A Bayesian Approach to Establishing a Reference Particle Size Distribution in the Presence of Outliers

Garritt L. Page

Pontificia Universidad Católica de Chile

Departmento de Estadística

Stephen B. Vardeman

Iowa State University

Department of Statistics

Draft

Supported by NSF grant DMS #0502347 EMSW21-RTG awarded to the Department of Statistics, Iowa State University

January 18, 2012

Abstract

The presence of observations or measurements that are unlike the majority is fairly common in studies conducted to establish particle size (or weight fraction) distributions. Therefore, there is a need to develop methods that are able to produce estimates of particle size distributions that are not overly sensitive to the presence of a few observations that might be considered outliers. This article proposes a type of contamination mixture model that probabilistically allocates each observation to either a "majority" component or a contamination component. Observations that are allocated to a contamination component are down-weighted when estimating the particle size distribution (while the uncertainty of contamination classification is automatically accounted for in estimation). Computational methods are developed and the utility of the proposed methodology is demonstrated via a simulation study. The method is then applied to data produced from an inter-laboratory study conducted to establish a particle size distribution in cement.

1 Introduction

Estimating a particle size distribution (PSD) is an important (and often necessary) exercise in a large range of disciplines. Studies in areas as diverse as environmental science (Zhang *et al.* [20]), cement composition (Ferraris *et al.* [4]), soil composition (Bah *et al.* [1]), and even mastication (Van der Bilt *et al.* [18]) are conducted with the purpose of learning about PSDs. Many methods are employed to gather data in a particle size study. The method that may be the least technical and therefore simplest to describe is that of sieving. In these studies, PSDs of specimens of a granular material are run through a set of progressively finer sieves. The fraction of specimen weight or the average size of particle that is retained at each sieve can then be used to establish a PSD. Other methods, though more complicated, still tend to focus on either a weight fraction or typical particle size at different discretized "levels." In the case of weight fraction analysis (which is the focus of this paper), "particle

size distribution" is a bit of a misnomer. It has the natural meaning of frequency distribution of size across particles. What is really under discussion is the cumulative weight fraction of the material as a function of particle size. Regardless, PSD is standard terminology in this area and we use it throughout this discussion.

Sieving is one of the five procedures that were employed in a round-robin (inter-laboratory) study conducted by an ASTM (American Society for Materials and Testing International) committee. (See Ferraris *et al.* [4] for more details.) The study was conducted to learn about cement particle size distributions and to establish a "single calibration curve that represents the average distribution for all methods inclusive." Twenty-one labs participated in the inter-laboratory study and were instructed to report a cumulative PSD measurement by employing the commonly used in-house technique (which resulted in the five measurement techniques). The results of the study can be found in Figure 1 (henceforth referred to as the NIST data set).

Each discipline has seemingly espoused its own method of estimating PSDs from data like that found in Figure 1. Some simply "average" observations at each sieve size (Ferraris *et al.* [4])). Others adopt a mathematical model (Van der Bilt *et al.* [18], Bah *et al.* [1]) and typically use some variation of nonlinear least squares to fit the model to the data. In the statistical literature, there is a surprisingly little dedicated specifically to estimating a "typical" PSD from data like those represented in Figure 1. Lwin [13] provides a method that is based on the bulk sampling, renewal process theory of Scheaffer [17]. The idea is to minimize a type of Kullback-Leibler distance between a theoretical weight size distribution and the sample weight-size distribution. Though the method provides reasonable results, estimation procedures are somewhat ad-hoc and uncertainty associated with the PSD estimate is not readily available. Lwin [14] proposes the direct use the approximate likelihood of Lwin [13] to formulate a model approach. Leyva *et al.* [11] develop maximum likelihood techniques and uncertainty estimates using the approximate likelihood of Lwin [13]. Their methodology has the virtue of being completely model-based and therefore PSDs and their

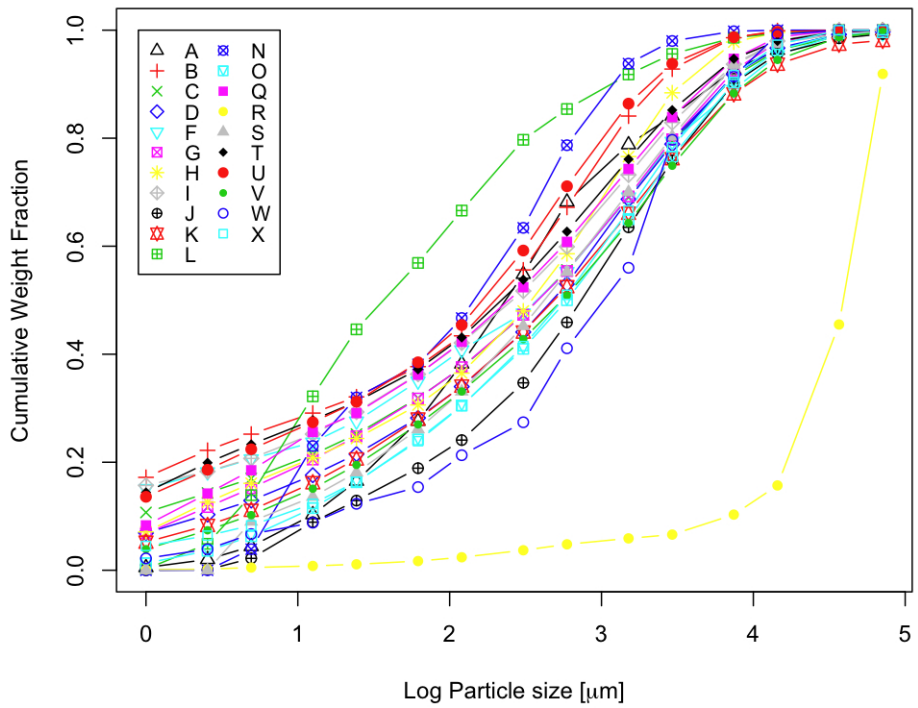

Figure 1: The raw cumulative weight fractions produced in the round-robin administered to measure the particle size distribution of cement.

uncertainty estimates are obtained in a principled and coherent way. In addition, Leyva *et al.* [10] develop a fully Bayesian methodology based on the same asymptotic multivariate normal structure, which provides sensible PSD estimates and readily available uncertainty quantification.

A common characteristic of all methods just described when applied to the NIST data set is the sensitivity of the PSD estimate to the presence of Labs R and L. It is not completely obvious how one might handle the measurements originating from these two labs. Simply discarding measurements from these two labs needs to be clearly justified (e.g., as coming in response to recording error). Since the Lwin procedures are not model-based, one would have to develop new estimators that are robust to the presence of outliers (something that doesn't appear to be straightforward). However, since the Bayesian approach outlined in

Leyva *et al.* [10] is model-based, it can be modified in direct fashion using the "robust" techniques found in Page *et al.* [15]. The robustified analysis includes all lab measurements and accommodates outliers probabilistically. Therefore, the uncertainty corresponding to outlier classification is incorporated in all estimation and prediction. The development of this robust methodology and the investigation of its properties are the focus of this paper.

Before proceeding, it is worth noting that because weight fraction data are constrained to sum to one, it might seem natural to consider methods found in the compositional data literature (see Aitchison [2] for an introduction and Filzmoser *et al.* [6] for robust compositional data methods). However, these methods ignore ordering of the size/weight classes (and implicit expectations of corresponding "smoothness" of PSD's) and are not applicable if for example different sets of size classes are used for different specimens.

The remainder of the article is organized as follows. Section 2 provides details from Leyva *et al.* [10]'s multivariate PSD modeling strategy that are necessary for developing the robust methodology. In Section 3 we develop a model that provides a robust PSD estimate and the computation necessary to carry out the analysis. Section 4 contains a simulation study comparing the robust Bayesian analysis to Leyva *et al.* [10]'s MLE approach and least squares fitting of a logistic curve. In Section 5 we apply the methodology to the NIST data set and Section 6 contains some concluding remarks.

2 Background: Models and Bayes Procedures for PSDs

In this section we summarize the parts of Leyva *et al.* [10]'s class of models that are required in the present study. First, multivariate normal likelihoods are detailed and then Bayesian extensions are given. For more background concerning what follows, the reader is referred to the paper.

2.1 Multivariate Normal Likelihoods for Particle Weight Fractions

Let particle size and weight be denoted by S and W respectively. Suppose that there are k particle size intervals $[C_{i-1}, C_i)$ for $i = 1, \ldots, k$ where $C_0 < C_1 < \cdots < C_k$. Let $p_1, p_2, \ldots, p_k$ be the corresponding specimen weight fractions. Then, from the assumption that $\log(S) \sim \mathrm{N}(\mu_s, \sigma_s^2)$ together with $\mathrm{E}[W|S = s] = \kappa s^\eta$ and $\mathrm{E}[W^2|S = s] = \kappa' s^{2\eta}$, for some constant η, Lwin [13] and Leyva *et al.* [11] argue that

$$(p_1, p_2, \ldots, p_k)' \dot{\sim} \mathrm{MVN}(\boldsymbol{\pi}, \boldsymbol{\Sigma}) \tag{1}$$

for the generation of specimens via random sampling of particles up to a fixed total weight. (Here $\dot{\sim}$ denotes "approximately distributed as.") $\boldsymbol{\pi}$ and $\boldsymbol{\Sigma}$ are functions of four parameters $(\mu_s, \sigma_s^2, \eta, \tau)$. The mean vector is

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)' \text{ for } \pi_i = \frac{\Phi\left(\frac{\log C_i - \mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_{i-1} - \mu_s^*}{\sigma_s}\right)}{\Phi\left(\frac{\log C_k - \mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_0 - \mu_s^*}{\sigma_s}\right)} \tag{2}$$

where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution and $\mu_s^* = \mu_s + \eta\sigma_s^2$. The variances and covariances in $\boldsymbol{\Sigma}$ are

$$\mathrm{Cov}(p_i, p_u) = \tau \begin{cases} \pi_i(1 - \pi_i)\gamma_i^* + \pi_i^2 \left[\sum_{\ell=1}^{k} \pi_\ell \gamma_\ell^* - \gamma_i^*\right] & \text{for } i = u \\ \pi_i \pi_u \left[\sum_{\ell=1}^{k} \pi_\ell \gamma_\ell^* - \gamma_i^* - \gamma_u^*\right] & \text{for } i \neq u \end{cases}$$

with

$$\gamma_i^* = e^{\eta(\mu_s^* + 0.5\eta\sigma_s^2)} \frac{\Phi\left(\frac{\log C_i - (\mu_s^* + \eta\sigma_s^2)}{\sigma_s}\right) - \Phi\left(\frac{\log C_{i-1} - (\mu_s^* + \eta\sigma_s^2)}{\sigma_s}\right)}{\Phi\left(\frac{\log C_i - \mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_{i-1} - \mu_s^*}{\sigma_s}\right)}. \tag{3}$$

The principle motivation for employing this method is that a potentially very-high dimensional multivariate normal distribution is parsimoniously parametrized with only four param-

eters. In addition, there is (at least in principle) an intuitive interpretation that accompanies each parameter. π_i is the fraction of log-normal mass that lies in the ith particle size interval, and the covariance pattern is vaguely similar to a multinomial covariance structure. μ_s^* and σ_s^2 are the mean and variance of the log particle sizes, τ is a scaling factor for the covariance matrix, and η can be thought of as potentially characterizing the "shape of average particles." Under the assumptions stated above, the cumulative weight fraction up to size s is

$$CW(s;\mu_s^*,\sigma_s^2) = \frac{\Phi\left(\frac{\log s-\mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_0-\mu_s^*}{\sigma_s}\right)}{\Phi\left(\frac{\log C_k-\mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_0-\mu_s^*}{\sigma_s}\right)}. \tag{4}$$

Though the approximate likelihood (1) is arrived at through the bulk sampling renewal ideas of Scheaffer [17], the fact that it is singular leads one to consider some of the transformation ideas found in the compositional data literature. We consider the additive log ratio transformation of Aitchison [2]. It has been noted that this transformation is not isometric (see Egozcue *et al.* [3]). That said, we proceed with this transformation, as all inference is done in the original space and MCMC procedures typically employed with Bayesian modeling (which we adopt) have the virtue of propagating uncertainty through transformations. Also, preliminary investigations (not shown) indicate that results are fairly insensitive to the basis used in the ratio transformation. Therefore, we fix the first particle weight fraction as a base and consider the model

$$\boldsymbol{q} = \begin{pmatrix} \log p_2 - \log p_1 \\ \log p_3 - \log p_1 \\ \vdots \\ \log p_k - \log p_1 \end{pmatrix} \dot{\sim} \text{MVN}_{k-1}(\boldsymbol{\delta}, \boldsymbol{\Delta}) \tag{5}$$

where

$$\boldsymbol{\delta} = (\delta_2, \delta_3, \ldots, \delta_k)' \text{ with } \delta_i = \log \pi_i - \log \pi_1 \quad i = 2, \ldots, k$$

and the entries of $\boldsymbol{\Delta}$ are

$$\text{Cov}(q_i, q_u) = \tau \begin{cases} \frac{1}{\pi_i}\gamma_i^* + \frac{1}{\pi_1}\gamma_1^* & \text{for } i = u \\ \frac{1}{\pi_1}\gamma_1^* & \text{for } i \neq u \end{cases}$$

(These moments follow from a "delta method" approximation based on the earlier modeling for $\boldsymbol{p}$) If we use this modeling for log ratios of weight fractions and assume that results from different labs are independent, then letting $h(\boldsymbol{q}_j|\mu_s^*, \sigma_s, \eta, \tau)$ denote the appropriate MVN_{k-1} pdf for the jth lab's results, the likelihood function based on the log weight fraction ratios is

$$L_{\boldsymbol{q}}(\mu_s^*, \sigma_s^2, \eta, \tau) = \prod_{j=1}^{L} h(\boldsymbol{q}_j|\mu_s^*, \sigma_s^2, \eta, \tau).$$

2.2 Bayesian Models

Often the parameters $\mu_s^*, \sigma_s^2, \eta, \tau$ are not of principle interest, but rather $CW(\cdot)$ (a function of μ_s^*, σ_s^2). One can use a plug-in approach to obtain a point estimate of $CW(\cdot)$, but its uncertainty and asymptotic distributional properties are more difficult to obtain. Additionally, it may be of interest to predict a PSD from a participating lab or possibly a lab whose measurements have yet to be taken. These predictions are readily available if one adopts a Bayesian modeling strategy. In light of this, using a Bayesian model may be of interest. Let $g(\mu_s^*, \sigma_s^2, \eta, \tau)$ be a joint (prior) density for the parameters $\mu_s^*, \sigma_s^2, \eta, \tau$. Then the posterior distribution for the parameters given the $\boldsymbol{q}_j$ has density

$$g(\mu_s^*, \sigma_s^2, \eta, \tau|\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_L) \propto L_{\boldsymbol{q}}(\mu_s^*, \sigma_s^2, \eta, \tau) g(\mu_s^*, \sigma_s^2, \eta, \tau).$$

The posterior distributions may be approximated using Markov Chain Monte Carlo (MCMC) techniques. Using the MCMC samples, approximate posterior distributions for parametric functions, $t(\mu_s^*, \sigma_s^2, \eta, \tau)$, are readily available (including values of $CW(\cdot)$, for example). Further, approximate posterior predictive distributions are available for an addi-

tional weight fraction vector $\boldsymbol{p}_{new}$.

3 Robust Models for PSDs

In this section we detail a method that robustifies the PSD estimate obtained from the Bayesian multivariate normal model of Leyva *et al.* [10] (and described in Section 2) to the presence of a few observed weight fraction vectors that are unlike the majority. These robust PSD estimates are obtained by incorporating mixture models like those in Page *et al.* [15]. These models can be thought of as contamination mixture models, since each observation is allocated to a "majority" component or a contamination component. Observations that are allocated to the contamination component are down-weighted when estimating a PSD.

3.1 Mixtures at the Observation Level

Modeling $\boldsymbol{q}$ with a two-component mixture requires modifying the likelihood described in the previous section. The likelihood in (5) becomes.

$$\boldsymbol{q}_j \overset{iid}{\sim} \rho \text{MVN}_{k-1}(\boldsymbol{\delta}, \boldsymbol{\Delta}) + (1-\rho)\text{MVU}(\boldsymbol{G}, \boldsymbol{H}). \tag{6}$$

where $\text{MVU}(\boldsymbol{G}, \boldsymbol{H})$ denotes a multivariate uniform distribution on a $(k-1)$-dimensional rectangle defined by lower limits in $\boldsymbol{G}$ and upper limits in $\boldsymbol{H}$ which are user-supplied and ρ is the probability that a randomly selected PSD is an outlier. As mentioned in Page *et al.* [15] one motivation for using a MVU distribution to model the contamination is to circumvent the label switching commonly met in Bayesian mixture analysis (see Jasra *et al.* [9]). We discuss our method of selecting entries of vectors $\boldsymbol{G}$ and $\boldsymbol{H}$ in Section 3.2

We assume that $\rho \sim \text{Beta}(9, 1)$. This follows the suggestion made by Verdinelli *et al.* [19] to make the probability of (a randomly selected weight fraction vector being) an outlier small (i.e., $1 - E(\rho) = 0.1$). To finish the model we employ the prior specifications used by Leyva *et al.* [10] save for σ_s^2 which we assume follows an inverse gamma distribution. The precise

assumptions we will use under the "first level" mixtures are presented below.

$$
\begin{aligned}
\boldsymbol{q}_j &\overset{iid}{\sim} \rho\text{MVN}_{k-1}(\boldsymbol{\delta}, \boldsymbol{\Delta}) + (1-\rho)\text{MVU}(\boldsymbol{G}, \boldsymbol{H}), \\
\mu_s^* &\sim \text{N}(0, 1000), \\
\sigma_s^2 &\sim \text{IG}(0.01, 0.01), \\
\eta &\sim \text{UN}(0, 3), \\
\tau &\sim \text{Exp}(0.001), \text{ and} \\
\rho &\sim \text{Beta}(9, 1).
\end{aligned}
$$

$\text{IG}(a, b)$ denotes an inverse gamma distribution with mean $b/(a-1)$. In addition, 0.001 is the mean of the exponential distribution and 1000 is the variance of the normal distribution. It was shown in Leyva *et al.* [10] that the priors employed here can be thought of as essentially "non-informative" priors. Note that we are assuming *a priori* independence between parameters (which is commonly done) but μ_s^* and σ_s^2 are highly correlated *a posteriori*. In subsequent sections we refer to this model as "partially specified contamination model" (PCM).

3.2 MCMC Algorithm with Mixtures at the Observation Level

The joint posterior distribution available from the PCM model is analytically intractable. In addition, independent random samples from this distribution are difficult if not impossible to obtain. In light of this, we use MCMC techniques to obtain (correlated) samples. As will be seen, the MCMC algorithm is a hybrid of a Metropolis within Gibbs algorithm.

To begin we consider the mixture portion of the model. To facilitate MCMC sampling in Bayesian mixtures it is common practice (Gelman *et al.* [7]) to incorporate a beta-binomial latent hierarchical structure to allocate each observation to a mixture component. That is,

the following latent classification variables

$$\zeta_j = \begin{cases} 1 & \text{if } \boldsymbol{q}_j \overset{iid}{\sim} \text{MVN}_{k-1}(\boldsymbol{\delta}, \boldsymbol{\Delta}) \\ 0 & \text{if } \boldsymbol{q}_j \overset{iid}{\sim} \text{MVU}(\boldsymbol{G}, \boldsymbol{H}) \end{cases}$$

such that $\zeta_j \overset{iid}{\sim} \text{Ber}(\rho)$ and $\rho \sim \text{Beta}(9, 1)$ are introduced for each observation. The ζ_j's could potentially be of interest, as they provide posterior probabilities of individual lab measurements being outliers. Augmenting the model with ζ_j's simplifies computation since the full conditionals of ζ_j and ρ are of recognizable form and very simple Gibbs sampler may be used to update their values in the MCMC algorithm. The four remaining parameters from the multivarate normal likelihood do not have recognizable full conditionals. Random walk Metropolis steps with a normal proposal distribution can be used to update them. An MCMC algorithm that provides samples from $h(\mu_s^*, \sigma_s^2, \eta, \tau, \boldsymbol{\zeta}, \rho | \boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_L)$ can then be constructed by updating each parameter on an individual basis using Gibbs steps for ζ_j's and ρ and Metropolis steps for the other parameters. The appendix contains derived full conditionals and details with regards to the performance of the algorithm.

The procedure developed here can be sensitive to the values selected for $\boldsymbol{G}$ and $\boldsymbol{H}$. Because of this, we suggest automating the process by letting the data be a guide. This can be done by setting the i^{th} entries of $\boldsymbol{G}$ and $\boldsymbol{H}$ to $g_i = \min_j q_{ji}$ and $h_i = \max_j q_{ji}$ respectively.

3.3 Estimating a Reference PSD

For the PCM, estimating a reference PSD is fairly straightfoward. Evaluating $CW(C_i)$ for each MCMC draw of μ_s^* and σ_s^2 provides MCMC iterates from the posterior distribution of the approximate mean cumulative weight fraction up to size C_i. Computing the empirical mean or median of these iterates provides an estimate for the mean cumulative weight fraction up to size C_i, characterizes the PSD.

4 Simulation Study

To investigate how the PCM performs in accommodating outlying empirical weight fraction vectors, we ran a small simulation study. The study consisted of generating many data sets that are similar to the NIST data set and for each estimating a reference PSD. We consider three methods of PSD estimation. The first is computing the MLEs for $\mu_s^*, \sigma_s^2, \eta, \tau$ (which we denote as $\hat{\mu}_s^*, \hat{\sigma}_s^2, \hat{\eta}, \hat{\tau}$) using the likelihood found in display (5). An estimate of $CW(C_i)$ is had by plugging $\hat{\mu}_s^*$ and $\hat{\sigma}_s^2$ into display (4). (Here and in what follows we use $\widehat{CW}(C_i)$ to denote a PSD estimates regardless of the methodology used). Secondly, we consider modeling the cumulative weight fractions directly with the commonly used logistic curve

$$f(C_i) = \frac{\exp(\beta_0 + \beta_1 C_i)}{1 + \exp(\beta_0 + \beta_1 C_i)}. \tag{7}$$

Estimation of β_0 and β_1 in (5) was carried out using robust nonlinear least squares. Lastly, a PSD estimate was obtained using the PCM model. To compare estimated PSDs from the three procedures to the "true" PSD, we used Aitchison's distance [2] and a type of total mean squared error (mse). In the subsequent two sections we detail the methods used to generate simulated data.

4.1 Using the NIST Data Set and Fitted q MVN Model to Generate Data Sets

Maximum likelihood estimates for $(\mu_s^*, \sigma_s^2, \eta, \tau)$ were computed using the $\boldsymbol{q}$ specification of a likelihood in Section 2.1 and the NIST data set (excluding labs that the NIST analysis classified as outliers). Then maximum likelihood estimates $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\Delta}}$ were obtained and used as parameters of a multivariate normal distribution from which vectors of log ratios of weight fractions were generated. To generate simulated measurements coming from outlying

laboratories, 10 was added to $\hat{\mu}_s^*$ and to $\hat{\sigma}_s^2$ prior to computing $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{\Delta}}$. Then the process just described was used to generate outlying weight fraction vectors. This method of generating vectors of particle weight fractions is referred to as the $\boldsymbol{q}$ data generating mechanism ($\boldsymbol{q}$-dgm). Since data simulated here is based on (3) the PCM and MLE may have an advantage over the logistic curve model in estimating a PSD in this data generating scenario. Figure 2 provides an example of a data set for 10 laboratories generated using the $\boldsymbol{q}$-dgm.

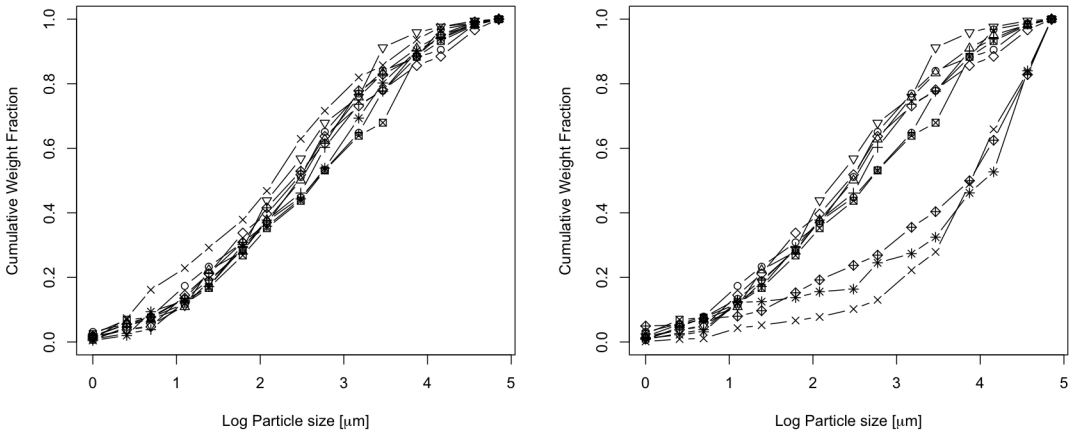

Figure 2: 10 cumulative particle weight fraction functions generated using the $\boldsymbol{q}$-dgm. The left plot is an example where no outliers are present. In the right plot three PSDs were randomly selected to be replaced by a PSD from the contamination distribution.

4.2 Using the NIST Data Set and Unrestricted MVN Models to Generate Data Sets

After discarding observations from laboratories that were considered outliers in the NIST analysis (Ferraris *et al.* [4]) and laboratories that had any weight fraction entries that were 0, an empirical mean vector ($\bar{\boldsymbol{q}}$) and covariance matrix ($\boldsymbol{S}_q$) were computed. Then $\bar{\boldsymbol{q}}$ and $\boldsymbol{S}_q$ were used as parameters of a multivariate normal distribution from which $\boldsymbol{q}$ vectors were generated. Outliers were generated in the same way except that $\bar{\boldsymbol{q}}$ and $\boldsymbol{S}_q$ were computed with observations from laboratories that were considered outliers in the NIST analysis. This

method of generating vectors of particle weight fractions is referred to as the empirical data generating mechanism (e-dgm) and should not provide an inherent advantage to any of the procedures. Figure 3 provides an example of a data set representing 10 laboratories that was generated using the e-dgm.

For both data generating methods, 200 data sets each containing 20 PSDs were generated. 100 of the data sets contained outliers and the remaining 100 were outlier free. For data sets that contained outliers, 5 PSDs were randomly selected to receive "outlier" PSDs.

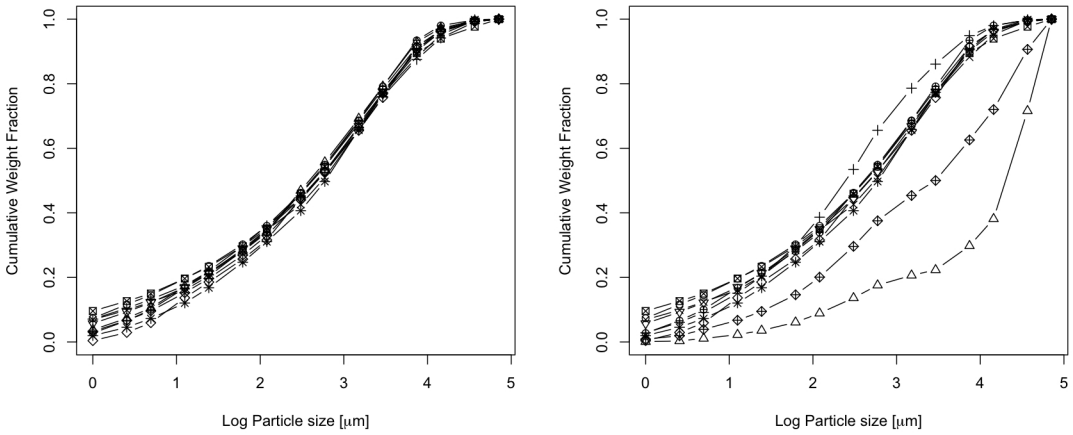

Figure 3: 10 cumulative particle weight fraction functions generated using the e-dgm. The left plot is an example where no outliers are present. In the right plot three PSDs were randomly selected to be replaced by a PSD from the contamination distribution. These three labs were given a particle weight fraction distribution simulated using a fitted "outlier" mean and covariance

4.3 Results of Simulation Study

In Table 1, the column "datgen" refers to the method that was used to generate vectors of weight fractions. The column "outlier" indicates whether measurements from outlying labs were included when generating vectors of weight fractions. To produce a type of total

mse of a procedure's PSD estimate, we computed

$$\frac{1}{D}\sum_{d=1}^{D}\left(\sum_{i=1}^{15}\left[CW(C_i)-\widehat{CW}_d(C_i)\right]^2\right). \tag{8}$$

Here $CW(C_i)$ is the i^{th} cumulative sum of the components of the mean vector that generated the data and d is an index for the $D = 100$ data sets that were generated. $\widehat{CW}_d(C_i)$ is the i^{th} component of an estimated reference cumulative PSD coming from the d^{th} data set. We use this metric, as the mse is a discrepancy statistic that incorporates both the bias and error of an estimator. This value can be found under the column header "mse."

The so-called Aitchison's distance [2] is commonly used in compositional data analysis to compute distance between two vectors that lie on the simplex and are constrained to sum to 1. We employ this distance as a means to compare how "close" the PSD estimates are to the weight fraction vector that was used to generate the data sets. This distance is

$$\frac{1}{D}\sum_{d=1}^{D}\left(\sum_{i=1}^{15}\left[\log\frac{\widehat{CW}_d(C_i)}{g(\widehat{CW}_d(\boldsymbol{C}))}-\log\frac{CW(C_i)}{g(CW(\boldsymbol{C}))}\right]\right), \tag{9}$$

where $g(\cdot)$ denotes the geometric mean function. This value was computed using the `robCompositions` [8] package found in the statistical computing software `R` [16]. The average Aitchison's distance across the 100 data sets for each estimate can be found under the header "Adist."

From Table 1 it appears that the PCM procedure performed best in terms of mse and Adist in estimating a reference PSD regardless of how the data were generated. This is particularly true when outlying weight fraction vectors were present.

Table 1: Results from the simulation study. 100 data sets were generated for each data generating scenario. The values under the header "mse" were computed using display (8) and those under the header "Adist" were computed using display (9).

		PCM		MLE		Logit	
datgen	outlier	mse	Adist	mse	Adist	mse	Adist
$\boldsymbol{q}$-dgm	No	0.001	0.142	0.004	0.222	0.003	0.458
	Yes	0.001	0.178	0.061	0.408	0.009	0.432
e-dgm	No	0.019	0.600	0.019	0.671	0.025	0.682
	Yes	0.026	0.825	0.052	2.804	0.028	1.033

5 Reference PSD from the Analysis of the NIST Data Set

When analyzing the ASTM-(committee C01.25.01)- sponsored inter-laboratory study, entries associated with particle weight ratios of "0" (which correspond to log weight ratios of $-\infty$) were imputed to be non-zero within the MCMC algorithm under the missing at random assumption (see Little *et al.* [12]). Posterior distributions from the PCM model were approximated using 3,000 MCMC iterates. These were obtained by pooling iterates from three chains with sparse (but carefully selected) starting values. Each chain was run for 100,000 iterations with the first 50,000 being discarded as burn-in and thinned by 50. The MCMC algorithm was written in the C programming language and the 100,000 MCMC iterates required 40 sec. to be collected using a desktop computer with 4 gigs of ram. Convergence was monitored graphically using MCMC iterate history plots (see the appendix). The marginal posterior distributions of $\mu_s^*, \sigma_s^2, \eta, \tau$ and ρ are provided in Figure 4. Using the MCMC iterates and methods discussed in Section 3 a reference cumulative PSD was estimated along with pointwise 95% posterior credible bands. These along with an MLE estimate and one obtained using robust nonlinear least squares to fit a logistic curve are provided in Table 2 and graphically in Figure 5.

From the Figure 4 it appears that *a posteriori* ρ lies somewhere between 0.65 and 0.95

Table 2: PSD estimates from the three procedures.

	PCM		MLE	Logit
Particle Size (μm)	Posterior Mean	95% Credible Bands	Estimate	Estimate
[0, 1)	0.014	(0.006, 0.024)	0.024	0.039
[1, 1.5)	0.034	(0.018, 0.051)	0.051	0.065
[1.5, 2)	0.056	(0.036, 0.082)	0.081	0.933
[2, 3)	0.109	(0.079, 0.144)	0.142	0.150
[3, 4)	0.164	(0.128, 0.204)	0.201	0.205
[4, 6)	0.266	(0.227, 0.313)	0.307	0.308
[6, 8)	0.356	(0.315, 0.400)	0.394	0.395
[8, 12)	0.496	(0.462, 0.535)	0.527	0.528
[12, 16)	0.598	(0.567, 0.629)	0.621	0.622
[16, 24)	0.731	(0.707, 0.754)	0.743	0.738
[24, 32)	0.812	(0.793, 0.828)	0.816	0.806
[32, 48)	0.899	(0.887, 0.912)	0.899	0.876
[48, 64)	0.943	(0.934, 0.952)	0.941	0.912
[64, 96)	0.983	(0.979, 0.986)	0.982	0.947
[96, 128)	1.000		1.000	0.963

(which is the 95% credible interval). Thus the probability that a randomly selected weight fraction is part of the majority is approximately 80%. In addition, the mean and variance of the log particles sizes (with 95% credible intervals) are availably by computing $E(\mu_s^*|\boldsymbol{p})$ and $E(\sigma_s^2|\boldsymbol{p})$ which turn out to be 2.52 (2.41, 2.62) and 1.32 (1.12, 1.57) repectively

Looking at Figure 5 and Table 2, it appears as if the the three PSD estimates are similar. All three procedures appear to have a slight lack-of-fit in the lower tail. The PSD estimate from the PCM model and that of the MLE are very similar in the upper tail with the estimate from the logistic model showing more lack-of-fit. Because all three estimates are derived from parametric models that are very parsimonious a slight restriction in flexibility (resulting in some lack-of-fit) should be expected. It does appear that both the MLE and LOGIT estimates are influenced by the two observations that lie above the majority.

It might be of interest to compare how each lab was classified under the PCM procedure compared to what was done in the ad-hoc NIST bootstrap analysis. Under the PCM we

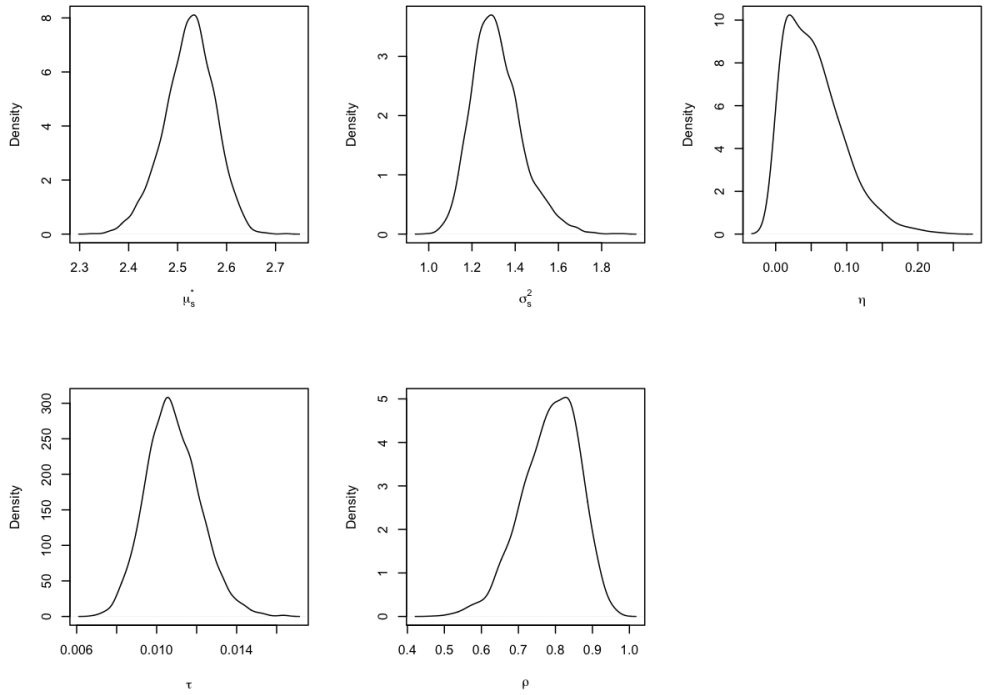

Figure 4: Marginal posterior distributions for for $\mu_s^*, \sigma_s^2, \eta, \tau$ and ρ. Time series plots of the MCMC iterates for each parameter are provided in the appendix.

classified the jth observation as an outlier if $E(\zeta_j|\boldsymbol{p}) = Pr(\zeta_j = 1|\boldsymbol{p}) < 0.1$. The results can be found in Table 3. If the entry in the "NIST" column is "Yes" then the corresponding laboratory was treated as an outlier in the NIST analysis. The opposite holds true for "No" entries. The columns "PCM" provide the posterior probabilities that laboratories were not outliers. (Perfect agreement between the NIST and Bayes analyses would pair "Yes" with 0 entries and "No" with 1.0 entries.) We again refer the reader to Figure 1 as a basis for qualitatively judging how a lab might be classified. Lab R is probably the most obviously outlying laboratory. The PCM produced posterior probability 1 that Lab R was an outlier. The labs that the NIST analysis and the PCM procedure classified differently are labs A, J and T. The NIST analysis classified them as outliers but the PCM procedure did not.

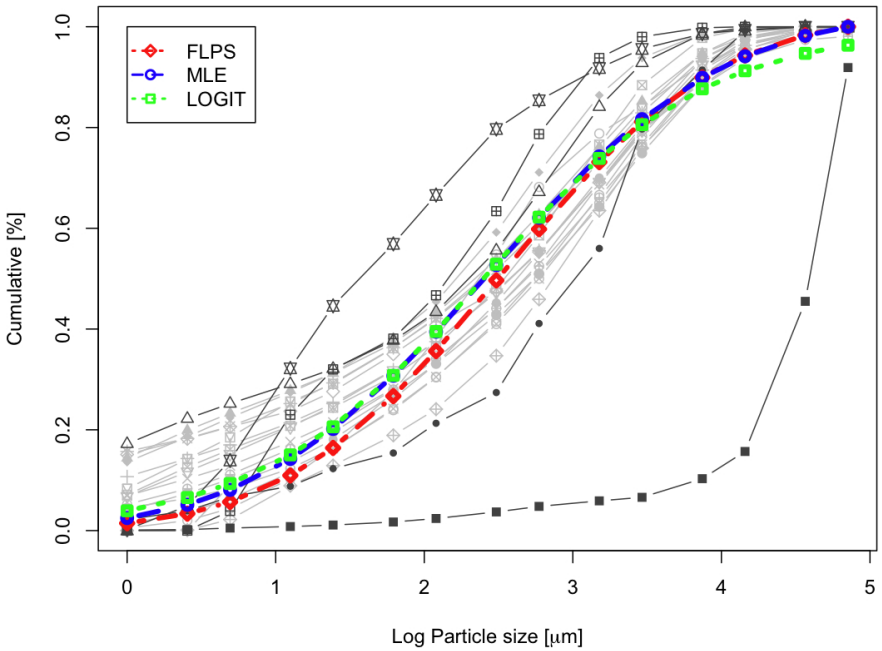

Figure 5: The estimated reference PSDs from the PCM procedure, MLE, and robust Logit model. The gray lines are the raw cumulative PSDs for all labs in the round-robin with the dark gray indicating labs that were considered outliers.

Table 3: Outlier lab classification obtained from the NIST analysis and using posterior probabilities obtained under the PCM

Lab	NIST	PCM	Lab	NIST	PCM
A	Yes	1.000	N	Yes	0.000
B	Yes	0.003	O	No	1.000
C	No	1.000	Q	No	0.998
D	No	1.000	R	Yes	0.000
F	No	0.997	S	No	0.999
G	No	0.999	T	Yes	0.998
H	No	0.731	U	Yes	0.179
I	No	0.999	V	No	1.000
J	Yes	0.861	W	Yes	0.071
K	No	1.000	X	No	1.000
L	Yes	0.001			

6 Extensions and Conclusions

A potentially huge benefit of the outlined methodology is that it is easily extended to a hierarchical setting. This would be very useful if, for example, each lab produced replicate weight fraction vectors. We could then potentially consider modeling the process level of the hierarchy with a mixture in addition (or in lieu of) the observation level. In this way, we assume the same basic kind of multivariate normal structure for observations from outlying labs and non-outlying labs (which is not the case for the PCM models where no multivariate normal structure is associated with outlying laboratory measurements). This may prove to be a useful extension in the presence of replication.

We have proposed a Bayes method that can be used to establish a reference PSD in the presence of outliers. Though the motivating example is from the perspective of an inter-laboratory study, this procedure could be used in any type of study conducted to estimate a PSD. The Bayesian methodology facilitates the addition of a mixture distribution to a very sensible likelihood that provides a PSD estimate that is robust to outliers. The methodology treats outliers rationally by virtue of the latent structure that is incorporated through the mixture. This latent structure provides nice probabilistic inference with regards to outlier classification and the uncertainty associated with this classification is accounted for in all estimation.

Appendices

A MCMC Algorigthm

Here we briefly describe the Metropolis-within-Gibbs MCMC algorithms used to simulate draws from the joint posterior distribution.

Let $\mu_s^{*(t)}$, $\sigma_s^{2(t)}$, $\eta^{(t)}$, $\tau^{(t)}$, $\rho^{(t)}$, and $\zeta_j^{(t)}$ denote the tth MCMC iterate for parameters μ_s^*, σ_s^2, η, τ, ρ, and ζ_j. We updated $\mu_s^{*(t)}$, $\sigma_s^{2(t)}$, $\eta^{(t)}$, and $\tau^{(t)}$ on an individual basis using random walk Metropolis steps with a normal proposal distribution. As an example we outline the process necessary to carrying out a Metropolis update for $\mu_s^{*(t)}$. Updates for the remaining parameters is similar. In what follows $\mu_s^{*(t)}$ is the first parameter within the t^{th} iteration to be updated. Let $L_(\mu_s^*, \sigma_s^2, \eta, \tau, \boldsymbol{\zeta}, \rho)$ be the likelihood function as described in display (6) then a Metropolis update can be had by the following 4 steps:

1. generate $\mu_s^{*new} \sim \text{N}(\mu_s^{*(t-1)}, \omega_{\mu_s})$ where ω_{μ_s} is a fixed known value,
2. compute $r_\mu = \dfrac{L_(\mu_s^{*new}, \sigma_s^{2(t-1)}, \eta^{(t-1)}, \tau^{(t-1)}, \boldsymbol{\zeta}^{(t-1)}, \rho^{(t-1)})\phi(\mu_s^{*new}; m_0, s_0^2)}{L_{\boldsymbol{q}}(\mu_s^{*(t-1)}, \sigma_s^{2(t-1)}, \eta^{(t-1)}, \tau^{(t-1)}, \boldsymbol{\zeta}^{(t-1)}, \rho^{(t-1)})\phi(\mu_s^{*(t-1)}; m_0, s_0^2)}$ where m_0 and s_0^2 are user-supplied constants (prior distribution values) and $\phi(\cdot)$ denotes the normal density,
3. generate $v \sim \text{Bernoulli}(\min(1, r_\mu))$, and
4. set $\mu_{s1}^{*(t)} = v\mu_{s1}^{*new} + (1-v)\mu_{s1}^{*(t-1)}$.

Values of ω_{μ_s} where chosen so that the proportion of MCMC iterates that resulted in accepting the proposed value was approximately 0.30. To update values for ρ and ζ_j a Gibbs step was used with the following full conditionals ($[\theta|-]$ denotes the distribution of θ conditioned on all parameters and data).

$$[\rho|-] \sim \text{Beta}\left(\sum \zeta_j + 9, L - \sum \zeta_j + 1\right)$$
$$[\zeta_j|-] \sim \text{Ber}(p^*) \text{ with } p^* = \frac{\boldsymbol{\phi}(\boldsymbol{q}_j; \boldsymbol{\delta}, \boldsymbol{\Delta})}{\boldsymbol{\phi}(\boldsymbol{q}_j; \boldsymbol{\delta}, \boldsymbol{\Delta}) \times \text{MVU}(\boldsymbol{q}_i; \boldsymbol{G}, \boldsymbol{H})}$$

The algorithms were numerically unstable if starting values for μ_s^* where chosen to be large (in absolute value) compared to σ_s^2. Therefore, the starting values should be carefully selected. In addition, occasionally small candidate values for $\sigma_s^{2(t)}$ where proposed. For these proposed values the likelihood could not be numerically evaluated because of zeros appearing in the denominators of ratios (2) and (3). To sidestep this, no proposed values

smaller than 0.01 were accepted in the Metropolis algorithm for $\sigma_s^{2(t)}$. This is reasonable because as $\sigma_s^{2(t)} \to 0$, $r_{\sigma_s} \to 0$.

References

[1] Bah, A. R., Kravchuk, O., Kirchhof, G. (2009) "Fitting Performances of Particle-size Distribution Models on Data Derived by Conventional and Laser Diffraction Techniques" *Soil Science Soceity of America Journal*, 73, 1101 – 1107.

[2] Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted 2003 with additional material by the Blackburn Press). 416 p.

[3] Egozcue, J. J., Powlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003) "Isometric Logratio Transformations for Compositional Data Analysis" *Mathematical Geology*, 35, 279 – 300.

[4] Ferraris, C. F., Hackley, V. A., Aviles, A. I., and Buchanan, C. E. (2002) "Analysis of the ASTM Round-Robin Test on Particle Size Distribution of Portland Cement: Phase I," Tech. rep., National Institute of Standards and Technology.

[5] Flegal, J. M., Haran, M., and Jones, G. L. (2008), "Markov Chain Monte Carlo: Can We Trust the Third Signicant Figure?" *Statistical Science*, 23, 250260.

[6] Filzmosor, P., Hron, K., and Templ, M. (2011), "Robust Compositional Data Analysis." In J. J. Egozcue, R. Tolosano-Delgado, amd M. I. Ortego, editors, *Proceedings of the 4th international Workshop on Compositional Data Analysis*, 4 pages, University of Girona, Girona, Spain, ISBN: 978-84-87867-76-7

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

[8] Hron, K., Matthias, T., and Filzmoser, P. (2010), "Imputation of Missing Values for Compositional Data Using Classical and Robust Methods" *Computational Statistics and Data Analysis*, 54.

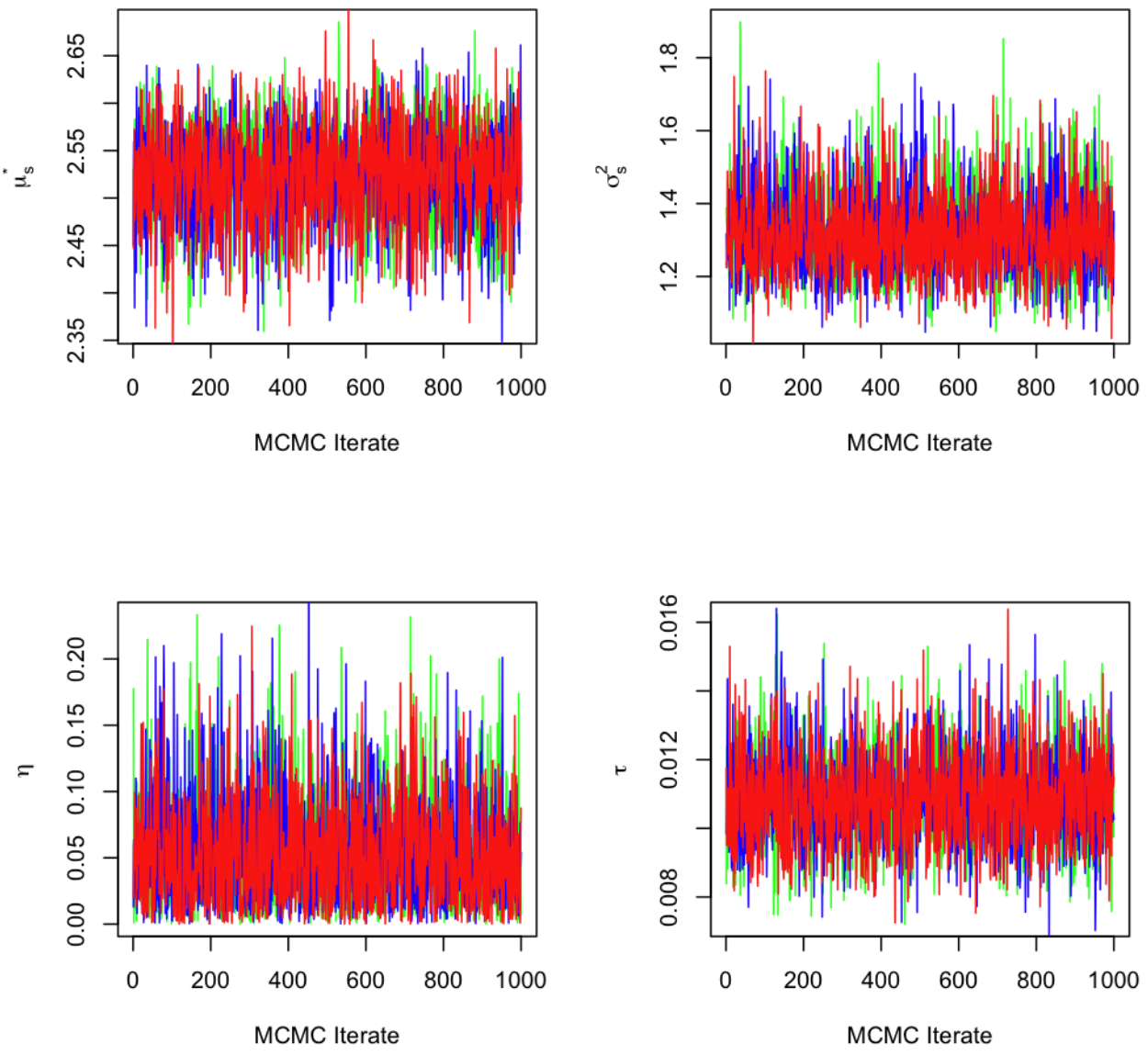

Figure 6: History plots of the MCMC iterates for $\mu_s^*, \sigma_s^2, \tau, \eta$

[9] Jasra, A., Holmes, C. C., and Stephens, D. (2005), "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling," *Statistical Science*, 20, 50 – 67.

[10] Leyva, N., Page, G. L., Vardeman, S. B., and Wendelberger, J. R. (2011), "Bayes Statistical Analyses for Particle Sieving Studies," Tech. rep., Los Alamos National Laboratory.

[11] Leyva, N., Vardeman, S. B. (2007), "Statistical inference for particle systems from sieving studies," Ph.D. Dissertation, Iowa State University.

[12] Little, R. J. A.., Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, New Jersey: John Wiley & Sons, 2nd ed.

[13] Lwin, T. (1994), "Analysis of Weight Frequency Distributions Using Replicated Data," *Technometrics*, 36, 28 – 36.

[14] Lwin, T. (2003), "Parameterization of Particle Size Distributions by Three Methods," *Mathematical Geology*, 35, 719 – 736.

[15] Page, G. L. and Vardeman, S. B. (2010), "Using Bayes methods and mixture models in inter- laboratory studies with outliers," *Accreditation and Quality Assurance*, 15, 379 – 389.

[16] R Development Core Team (2010), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[17] Scheaffer, R. L. (1969), "Sampling Mixtures of Multi-sized Particles: An Application of Renewal Theory," *Technometrics*, 11, 285 – 298.

[18] Van der Bilt, A., Fontign-Tekamp, F. A. (2004), "Comparison of Single and Multiple Sieve Methods for the Determination of Masticatory Performance," *Archives of Oral Biology*, 49, 193 – 198.

[19] Verdinelli, I., Wasserman, L. (1991), "Bayesian Analysis of Outlier Problems Using the Gibbs Sampler," *Statistics and Computing*, 1, 105 – 117.

[20] Zhang, H., Hu, D., Chen, J., Ye, X., Wang, S. X., Hao, J. M., Wang, L., Zhang, R., and An, Z., (2011), "Particle Size Distribution and Polycyclic Aromatic Hydrocarbons Emissions from Agricultural Crop Residue Burning," *Environmental Science and Technology*, 45, 5477 – 5482.