldsymbol{\Sigma}_0$ is fixed to be a diagonal matrix whose $k$th entry is set by computing the variance between the empirical lab-specific means for the $k$th element. For the data generation process, $\boldsymbol{\Sigma}$ is also a diagonal matrix whose $k$th entry is fixed at the average of the lab-specific sample variances of the $k$th element.

For data sets containing outliers, 10 of the 30 labs are randomly selected to have outliers in at least one entry of $\boldsymbol{\mu}_i$. Five of the ten have one element randomly selected to be an outlier, two have two elements as outliers, two have five elements and one has ten elements. This outlier structure is similar to that found in the marine mammal data set. An

outlier is generated by setting the mean for an outlying element to $\mu^*_{0k} + 6\sigma_{0kk}$. Figure 3 provides an example of the data sets used in the simulation study with and with out outliers.

## 4.3 Results

Using the MVN, 500 data sets with and without outliers are generated. This is repeated using MVT, giving a total of 2000 data sets. For each data set, posterior distributions of $\boldsymbol{\mu}^*_0$ were obtained using the four procedures. Posterior distributions are summarized by posterior mean vectors and 15-dimensional 95% credible regions. The credible regions are 15-dimensional rectangles formed as the product of fifteen $(0.95)^{1/15} \times 100\%$ credible intervals, one for each of $\mu^*_{01}, \ldots \mu^*_{015}$. We refer to the posterior means as $E[\boldsymbol{\mu}^*_0|\boldsymbol{y}]$. To compare the four procedures' performance in estimating $\boldsymbol{\mu}^*_0$ , we use empirical coverage ratios and credible region volume along with two metrics related to the frequentist bias and MSE. As a type of total absolute bias the following was computed for each procedure

$$\text{bias} = \frac{1}{D}\sum_{d=1}^{D}\left(\sum_{k=1}^{15}|E[\boldsymbol{\mu}^*_0|\boldsymbol{y}]_{dk} - \boldsymbol{\mu}^*_{0k}|\right).$$

Here, $E[\boldsymbol{\mu}^*_0|\boldsymbol{y}]_{dk}$ is the $k$th element of the $E[\boldsymbol{\mu}^*_0|\boldsymbol{y}]$ and $d$ is an index for the $D = 500$ data sets that were generated. Values in the MSE column were computed using

$$\text{MSE} = \frac{1}{D}\sum_{d=1}^{D}\left(\sum_{k=1}^{15}Var[\boldsymbol{\mu}^*_0|\boldsymbol{y}]_{dk} + (E[\boldsymbol{\mu}^*_0|\boldsymbol{y}]_{dk} - \boldsymbol{\mu}^*_{0k})^2\right).$$

This might be thought of as a type of total MSE averaged over the 500 data sets. Results can be found in Table 1.

When no outliers are present and the data are generated using the MVN model, results from

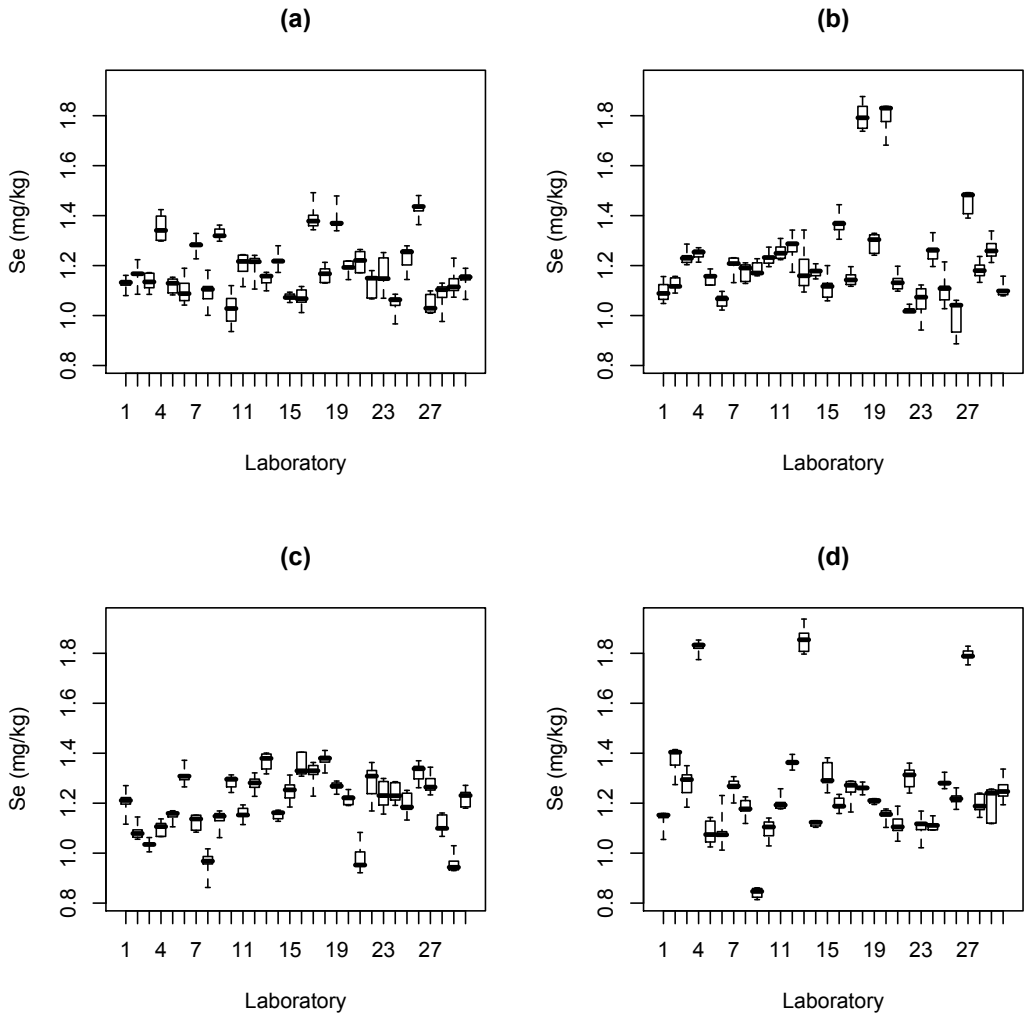


Figure 3: Boxplots of concentrations of Se from a randomly generated data set used in the simulation study. Plots (a) and (c) contain no outliers while plots (b) and (d) do. The data in plots (a) and (b) are generated using the MVN model and those in plots (c) and (d) are generated using MVT. In plot (b) two of the ten outlier labs have outlying Se values. Three of the ten outlying labs has outlying Se values in plot (d).

Table 1: Results from the simulation study. Entries correspond to averages computed across the 500 generated data sets.

| True Model | | procedure | coverage | volume $\times 10^{11}$ | bias | MSE |
|---|---|---|---|---|---|---|
| MVN | No Outliers | MVN model | 0.902 | 0.001 | 0.302 | 0.034 |
| | | MVT model | 0.908 | 0.002 | 0.318 | 0.038 |
| | | DP model | 0.916 | 0.003 | 0.316 | 0.038 |
| | | LC model | 0.890 | 0.001 | 0.311 | 0.035 |
| | 10 Outliers | MVN model | 0.890 | 0.647 | 0.799 | 0.156 |
| | | MVT model | 0.944 | 0.026 | 0.410 | 0.065 |
| | | DP model | 0.922 | 1.410 | 0.722 | 0.145 |
| | | LC model | 0.920 | 0.001 | 0.312 | 0.038 |
| MVT | No Outliers | MVN model | 0.926 | 1.500 | 0.406 | 0.065 |
| | | MVT model | 0.946 | 0.005 | 0.318 | 0.042 |
| | | DP model | 0.912 | 0.224 | 0.410 | 0.067 |
| | | LC model | 0.906 | 0.007 | 0.338 | 0.045 |
| | 10 Outliers | MVN model | 0.856 | 31.41 | 0.855 | 0.194 |
| | | MVT model | 0.960 | 0.159 | 0.443 | 0.077 |
| | | DP model | 0.886 | 19.29 | 0.796 | 0.188 |
| | | LC model | 0.932 | 0.039 | 0.375 | 0.056 |

the LC procedure are very much comparable to the other four other procedures in terms of coverage, volume, bias, and MSE. The fact that these intervals didn't attain the nominal 95% coverage is to be expected as we are working with Bayesian posterior probability intervals which have a different interpretation than confidence intervals in a finite sample setting. When outliers are introduced, the MVN and DP models are more negatively affected than the MVT and LC models and the LC model performs much better in all four metrics. When the data are generated using MVT and there are no outliers the MVT model performs the best out of the four procedures in terms of bias and MSE. This is to be expected. However, when outliers are introduced, the LC procedure has a lower MSE and bias than the MVT. It appears that the LC procedure has a clear advantage over the other three procedures at estimating $\boldsymbol{\mu}_0^*$ when outlying laboratories are present. In other simulation results (not shown) the advantages of using the LC model (in terms of smaller bias and MSE) compared to the MVT model become more obvious as the ratio of outlying labs to participating labs increases or as the distance between an outlying lab mean and the majority component mean increases.

Table 2: Results from the simulation study with regards to the accuracy in which the LC model allocates elements to the contamination and the number of occupied clusters of the contamination

| | Procedure | clusters occupied | outlying elements | outlying labs |
|---|---|---|---|---|
| MVN | No Outliers | 2 | 0 | 0 |
| | 10 Outliers | 3 | 29 | 10 |
| MVT | No Outliers | 3 | 10 | 4 |
| | 10 Outliers | 4 | 35 | 12 |

To demonstrate the utility of a contamination model with more than one cluster, we report in Table 2 several summaries of the outlying clusters identified by the LC model. At each iteration of the MCMC algorithm we count the number of clusters that contained an ele-

ment that had been allocated to the contamination. The 'clusters occupied' column reports the median of this value across MCMC iterates and the 500 data sets. For data generated using the MVN with outliers, approximately 3 clusters of the contamination were occupied while 4 were occupied for data generated from MVT with outliers. We also enumerated the total number of elements that were classified as outliers (using $E(\gamma_{ik}|\boldsymbol{y}) < 0.1$ as the outlier criterion). The values found in the column 'outlying elements' is the median number of elements allocated to the contamination across the 500 data sets. Recall that 29 elements were randomly selected to be outliers for data sets that contained outliers. The median number of labs that had at least one outlying element is also consistent with the true value of 10 or 0 labs (as is seen in the 'outlying labs' column). Table 2 indicates that the LC model is fairly accurate in its allocation of outlying elements.

# 5 Data Analysis of the NIST/NOAA Sponsored Inter-Laboratory Study

In this section we briefly describe the data that was produced from the marine mammal inter-laboratory experiment, provide results of the analysis from the LC model along with the three other methods outlined in the previous section, and assess model fit using cross validation.

## 5.1 Description of NIST/NOAA inter-laboratory study Data

In 2005 a NIST/NOAA sponsored inter-laboratory study was conducted to improve the quality of trace element measurements in marine environmental systems. The NIST prepared fresh-frozen marine mammal control materials (white-sided dolphin liver homogenate

(QC04LH4)). A glass jar containing approximately 8-10 grams of the frozen material was distributed to 33 participating laboratories. Each lab was asked to keep the material in an environment that would preserve its authenticity and to divide the material into five aliquots. Measurements of 15 trace elements (Ag, As, Cd, Co, Cs, Cu, Fe, Hg, Mn, Mo, Rb, Se, Sn, V and Zn) using in-house analytical techniques were to be taken on each aliquot. The raw measurement results were submitted to the NIST. Figure 1 provides a graphical display of measurement results for As and Se. Additional details are provided in Christopher et al. (2007).

Some labs didn't measure all 15 trace element concentrations on each aliqout. We imputed values for the missing data under the missing at random (MAR) assumption (Gelman et al. (2004)) within our proposed MCMC algorithm. The core assumption of MAR is that the missingness mechanism does not depend on the missing data. Information regarding the rationale behind a labs decision not to measure the concentration of all 15 trace elements on each aliqout is not available. However, there is no obvious indicator of a MAR assumption violation and it seems completely plausible that the missingness mechanism didn't depend on the trace element.

Algorithms to impute missing values and update posterior draws for the four procedures were written in the C programming language. For each procedure 10000 posterior draws were collected after a burn in of 20000 and thinning of 5. With regards to computation time, the computer code associated with the two simpler models required slightly less time (MVN, 584 seconds and MVT, 602 seconds) to run than that of the more complicated models (DP, 980 seconds and LC, 818 seconds). The added flexibility afforded by the LC model comes at a minimal computational cost.

## 5.2 Results From the Marine Mammal Data Analysis

Prior to analysis the data were normalized and log transformed. After transforming back to the original scale, marginal posterior means and 95% credible intervals were calculated for each trace element and are provided in Table 3.

Table 3: Posterior means and 95% credible intervals for the fifteen trace elements measured in the marine mammal inter-laboratory study using the four procedures.

| | MVN | | MVT | | DP | | LC | |
|---|---|---|---|---|---|---|---|---|
| | Mean | 95%CI | Mean | 95%CI | Mean | 95%CI | Mean | 95%CI |
| Ag | 0.50 | (0.42, 0.60) | 0.48 | (0.43, 0.54) | 0.52 | (0.38, 0.66) | 0.47 | (0.43, 0.50) |
| As | 0.30 | (0.26, 0.34) | 0.28 | (0.26, 0.30) | 0.30 | (0.25, 0.35) | 0.29 | (0.26, 0.31) |
| Cd | 0.25 | (0.20, 0.30) | 0.22 | (0.20, 0.23) | 0.26 | (0.18, 0.36) | 0.22 | (0.20, 0.23) |
| Co | 0.03 | (0.00, 0.06) | 0.02 | (0.01, 0.03) | 0.03 | (0.00, 0.08) | 0.01 | (0.01, 0.02) |
| Cs | 0.03 | (0.02, 0.05) | 0.03 | (0.02, 0.04) | 0.03 | (0.02, 0.05) | 0.03 | (0.03, 0.03) |
| Cu | 5.52 | (4.84, 6.16) | 5.26 | (4.99, 5.53) | 5.56 | (4.60, 6.48) | 5.25 | (5.04, 5.45) |
| Fe | 366.2 | (333.7, 398.9) | 355.3 | (344.1, 367.6) | 360.6 | (325.3, 398.0) | 354.7 | (344.7, 364.1) |
| Hg | 4.00 | (3.42, 4.62) | 3.67 | (3.42, 3.95) | 4.05 | (3.08, 5.03) | 3.62 | (3.43, 3.82) |
| Mn | 3.20 | (2.86, 3.55) | 3.10 | (2.97, 3.23) | 3.07 | (2.69, 3.45) | 3.13 | (3.03, 3.23) |
| Mo | 0.64 | (0.25, 1.09) | 0.42 | (0.32, 0.52) | 0.41 | (0.17, 0.70) | 0.39 | (0.32, 0.47) |
| Rb | 1.25 | (1.14, 1.36) | 1.18 | (1.13, 1.24) | 1.23 | (1.09, 1.39) | 1.19 | (1.14, 1.24) |
| Se | 3.58 | (3.13, 4.03) | 3.32 | (3.09, 3.54) | 3.68 | (3.07, 4.33) | 3.32 | (3.09, 3.57) |
| Sn | 0.12 | (0.04, 0.21) | 0.08 | (0.05, 0.11) | 0.15 | (0.03, 0.30) | 0.09 | (0.05, 0.13) |
| V | 0.05 | (0.04, 0.06 | 0.04 | (0.04, 0.05) | 0.05 | (0.04, 0.07) | 0.05 | (0.04, 0.06) |
| Zn | 29.49 | (24.26, 35.50) | 30.90 | (29.04, 32.77) | 30.38 | (24.89, 36.07) | 30.92 | (29.69, 32.17) |

From Table 3 it can be seen that the 95% credible interval widths are generally the smallest under the LC model. In fact, the average credible interval width is smaller for the LC (1.58) compared to the DP model (6.10), the MVT model (1.97) and the MVN model (5.49). Notice that in terms of estimating the location of $\mu_{0k}^*$, the LC and MVT procedures are very similar. The advantage that the LC procedure has over the MVT is mainly in the estimation of the variability associated with $\mu_{0k}^*$.

We set $L = 10$ when fitting the LC model to the marine mammal data set . Table 4 gives the posterior probability distribution on the number of occupied contamination clusters. Five

occupied contamination clusters turned out to have the highest posterior probability.

Table 4: Fraction of MCMC iterates for which the number of contamination clusters were occupied

| | Number of occupied contamination clusters | | | | | | |
|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Posterior Probability | 0.0509 | 0.2717 | 0.3888 | 0.2253 | 0.0564 | 0.0065 | 0.0004 |

A nice characteristic of the LC model is that it is possible to estimate the posterior probability of the $i$th lab being part of the majority. This can be done by computing $E(\pi_i|\boldsymbol{y})$, the posterior mean of $\pi_i$. In addition, $E(\gamma_{ik}|\boldsymbol{y})$, the posterior mean of $\gamma_{ik}$, provides the posterior probability that the $k$th element corresponding to the $i$th lab is part of the majority. These probabilities are listed in Table 5. The number of trace elements for each lab such that $E(\gamma_{ik}|\boldsymbol{y}) < 0.1$, $E(\gamma_{ik}|\boldsymbol{y}) < 0.5$, and $E(\gamma_{ik}|\boldsymbol{y}) < 0.9$ are found in parenthesis. Lab 18 reported measurements for multiple trace elements that might be considered to be far from the majority and this is reflected in its posterior probability of being an outlier. With regards to the trace elements referenced in Section 1, the posterior probability of As belonging to the majority component for lab 17 is 0.87, for lab 18 is 0.00, and for lab 26 is 0.05, while the posterior probability of Se belonging to the majority component for lab 17 is 0.025, for lab 18 is 0.00, and for lab 26 is 0.80.

We assessed model fit using cross-validation. A testing partition of the marine mammal data was created by removing 2 randomly selected observations from 10 randomly selected laboratories. The MVN, MVT, DP, and LC models were fit to the remaining portion of the marine mammal data set and the removed observations were imputed within the MCMC algorithm.

Table 5: The posterior probability of each laboratory being part of the majority component ($E(\pi_i|\boldsymbol{y})$). The values in parenthesis are the number of trace elements for each lab whose posterior means of $\gamma_{ik}$ are less than 0.1, 0.5, and 0.9. The posterior mean of $\gamma_{ik}$ is the probability that the $k$th element corresponding to the $i$th lab is part of the majority component.

| Lab | $E(\pi_i\|\boldsymbol{y})$ | Lab | $E(\pi_i\|\boldsymbol{y})$ | Lab | $E(\pi_i\|\boldsymbol{y})$ | Lab | $E(\pi_i\|\boldsymbol{y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.98 (0,0,0) | 11 | 0.68 (4,5,14) | 21 | 0.97 (0,0,0) | 31 | 0.97 (0,0,1) |
| 2 | 0.98 (0,0,0) | 12 | 0.92 (1,1,2) | 22 | 0.98 (0,0,0) | 32 | 0.92 (1,1,2) |
| 3 | 0.98 (0,0,0) | 13 | 0.97 (0,0,0) | 23 | 0.90 (1,2,3) | 33 | 0.94 (0,0,1) |
| 4 | 0.97 (0,0,0) | 14 | 0.97 (0,0,0) | 24 | 0.98 (0,0,0) | | |
| 5 | 0.96 (0,0,0) | 15 | 0.87 (1,1,6) | 25 | 0.98 (0,0,0) | | |
| 6 | 0.97 (0,0,0) | 16 | 0.97 (0,0,0) | 26 | 0.82 (1,2,13) | | |
| 7 | 0.97 (0,0,0) | 17 | 0.89 (1,2,4) | 27 | 0.96 (0,0,1) | | |
| 8 | 0.97 (0,0,0) | 18 | 0.46 (12,13,14) | 28 | 0.75 (3,5,9) | | |
| 9 | 0.97 (0,0,0) | 19 | 0.91 (1,1,3) | 29 | 0.98 (0,0,0) | | |
| 10 | 0.97 (0,0,0) | 20 | 0.97 (0,0,0) | 30 | 0.92 (1,1,1) | | |

Typically the goal in using out of sample prediction to assess model fit is to determine how concentrated the posterior distribution for the predictive values is around the truth. Often this is done by computing the mean squared prediction error (MSPE). However, the MSPE only provides a measure of how close a point estimate of the predicted value is to the true value. That is, it doesn't consider in any way how concentrated the posterior distribution is around the truth. In addition, when outliers are present it is not completely clear that the mpse is a good metric to assess out of sample prediction. Since outliers make up a small minority of the observations, overly-simple models that under-estimate the true uncertainty may do well in terms of MSPE as the bias introduced by the outliers is dominated by the smaller variance. Because of this we propose the following measure of how concentrated the

posterior distribution for a predicted value is to the truth:

$$\frac{1}{p}\sum_{k=1}^{p}\left[\frac{1}{T}\sum_{t=1}^{T}\left(y_{pred_{tk}}-y_{test_k}\right)^2\right]. \tag{13}$$

Here $y_{pred_{tk}}$ is the $t^{\text{th}}$ MCMC iterate of the $k$th entry of the predicted vector and $y_{test_k}$ is the the $k$th entry of a data vector from the testing data set. This metric incorporates each MCMC iterate to assess concentration around the truth. The value of equation (12) averaged over the twenty predictions under the LC model (0.062) was slightly smaller than that of the other three models (MVN (0.064), MVT (0.064), and DP (1.130)). The average prediction interval width was also slightly smaller for the LC model (1.59) compared to the other three (MVN (1.63), MVT (1.60), and DP (5.18).

## 5.3 Univariate Analyses

A very simple approach to the analysis of these type of data is to perform an independent analysis for each trace element. Indeed this was the approach the NIST originally used to establish a reference value (Christopher et al. (2007)). Here we compare the performance of the LC model in establishing a reference value to that obtained by performing 15 independent univariate analyses. There are any number of approaches one might take to conduct a univariate analysis that would be robust (accommodating) to the presence of outliers. None of these approaches are exactly comparable/analogous to the LC model. Because of this we

use the following univariate model that accommodates univariate outliers very flexibly.

$$\begin{aligned} y_{ij} &\sim N(\mu_i, \sigma_i^2) \\ (\mu_i, \sigma_i^2) &\sim \mathcal{P} \\ \mathcal{P} &\sim DP(\alpha, P_0) \text{ with } P_0 = N(\mu_0^*, \sigma_0^{2*}) \end{aligned}$$

In addition a normal-inv-Gamma prior was used for $(\mu_0^*, \sigma_0^{2*})$ and a inv-Gamma prior was used for $\sigma_i^2$. Hyperprior values equaling those used for the LC model were used. Using a Dirichlet process prior to model the density of $(\mu_i, \sigma_i^2)$ provides a great deal of flexibility and hopefully this results in the ability to make reasonable comparisons between the LC model and 15 independent univariate analyses. The above model was fit to each of the 15 trace elements found in the marine mammal data set. Posterior means and 95% credible intervals associated with $\boldsymbol{\mu}_0^*$ were computed for each trace element and are listed in Table 6.

Apart from the fact that performing separate univariate analyses effectively ignores the dependence structure and therefore can produce misleading results, there appear to be some pragmatic gains to using the LC procedure. The credible interval widths associated with the LC model are shorter compared to those coming from the 15 independent analyses. Obviously shorter credible intervals are desirable only if point estimates corresponding to the intervals are near the truth. We argue that this is indeed the case in the present setting using the results from the simulation study (the bias associated with estimates using the LC procedure was very small in the presence of outliers) and the fact that the LC model and 15 independent univariate analyses produce point estimates that are fairly similar. In the cases for which the point estimates from the two procedures differ (e.g. Se), those associated with the 15 independent univariate analyses are greater than those from LC procedure. This could be an indication that results from the univariate analyses are still influenced by outlying labs

as outlying labs produced measurements greater than the majority in the marine mammal data set.

Table 6: Posterior means and 95% credible intervals for the fifteen trace elements measured in the marine mammal inter-laboratory study using 15 independent univariate analyses, $t$-residual LC model and the normal residual LC model

| | Univariate | | $t$-residual LC Model | | LC Model | |
|---|---|---|---|---|---|---|
| | Mean | 95%CI | Mean | 95%CI | Mean | 95%CI |
| Ag | 0.48 | (0.45, 0.53) | 0.46 | (0.44 , 0.48) | 0.47 | (0.43, 0.50) |
| As | 0.34 | (0.26, 0.43) | 0.29 | (0.27 , 0.31) | 0.29 | (0.26, 0.31) |
| Cd | 0.26 | (0.21, 0.30) | 0.22 | (0.21 , 0.23) | 0.22 | (0.20, 0.23) |
| Co | 0.02 | (0.01, 0.04) | 0.01 | (0.01 , 0.02) | 0.01 | (0.01, 0.02) |
| Cs | 0.03 | (0.02, 0.05) | 0.03 | (0.03 , 0.03) | 0.03 | (0.03, 0.03) |
| Cu | 5.44 | (4.89, 6.35) | 5.18 | (5.01 , 5.35) | 5.25 | (5.04, 5.45) |
| Fe | 355.8 | (333.8, 378.7) | 354.8 | (344.7 , 364.3) | 354.6 | (344.7, 364.1) |
| Hg | 3.71 | (3.36, 4.14) | 3.61 | (3.44 , 3.79) | 3.62 | (3.43, 3.82) |
| Mn | 3.16 | (2.94, 3.39) | 3.10 | (3.01 , 3.19) | 3.13 | (3.03, 3.23) |
| Mo | 0.41 | (0.38, 0.45) | 0.39 | (0.34 , 0.45) | 0.39 | (0.32, 0.47) |
| Rb | 1.35 | (1.20, 1.53) | 1.18 | (1.12 , 1.22) | 1.19 | (1.14, 1.24) |
| Se | 3.99 | (3.06, 5.05) | 3.35 | (3.09 , 3.60) | 3.32 | (3.09, 3.57) |
| Sn | 0.11 | (0.05, 0.19) | 0.07 | (0.05 , 0.09) | 0.09 | (0.05, 0.13) |
| V | 0.05 | (0.04, 0.07) | 0.05 | (0.04 , 0.05) | 0.05 | (0.04, 0.06) |
| Zn | 30.03 | (27.84, 32.32) | 30.96 | (29.76 , 32.14) | 30.92 | (29.69, 32.17) |

## 5.4 Using a $t$-distribution Residual to Accommodate Within Lab Outliers

In addition to outlying labs, it is possible for within-lab outliers to exist. These types of outliers can easily be accommodated by changing (1) to

$$\boldsymbol{y}_{ij} \overset{ind}{\sim} t_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \kappa)$$

where $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \kappa)$ denotes a shifted ($\boldsymbol{\mu}$) and scaled ($\boldsymbol{\Sigma}$) $p$-dimensional $t$-distribution with $\kappa$ degrees of freedom. Conceptually this is a very straightforward modification of the LC model. Computationally, we lose conjugacy if the $t$ density is used directly. However, the $t$-distribution can be represented as a scale mixture of normals and this characterization of the $t$ preserves full conjugacy. Thus, a $t$-density residual can be obtained with the following hierarchy

$$\begin{aligned}\boldsymbol{y}_{ij} &\overset{ind}{\sim} \mathrm{N}_p(\boldsymbol{\mu}_i, \lambda_{ij}\boldsymbol{\Sigma}_i)\\ \lambda_{ij} &\overset{iid}{\sim} \mathrm{IG}(\kappa/2, 2/\kappa).\end{aligned}$$

With this representation of the $t$-distribution (2) and (3) change in the following ways

$$\begin{aligned}[\boldsymbol{\mu}_i|-] &\sim \mathrm{N}_p\left(\left[\Sigma_i^{-1}\sum_{j=1}^{n_i}\lambda_{ij}^{-1} + \boldsymbol{T}_{\boldsymbol{\gamma}_i}^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n_i}\lambda_{ij}^{-1}\boldsymbol{y}_{ij} + \boldsymbol{T}_{\boldsymbol{\gamma}_i}^{-1}\boldsymbol{\delta}_{\boldsymbol{\gamma}_i}\right], \left[\boldsymbol{\Sigma}_i^{-1}\sum_{j=1}^{n_i}\lambda_{ij}^{-1} + \boldsymbol{T}_{\boldsymbol{\gamma}_i}^{-1}\right]^{-1}\right),\\ [\boldsymbol{\Sigma}_i|-] &\sim \mathrm{IW}\left(n_i + \upsilon, \sum_{j=1}^{n_i}\lambda_{ij}^{-1}\boldsymbol{y}_{ij}\mathrm{diag}(\boldsymbol{\psi}_i)^{-1}(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i)(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i)'\mathrm{diag}(\boldsymbol{\psi}_i)^{-1} + \boldsymbol{A}\right).\end{aligned}$$

In addition, the full conditional for $\lambda_{ij}$ is

$$[\lambda_{ij}|-] \sim \mathrm{IG}\left(\frac{1}{2}(\kappa + p), \left(\frac{1}{2}[(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_{ij} - \boldsymbol{\mu}_i) + \kappa]\right)^{-1}\right)$$

One can assign a prior to $\kappa$ but for simplicity we set $\kappa = 4$. Using the same prior values for the $t$-residual LC model as those used for the Gaussian residual LC model, we fit the $t$-residual LC model to the marine mammal data set. Posterior means and 95% credible intervals for $\mu_{0k}^*$, $k = 1, \ldots, p$ are listed in Table 6. The results are fairly comparable to a Gaussian LC model. The average credible interval length across the 15 trace elements turned out to be 1.58 with all 15 point estimates being very similar to those from the Gaussian residual LC model. Using $E(\gamma_{ik}|\boldsymbol{y}) < 0.1$ as a criteria to classify the $k$th element of the $i$th lab as an

outlier, there were 29 outlying elements (compared to 27 using the Gaussian residual) and 11 labs with at least one outlying element (compared to 11 using the Gaussian residual).

# 6 Conclusion

We have developed a methodolgy that does very well in accommodating multivariate outliers in a multi-level/hierarchical modeling framework. Although the presentation of the model is necessarily a bit notation heavy, the fundamental idea to handling multivariate outliers (locally allocating multivariate vector entries to a contamination) could not be more natural and is quite simple and intuitive. In addition to being robust to the presence of outliers, the methodology provides probabilistic inference on lab/element outlier classification. This type of information should be of interest to practitioners. Also, the methodology incorporates the uncertainty associated with lab/element outlier classification in parameter estimation and inference. These nice features are available at a minimal computational cost, as a straightforward Gibbs sampler is all that is required to implement the methodology. Computer code is available from the first author by request.

# Appendices

## A Derivation of full conditionals

Here we only include details regarding $[\boldsymbol{\mu}_0^*|-]$ and $[\sigma_0^{2*}|-]$ as the derivations of $[\boldsymbol{\mu}_h^*|-]$ and $[\sigma_h^{2*}|-]$ follow similar arguments. We begin with $[\boldsymbol{\mu}_0^*|-]$.

$$
\begin{aligned}
[\boldsymbol{\mu}_0^*|-] &\propto \prod_{i=1}^{m} \boldsymbol{\phi}(\boldsymbol{\mu}_i; \boldsymbol{\delta}_{\gamma_i}, \boldsymbol{T}_{\gamma_i}) \boldsymbol{\phi}(\boldsymbol{\delta}_{\gamma_i}; \boldsymbol{m}_0, \boldsymbol{S}_0) \\
&\propto \prod_{i=1}^{m} \exp\{-0.5(\boldsymbol{\mu}_i - \boldsymbol{\delta}_{\gamma_i})' \boldsymbol{T}_{\gamma_i}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\delta}_{\gamma_i}) - 0.5(\boldsymbol{\delta}_{\gamma_i} - \boldsymbol{m}_0)' \boldsymbol{S}_0^{-1} (\boldsymbol{\delta}_{\gamma_i} - \boldsymbol{m}_0)\} \\
&\propto \exp\big\{ -0.5\big( \sum_{i=1}^{m} \big[ -\boldsymbol{\mu}_i' \boldsymbol{T}_{\gamma_i}^{-1} (\boldsymbol{\mu}_0^* \otimes \boldsymbol{\gamma}_i + \boldsymbol{\mu}_{S_i}^* \otimes (1 - \boldsymbol{\gamma}_i)) - \\
&\qquad (\boldsymbol{\mu}_0^* \otimes \boldsymbol{\gamma}_i + \boldsymbol{\mu}_{S_i}^* \otimes (1 - \boldsymbol{\gamma}_i))' \boldsymbol{T}_{\gamma_i}^{-1} \boldsymbol{\mu}_i' + \\
&\qquad (\boldsymbol{\mu}_0^* \otimes \boldsymbol{\gamma}_i + \boldsymbol{\mu}_{S_i}^* \otimes (1 - \boldsymbol{\gamma}_i))' \boldsymbol{T}_{\gamma_i}^{-1} (\boldsymbol{\mu}_0^* \otimes \boldsymbol{\gamma}_i + \boldsymbol{\mu}_{S_i}^* \otimes (1 - \boldsymbol{\gamma}_i)) \big] + \\
&\qquad \boldsymbol{\mu}_0^{*'} \boldsymbol{S}_0^{-1} \boldsymbol{\mu}_0^* - \boldsymbol{\mu}_0^{*'} \boldsymbol{S}_0^{-1} \boldsymbol{m}_0 - \boldsymbol{m}_0' \boldsymbol{S}_0^{-1} \boldsymbol{\mu}_0^* \big) \big\} \\
&\propto \exp\big\{ -0.5\big( \boldsymbol{\mu}_0^{*'} \big[ \sum_{i=1}^{m} \boldsymbol{T}_{\gamma_i=0}^{-1} + \boldsymbol{S}_0^{-1} \big] \boldsymbol{\mu}_0^* - \boldsymbol{\mu}_0^{*'} \big[ \sum_{i=1}^{m} \boldsymbol{T}_{\gamma_i=0}^{-1} \boldsymbol{\mu}_i + \boldsymbol{S}_0^{-1} \boldsymbol{m}_0 \big] - \\
&\qquad \big[ \sum_{i=1}^{m} \boldsymbol{\mu}_i' \boldsymbol{T}_{\gamma_i=0}^{-1} + \boldsymbol{m}_0' \boldsymbol{S}_0^{-1} \big] \boldsymbol{\mu}_0^* \big) \big\} \\
&\sim \mathrm{N}_p \left( \left[ \sum_{i=1}^{m} \boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}^{-1} + \boldsymbol{S}_0^{-1} \right]^{-1} \left[ \sum_{i=1}^{m} \boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}^{-1} \boldsymbol{\mu}_i + \boldsymbol{S}_0^{-1} \boldsymbol{m}_0 \right], \left[ \sum_{i=1}^{m} \boldsymbol{T}_{0:\boldsymbol{\gamma}_i=0}^{-1} + \boldsymbol{S}_0^{-1} \right]^{-1} \right).
\end{aligned}
$$

Now $[\sigma_0^{2*}|-]$ can be obtained by

$$
\begin{aligned}
[\sigma_{0k}^{2^*}|-] &\propto \prod_{i=1}^{m} \boldsymbol{\phi}(\boldsymbol{\mu}_i; \boldsymbol{\delta}_{\gamma_i}, \boldsymbol{T}_{\gamma_i}) IG(\sigma_{0k}^{2*}; a_{\sigma_k}, b_{\sigma_k}) \\
&\propto (\sigma_{0k}^{2*})^{-0.5\sum_{i=1}^{m}\gamma_i - a_{\sigma_k} - 1} \exp\{-0.5 \sum_{i=1}^{m} \gamma_{ik}(1/\sigma_{0k}^{2*})(\mu_{ik} - \mu_{0k}^{*})^2 - 1/(b_{\sigma_k}\sigma_{0k}^{2*})\} \\
&\sim \text{IG}\left( a_{\sigma_k} + \frac{1}{2}\sum_{i=1}^{m}\gamma_{ik}, \left[ \frac{1}{b_{\sigma_k}} + \frac{1}{2}\sum_{i=1}^{m}\gamma_{ik}(\mu_{ik} - \mu_{0k}^{*})^2 \right]^{-1} \right).
\end{aligned}
$$

# References


Bayarri, M. and Morales, J. (2003), "Bayesian Measures of Surprise for Outlier Detection," *Journal of Statistical Planning and Inference*, 111, 3–22.

Bednarski, T. and Zontek, S. (1996), "Robust estimation of parameters in a mixed unbalanced model," *The Annals of Statistics*, 24, 1493–1510.

Box, G. and Tiao, G. (1968), "A Bayesian approach to some outlier problems," *Biometrika*, 55, 119–129.

Christopher, S. J., Pugh, R. S., Ellisor, M. B., Mackey, E. A., Spatz, R. O., Porter, B. J., Bealer, K. J., Kucklick, J. R., Rowles, T. K., and Becker, P. R. (2007), "Description and results of the NIST/NOAA 2005 Interlaboratory Comparison Exercise for Trace Elements in Marine Mammals," *Accreditation and Quality Assurance*, 12, 175–187.

Dunson, D. B. (2009), "Nonparametric Bayes Local Partition Models for Random Effects," *Biometrika*, 96, 249–262.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, New York: Chapman and Hall/ CRC, 2nd ed.

Gleser, L. J. (1998), "Assessing Uncertainty in Measurement," *Statistical Science*, 13, 277–290.

Hoff, P. D. (2006), "Model-based Subspace Clustering," *Bayesian Analysis*, 1, 321–344.

Ishwaran, H. and Zarepour, M. (2002), "Dirichlet Prior Sieves in Finite Normal Mixtures," *Statistica Sinica*, 12, 941–963.

Lischer, P. (1996), "Robust statistical methods in interlaboratory analytical studies," in *Robust Statistics, Data Analysis, and Computer Intensive Methods*, ed. Rieder, H., New York: Springer Verlag, pp. 251–265.

Mandel, J. (1995), "Structure and Outliers in Interlaboratory Studies," *Journal of testing and evaluation*, 23, 364–369.

Muller, C. H. and Uhlig, S. (2001), "Estimation of variance components with high breakdown point and high efficiency," *Biometrika*, 88, 353–366.

Peña, D. and Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics*, 43, 286–300.

Penny, K. I. and Jolliffe, I. T. (2001), "A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Saftey Data," *The Statistician*, 50, 295–308.

Roberts, G. and Rosenthal, J. (2007), "Coupling and ergodicity of adaptive MCMC," *Journal of Applied Probability*, 44, 458–475.

Rocke, D. M. (1983), "Robust Statistical Analysis of Interlaboratory Studies," *Biometrika*, 70, 421–31.

Rukhin, A. L. (2007), "Estimating Common Vector Parameters in Interlaboratory Studies," *Journal of Multivariate Analysis*, 98, 435–454.

Rukhin, A. L. and Vangel, M. G. (1998), "Estimation of a Common Mean and Weighted Means Statistics," *Journal of the American Statistical Association*, 93, 303–308.

Song, P. X.-K., Zhang, P., and QU, A. (2007), "Maximum Likelihood Inference in Robust Linear Mixed-Effects Models Using Multivariate $t$ Distributions," *Statistica Sinica*, 17, 929–943.

Toman, B. (2007), "Bayesian Approaches to Calculating a Reference Value in Key Comparison Experiments," *Technometrics*, 29, 81–87.

Verdinelli, I. and Wasserman, L. (1991), "Bayesian analysis of outlier problems using the Gibbs sampler," *Statistics and Computing*, 1, 105–117.