Bayes Statistical Analyses for Particle Sieving Studies

Norma Leyva [*]
Nonclinical Statistics
Teva Pharmaceuticals
norma.leyva@tevapharm.com

Garritt L. Page
Departamento de Estadística
Pontificia Universidad Católica de Chile
page@mat.puc.cl

Stephen B. Vardeman [†]
Statistics and IMSE Departments
Iowa State University
vardeman@iastate.edu

Joanne R. Wendelberger
Statistical Sciences Group
Los Alamos National Laboratory
joanne@lanl.gov

December 12, 2012

[*] This research was conducted primarily while the author was a student at Iowa State University, is independent of the authors work as a Teva employee, and does not reflect on her responsibilities at Teva or incorporate Teva confidential or proprietary information.

[†] Stephen Vardemans work was partially supported by NSF Grant DMS No. 0502347 EMSW21-RTG awarded to the Department of Statistics, Iowa State University.

Abstract

Particle size is commonly used to determine quality and predict performance of particle systems. We consider particle size distributions inferred from a material sample using a fixed number of sieves with progressively smaller size openings, where the weight of the particles in each size interval is measured. In this article we propose Bayes analyses for data from particle sieving studies based on parsimonously parameterized multivariate normal approximate models for vectors of log weight fraction ratios. Additionally, we observe that the basic approach extends directly to modeling mixture contexts, which provides added model flexibility and is a very natural extension when physical mixtures of materials with fundamentally different particle sizes are encountered. We also consider hierarchical modeling, where a single process produces lots of particles and the data available are (replicated) weight fraction vectors from several different lots. Supplementary materials for this article are available online.

1 Introduction

We consider contexts where specimens of a granular material are taken from a large supply of the material and run through a set of progressively finer sieves, and the fractions of the specimen weight captured on each sieve are measured to provide the basis for a characterization of the material through its "particle size distribution." As an aside, we note that in the case of weight fraction analysis, this 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. But, "particle size distribution" is standard terminology in this area and we use it throughout this discussion.

Particle size distributions are of interest in a number of fields involving powders and other bulk materials. For background on statistical and other issues in bulk sampling see Duncan

(1962), Sommer (1986), Gy (1992), Pitard (1993), and Smith (2001). The breadth of interest in this general problem is indicated by the applications of Maricq et al. (1999) who describe particle size distributions of emissions from gasoline vehicles, Dalby and Byron (1988) who compare particle size distributions from pressurized aerosols, and Van der Bilt et al. (1993) who discuss data analysis methods for studies of chewed food particles. A simple example of this type of data is found in Lwin (1994) represented here in Figure 1, and available in the supplemental materials in tabular form. This data set consists of sieving results from 6 specimens of about 2g each of scheelite-ore fines, pre-truncated to contain only particles below 9.00μm in size. The data are presented in a cumulative weight function form (as is quite common).

When dealing with bulk materials, it is not uncommon to encounter mixtures of fundamentally different particle types and basic sizes. Examples include mixtures of sediments and larger rock particles found in stream beds, heterogeneous road paving materials, blended polymers in industrial applications, fertilizer mixtures, or in pulmonary administration of drugs, drug micronized particles are usually blended with coarse and fine carrier particles to improve flowability. So modeling and data analysis methods for mixtures of basic particle size distributions are also important. For example, Smyth and Hickey (2003) treat a pharmaceutical example involving a mixture obtained from pressurized metered-dose inhalers.

The basic modeling adopted here is based on Scheaffer (1969), Lwin (1994), and Leyva (2006). Scheaffer (1969) provided a basis for probabilistic modeling and statistical analysis of particle size data based on standard renewal theory where instead of accumulating time, the process refers to accumulating weight. Lwin (1994) used the probability structure of Scheaffer (1969) and considered problems of inference from (replicated) sample weight fraction data. He noted that a model of random sampling of particles up to a fixed target

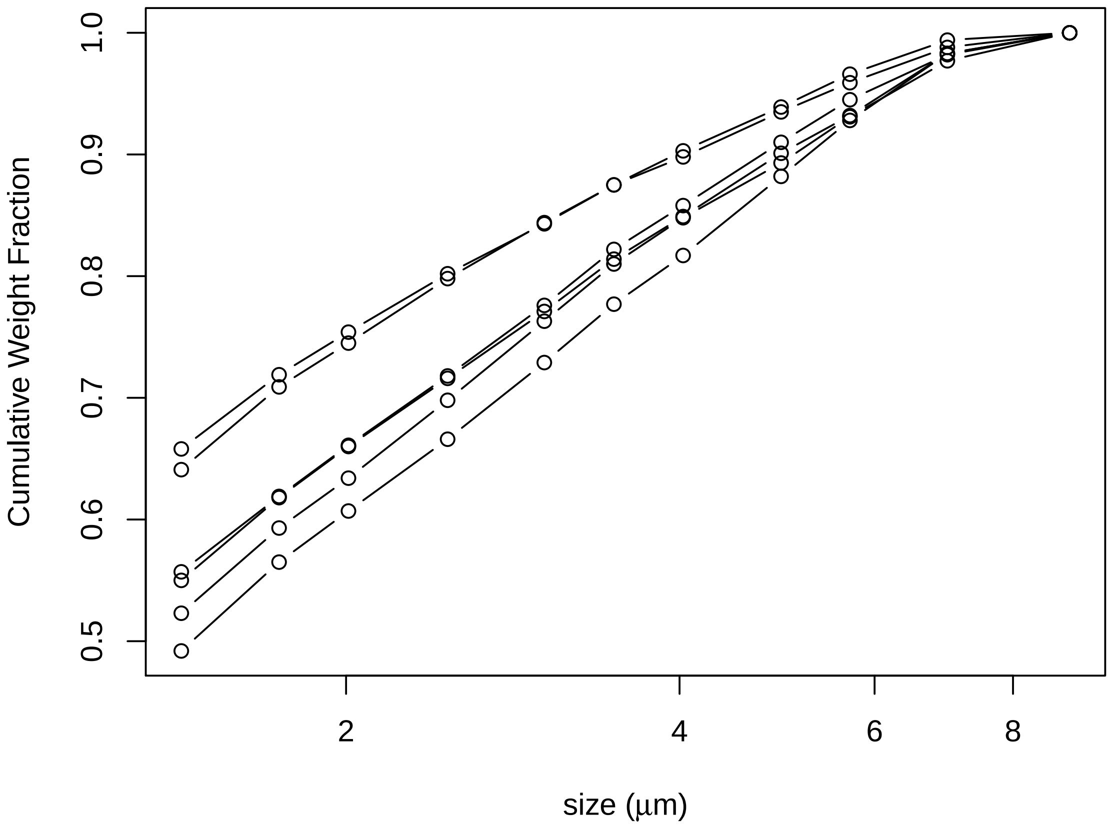

Figure 1: 6 Vectors of Cumulative Weight Fractions, from Lwin (1994)

.

specimen weight produces an approximately multivariate normal distribution for weight fraction vectors, whose parameters depend upon the bivariate distribution of sizes and weights of particles in the supply of material. Leyva (2006) additionally noted that employing a log-normal assumption on the (marginal) distribution of particle sizes (which is fairly common, see Allen 2003) and a power law assumption on the mean and standard deviation of weight for particles of a given size, one arrives at a relatively simple and fairly standard 4-parameter multivariate normal approximate distribution for weight fraction vectors.

The remainder of the article is organized as follows. In Section 2 we provide a review of the development and notation of Leyva (2006) needed in this paper. Section 3 contains Bayesian extensions in a one-sample sieving study setting. Section 4 provides a generalization and development of the mixture ideas with some motivations. Section 5 provides extensions to a hierarchical setting. Finally, some concluding remarks are provided in Section 6.

2 Background: Log Normal Modeling of Particle Size Distributions

Let an individual particle size be denoted by S and the corresponding particle weight be denoted by W. Suppose that a specimen of approximate weight m is to be sampled from a reservoir of particles and run through a series of $k-1$ successively finer sieves. Let $C_1, \ldots, C_{k-1}$ denote sieve sizes with $0 \leq C_0 < C_1 < \cdots < C_k \leq \infty$. Then the k particle size classes are defined by the intervals $[C_{i-1}, C_i)$ for $i = 1, \ldots, k$, and the observed specimen weight fractions for the k classes are $p_1, p_2, \ldots, p_k$. Often the smallest size sieve does not retain all the material and other methods (such as sedimentation) are employed to characterize the particle distribution for fine material. For this reason C_0 is potentially strictly positive. Now, let S have a probability density $f(s|\theta)$ for some parameter vector θ such that $\mathrm{E}[W|S=s] = \kappa s^\eta$ and $\mathrm{E}[W^2|S=s] = \kappa' s^{2\eta}$, for positive constants η, κ, and κ'. The moment conditions ensure that the conditional standard deviation of an observed particle weight is proportional to its conditional mean. The the first two moments of the weight distributions for particles in the ith sieve class are then

$$b_i = \kappa \frac{\int_{C_{i-1}}^{C_i} s^\eta f(s|\theta) ds}{\int_{C_{i-1}}^{C_i} f(s|\theta) ds} \quad \text{and} \quad b_i^2 = \kappa' \frac{\int_{C_{i-1}}^{C_i} s^{2\eta} f(s|\theta) ds}{\int_{C_{i-1}}^{C_i} f(s|\theta) ds} \tag{1}$$

If $\log(S) \sim N(\mu_s, \sigma_s^2)$ (or possibly truncated lognormal), and one adopts the model just described for the generation of specimens consisting of random sampling of particles up to a fixed specimen (total particle) weight m, it follows directly from the analysis of Scheaffer (1969) (with more details provided in Lwin 1994) that the limiting distribution (as $m \to \infty$) of the vector of observed weight fractions is $\boldsymbol{p} \equiv (p_1, p_2, \ldots, p_k)' \dot{\sim} \text{MVN}(\boldsymbol{\pi}, \boldsymbol{\Sigma})$. Here $\boldsymbol{\pi}$ and $\boldsymbol{\Sigma}$ are relatively simple functions of four parameters, (μ_s, σ_s^2, η, τ), where $\tau = \kappa'/(m\kappa)$. The mean vector is $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)'$ for

$$\pi_i = \frac{\Phi(\frac{\log C_i - \mu_s^*}{\sigma_s}) - \Phi(\frac{\log C_{i-1} - \mu_s^*}{\sigma_s})}{\Phi(\frac{\log C_k - \mu_s^*}{\sigma_s}) - \Phi(\frac{\log C_0 - \mu_s^*}{\sigma_s})}, \tag{2}$$

where $\Phi(\cdot)$ denotes a standard normal cdf and $\mu_s^* = \mu_s + \eta\sigma_s^2$ (which is a consequence of the partial expectations found in (1)). The entries of $\boldsymbol{\Sigma}$ are also functions of the four parameters and are

$$\text{Cov}(p_i, p_u) = \tau \cdot \begin{cases} \pi_i(1-\pi_i)\gamma_i^* + \pi_i^2 \left[\sum_{j=1}^{k} \pi_j \gamma_j^* - \gamma_i^* \right] & \text{for } i = u \\ \pi_i \pi_u \left[\sum_{j=1}^{k} \pi_j \gamma_j^* - \gamma_i^* - \gamma_u^* \right] & \text{for } i \neq u \end{cases} \tag{3}$$

with

$$\gamma_i^* = b_i^2/b_i = e^{\eta(\mu_s^* + 0.5\eta\sigma_s^2)} \frac{\Phi\left(\frac{\log C_i - (\mu_s^* + \eta\sigma_s^2)}{\sigma_s}\right) - \Phi\left(\frac{\log C_{i-1} - (\mu_s^* + \eta\sigma_s^2)}{\sigma_s}\right)}{\Phi\left(\frac{\log C_i - \mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_{i-1} - \mu_s^*}{\sigma_s}\right)}. \tag{4}$$

Finally, under the assumptions above, the cumulative weight fraction up to size s is

$$CW(s) = \frac{\Phi\left(\frac{\log s - \mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_0 - \mu_s^*}{\sigma_s}\right)}{\Phi\left(\frac{\log C_k - \mu_s^*}{\sigma_s}\right) - \Phi\left(\frac{\log C_0 - \mu_s^*}{\sigma_s}\right)} \tag{5}$$

which is a form employed regularly by practitioners. Though not explicitly investigated here, it seems reasonable to expect that as size of particles being analyzed decreases, the amount of material m needed to effectively characterize the particle size distribution would decrease.

The logic just outlined has the virtue of providing a model for a potentially very-high dimensional random vector using a multivariate normal distribution that is parsimoniously parameterized with only four parameters. Also, the four parameters (at least in principle) have intuitively appealing interpretations. μ_s^* and σ_s describe the normal or truncated normal "cumulative weight fraction as a function of log size" function. η, which other authors have pointed to as potentially characterizing the "shape of average particles" and suggested should be between 0 and 3, and τ, that is some scaling factor that is inversely proportional to the total weight of the specimen being sieved, are additionally needed to determine the covariance matrix. It is worth noting that Lwin (1994) arrives at (and employs) the $\eta = 3$ version of this approximate distribution from the (very much stronger assumptions) that $(\log S, \log W)$ is bivariate normal, upon assuming that $\text{Cov}(S, W) = 3\text{Var}(S)$.

The final form of the (large amount of material sampled) limiting distribution is very attractive. It is common practice in the analysis of sieving data to treat the cumulative weight fraction function as having (a parametric sigmoidal shape and most often) the shape of a normal cdf in the argument log s. That the expected value of $\boldsymbol{p}$ is $\boldsymbol{\pi}$ cannot be more natural. Large m multivariate normality is surely plausible. And form (3) provides a defensible patterned structure for $\boldsymbol{\Sigma}$, *without which for typical large numbers of sieves and typical (very small) numbers of observation vectors, anything like formal multivariate statistical inference seems quite hopeless.*

Probably the biggest drawback of the basic multivariate normal modeling just described is that although the approximate model guarantees that with probability 1 the observed weight fractions sum to 1, there is positive (and appreciable for τ large enough) probability assigned to the event that at least one observed weight fraction is negative. An alternative to direct use of the limiting distribution for $\boldsymbol{p}$ under the Leyva (2006) assumptions that does not have this failing is to consider instead the limiting distribution of the vector of log ratios of $k-1$ elements of $\boldsymbol{p}$ to the other (fixed) element of $\boldsymbol{p}$. For example, choosing p_1 as a reference weight fraction and using the delta method, the limiting distribution (as $m \to \infty$) of $\boldsymbol{q} = (\log p_2 - \log p_1, \log p_3 - \log p_1, \ldots, \log p_k - \log p_1)' \mathrel{\dot\sim} \text{MVN}_{k-1}(\boldsymbol{\delta}, \boldsymbol{\Delta})$ where $\boldsymbol{\delta} = (\delta_2, \delta_3, \ldots, \delta_k)'$ with $\delta_i = \log \pi_i - \log \pi_1$, for $i = 2, \ldots, k$, and the entries of $\boldsymbol{\Delta}$ are

$$\text{Cov}(q_i, q_u) = \tau \cdot \begin{cases} \dfrac{1}{\pi_i}\gamma_i^* + \dfrac{1}{\pi_1}\gamma_1^* & \text{for } i = u \\ \dfrac{1}{\pi_1}\gamma_1^* & \text{for } i \neq u \end{cases} . \tag{6}$$

In what follows we focus solely on the $\boldsymbol{q}$-likelihood as (at least in our examples) it seems to be more numerically stable than the $\boldsymbol{p}$-likelihood and doesn't suffer from the drawback just described.

Finally, there are some formal similarities between the ultimate multivariate normal modeling here for $\boldsymbol{p}$ and $\boldsymbol{q}$ and the compositional data analysis of Aitchison (1986) and others. But in these works the indexing on entries of $\boldsymbol{p}$ and $\boldsymbol{q}$ is essentially arbitrary (unlike our situation) and highly specialized parametric forms for means and covariance matrices like those we will employ is lacking.

3 Bayes Analyses of One Sample Sieving Studies

First consider a situation where n specimens of the same supply of material are sieved and weight fraction vectors $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ are observed. For an arbitrary choice of reference size class (we will use size class 1), let $\boldsymbol{q}_i$ be the $(k-1)$-dimensional vector of log weight fraction ratios corresponding to $\boldsymbol{p}_i$ as considered in the previous section.

In particle size studies there is often more practical interest in parametric functions and predictions associated with the model than there is in the values of individual model parameters (μ_s^*, σ_s^2, η, τ). In particular, there is scientific interest in values of the parametric functions $CW(s)$ and $CW^{-1}(p)$ (the cumulative weight fraction function and the function giving particle sizes corresponding to input cumulative weight fractions), perhaps the $\pi_i = CW(C_i) - CW(C_{i-1})$, and for $\boldsymbol{p}_{\text{new}}$ an additional (unobserved) weight fraction vector, the values

$$p_{\text{new},i} \text{ and } t_{\text{new},i} \equiv \sum_{j=1}^{i} p_{\text{new},j}. \tag{7}$$

The last of these is the set of empirical cumulative weight fractions associated with $\boldsymbol{p}_{\text{new}}$. For a given s, p, or i, all of the values $CW(s)$, $CW^{-1}(p)$ and π_i are very easily estimated under the Bayesian paradigm which we adopt here.

Under the models of Section 2, the $\boldsymbol{q}_i$ have (non-singular) approximately MVN_{k-1} distributions. So, let $h(\boldsymbol{q}|\mu_s^*, \sigma_s, \eta, \tau)$ be the $\text{MVN}_{k-1}(\boldsymbol{\delta}, \boldsymbol{\Delta})$ pdf. With this notation, a likelihood function based on the vectors of log ratios of the weight fractions is

$$L_{\boldsymbol{q}}(\mu_s^*, \sigma_s^2, \eta, \tau) = \prod_{i=1}^{n} h(\boldsymbol{q}_i|\mu_s^*, \sigma_s^2, \eta, \tau). \tag{8}$$

For $g(\mu_s^*, \sigma_s^2, \eta, \tau)$ a joint prior density for the parameters μ_s^*, σ_s, η, and τ, posterior densities for the parameters are

$$g_{\boldsymbol{q}}(\mu_s^*, \sigma_s^2, \eta, \tau | \boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n) \propto L_{\boldsymbol{q}}(\mu_s^*, \sigma_s^2, \eta, \tau) g(\mu_s^*, \sigma_s^2, \eta, \tau). \tag{9}$$

Upon employing Markov Chain Monte Carlo to sample from one of these posteriors, credible intervals for parametric functions of $(\mu_s^*, \sigma_s, \eta, \tau)$ are immediate, as are approximate predictive posteriors of an additional weight fraction vector $\boldsymbol{p}_{\text{new}}$.

3.1 An Example

As mentioned in the Introduction, Lwin (1994) used a data set consisting of sieving results from 6 specimens of about 2g each of scheelite-ore fines, pre-truncated to contain only particles below 9.00μm in size that was represented here in Figure 1. A relatively ad hoc analysis of Lwin suggests that roughly $\mu_s^* \approx 0.25$ and $\sigma_s \approx 1.29$ are consistent with the data. The maximum likelihood analysis of Leyva (2006) based on the $\boldsymbol{p}$-likelihood produces point estimates $\widehat{\mu_s^*} = 0.217$, $\hat{\sigma}_s = 1.286$, $\hat{\eta} = 1.177$, and $\hat{\tau} = 0.001$. We here consider Bayes analyses of the Lwin data based on the $\boldsymbol{q}$-likelihood.

We considered a number of different prior distributions for $(\mu_s^*, \sigma_s^2, \eta, \tau)$. For each we assumed the parameters to be *a priori* independent. The resulting posterior distributions under all the prior specifications were similar, providing some empirical evidence the analysis is more or less robust to the details of the prior. Because of this, we report only the results obtained employing $\mu_s^* \sim N(0, 10)$, $\sigma_s^2 \sim \text{inv-}\Gamma(0.1, 0.1)$, $\eta \sim \text{U}(0, 3)$, and $\tau \sim \text{Exp}(0.01)$ as priors. (The second parameter of the normal distribution is the variance and the parameter of the exponential distribution is the mean.) The prior distribution used for σ_s^2 is a commonly used approximately non-informative prior for variances. The prior for μ_s^* could be thought of as non-informative in the range of values that are plausible for μ_s^* (clearly μ_s^* is highly unlikely to be greater than 10). A random walk Metropolis-within-Gibbs algorithm wherein each parameter is updated individually was used to obtain 1000 MCMC iterates after discarding the first 50,000 as burn-in and thinning by 100 (more details regarding the MCMC algorithm employed can be found in the supplemental materials). The posterior medians and 95% credible intervals for the model parameters are recorded in Table 1.

Table 1: Posterior inferences for model parameters for the Lwin data.

Parameter	Median	95% Credible Interval
μ_s^*	0.2206	(0.1411, 0.3166)
σ_s	1.2902	(1.1632, 1.4218)
η	1.1482	(0.4627, 1.8991)
τ	0.0010	(0.0003, 0.0022)

The two parameters μ_s^* and σ_s completely determine $CW(s)$ and are determined by $CW(s)$ while sample information about η and τ is available only through the variance-covariance structure of the sample weight fraction vectors. So it is not surprising (at least in light of the fact that $\log(9.00) = 2.20$ and $\log(1.42) = 0.35$) that μ_s^* and σ_s seem far more precisely determined than the parameters η and τ. It is also worth noting that the inferences shown in Table 1 are reasonably compatible with the maximum likelihood estimates of the model parameters based on the $\boldsymbol{q}$-likelihood.

Figure 2 provides posterior estimates of $CW^{-1}(p)$, $CW(s)$, and $\boldsymbol{p}_{\text{new}}$. One general impression provided by the figure is that for the Lwin data the methodology provides a plausible fit based on the fairly simple 2-parameter form of $CW(s)$ and allows for enough uncertainty in predictions to make the observed data themselves plausible realizations under the posterior model. However, there is a slight lack of fit for large size, and for these data the prediction

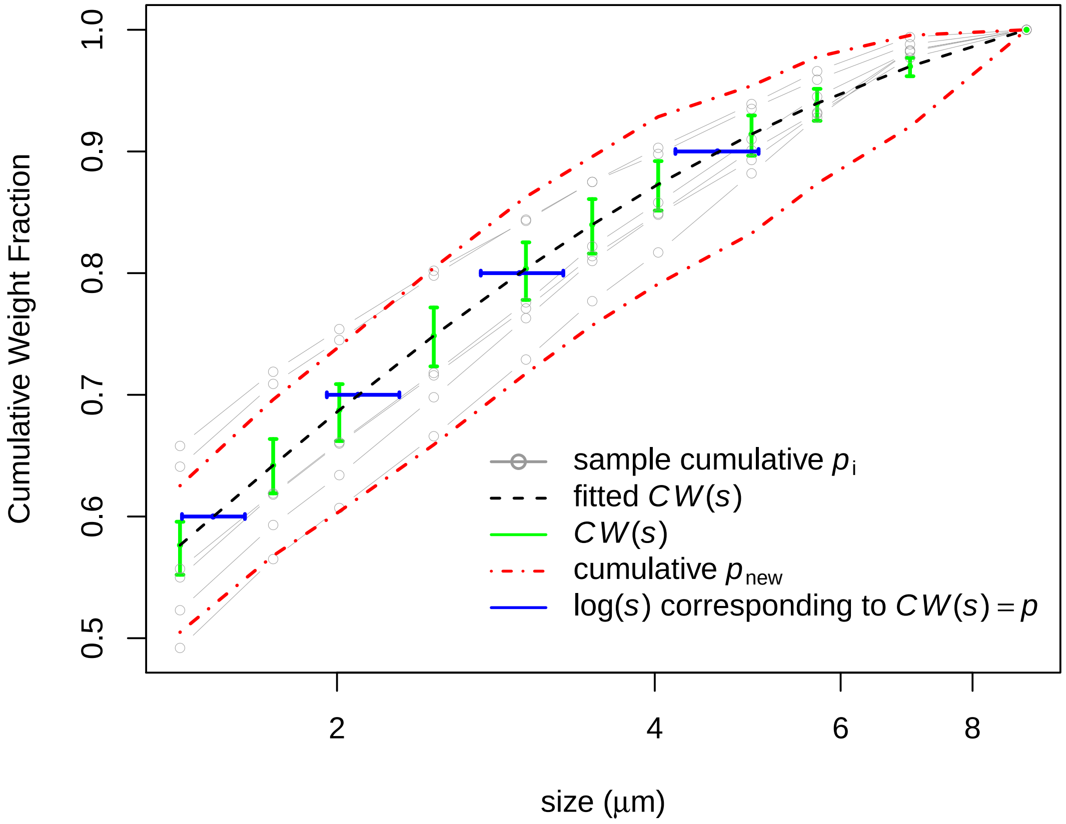

Figure 2: Posterior inferences based on Lwin's data. The dash-dotted lines correspond to 95% error bands for $\boldsymbol{p}_{\text{new}}$. The vertical bars correspond to 95% point-wise credible intervals for $CW(s)$. The horizontal bars indicate 95% credible intervals for the value of $\log(s)$ corresponding to a particular $p = CW(s)$.

bounds don't seem to mimic the narrowing of the sample paths as log size increases. This suggests that the 2-parameter form employed for $CW(s)$ may not be quite flexible enough for the present application, and that the assumptions here are not quite adequate to capture the pattern of variation in the cumulative sums. In the next section we address an extension that is able to more flexibly capture the natural sample variability.

This Bayesian approach to the analysis of one-sample sieving experiments (when n specimens are drawn from one batch of material) can be generalized to hierarchical modeling for situations where specimens are drawn from more than one batch of material taken from a single process. It seems reasonable in such a case that there is an underlying central weight fraction distribution for the process, and that the model parameters for each single batch of material somehow share the characteristics of the process. We elaborate on this line of thinking in Section 5.

4 Extension to a Mixture of Two Particle Types, Each with a Lognormal Distribution of Particle Sizes

While it is common to treat a cumulative weight fraction function as having the shape of a normal cdf in the argument log s, there are applications where this is simply not adequate. It is also the case that there are contexts in which it is common to purposely mix materials with very different particle sizes in order to achieve a desired performance of the blend, and there are other contexts where unintentional mixtures of particle sizes occur when there are agglomerates. In particular, we have encountered an example where two particle types have been manufactured and later mixed, and the above modeling clearly needs generalization.

In generalization of the basic modeling of Leyva (2006) suppose that a supply of material contains two types of particles mixed together in fractions (of *particles, not weight*) $(1-\rho)$ and ρ for $0<\rho<1$. Then, suppose that for particles of Types 1 and 2, $\log S \sim \mathrm{N}(\mu_{si}, \sigma_{si}^2)$, for $i = 1, 2$, (or $\log S \sim$ truncated-$\mathrm{N}(\mu_{si}, \sigma_{si}^2)$ in the event that $0 < C_0$ and/or $C_k < \infty$), and

$$\mathrm{E}[W|S=s] = \kappa_i s^{\eta_i} \text{ and } \mathrm{E}[W^2|S=s] = \kappa_i' s^{2\eta_i},\ i = 1, 2$$

We are here allowing both different distributions of size and different relationships between size and weight for the two types of particles. Under the model of random sampling of particles up to a fixed target specimen weight, one may still apply the Scheaffer renewal theory argument to produce MVN approximations to distributions for $\boldsymbol{p}$ and $\boldsymbol{q}$. The forms for the π_i and the γ_i simply have to be changed to reflect the more complicated scenario.

For $f(s|\mu_s, \sigma_s)$ a lognormal density with parameters μ_s and σ_s, let

$$\begin{aligned} I(a, b, \eta, \mu_s, \sigma_s) &= \int_a^b s^\eta f(s|\mu_s, \sigma_s) ds \\ &= \exp\left(\eta\mu_s + \frac{1}{2}\eta^2\sigma_s^2\right)\left(\Phi\left(\frac{\log b - (\mu_s + \eta\sigma_s^2)}{\sigma_s}\right) - \Phi\left(\frac{\log a - (\mu_s + \eta\sigma_s^2)}{\sigma_s}\right)\right). \end{aligned}$$

Note, for example, that in this notation the fraction of the weight of a large supply of the material attributed to particles of Type 1 is

$$\textit{Type 1 Weight Fraction} = \frac{(1-\rho)\kappa_1 I(C_0, C_k, \eta_1, \mu_{s1}, \sigma_{s1})}{(1-\rho)\kappa_1 I(C_0, C_k, \eta_1, \mu_{s1}, \sigma_{s1}) + \rho\kappa_2 I(C_0, C_k, \eta_2, \mu_{s2}, \sigma_{s2})}$$

and the cumulative weight fraction function is

$$CW(s) = \frac{(1-\rho)\kappa_1 I(C_0, s, \eta_1, \mu_{s1}, \sigma_{s1}) + \rho\kappa_2 I(C_0, s, \eta_2, \mu_{s2}, \sigma_{s2})}{(1-\rho)\kappa_1 I(C_0, C_k, \eta_1, \mu_{s1}, \sigma_{s1}) + \rho\kappa_2 I(C_0, C_k, \eta_2, \mu_{s2}, \sigma_{s2})}, \tag{10}$$

and this latter (for fixed C_0 and C_k) is a generalization of the essentially 2-parameter form (5). Clearly, with form (10) again one has $\pi_i = CW(C_i) - CW(C_{i-1})$. Then taking

$$\gamma_i = \frac{(1-\rho)\kappa_1' I(C_{i-1}, C_i, 2\eta_1, \mu_{s1}, \sigma_{s1}) + \rho\kappa_2' I(C_{i-1}, C_i, 2\eta_2, \mu_{s2}, \sigma_{s2})}{(1-\rho)\kappa_1 I(C_{i-1}, C_i, \eta_1, \mu_{s1}, \sigma_{s1}) + \rho\kappa_2 I(C_{i-1}, C_i, \eta_2, \mu_{s2}, \sigma_{s2})} \tag{11}$$

we may write the entries of $\boldsymbol{\Sigma}$ as in form (3), but substituting γ_i^* by γ_i, and τ by $1/m$. The resultant parametric form substantially generalizes the earlier form (3) (depending, as it does, on many more parameters). This development extends directly to an approximate distribution for $\boldsymbol{q}$. In particular, the generalization of the covariance form (6) replaces γ_i^* with γ_i, and τ with $1/m$. Important choices and complications arise as one considers Bayes inference under the added modeling complexity afforded by the mixture structure. We briefly note two of these.

A basic question is how much commonality one wants to assume about particle types. In an example that we will consider in Section 5, the chemical make-up and to some degree the physical structure of small and large particles were meant to be similar. Such may not always be the case, but when it is, it suggests the possibility of reducing a list of 11 potential model parameters $\rho, \mu_{s1}, \sigma_{s1}, \eta_1, \kappa_1, \kappa_1', \mu_{s2}, \sigma_{s2}, \eta_2, \kappa_2, \kappa_2'$ to a smaller (and operationally more manageable) list of 8 parameters $\rho, \mu_{s1}, \sigma_{s1}, \mu_{s2}, \sigma_{s2}, \eta, \kappa, \kappa'$ through the assumptions that $\eta_1 = \eta_2$, $\kappa_1 = \kappa_2$, and $\kappa_1' = \kappa_2'$. In words, there is a single relationship between particle size and weight operating for both particle types.

But even if one reduces to the set of 8 model parameters, a basic statistical complication of mixture modeling will remain, in that without imposing some restrictions on the two parameter pairs (μ_{s1}, σ_{s1}) and (μ_{s2}, σ_{s2}), the model will not be identifiable. One way of dealing with this is to assume that $\mu_{s1} < \mu_{s2}$ and in a Bayes analysis, to use a prior distribution for μ_{s1} and μ_{s2} that places mass 1 on the event that $\mu_{s1} < \mu_{s2}$. A tractable type of prior that can be used in this context is truncated bivariate normal, where the truncation is to the set of (μ_{s1}, μ_{s2}) satisfying the inequality.

In addition to handling weight fraction vectors that originate from material that is knowingly a mixture of different particle sizes, the mixture extension can simply be used as means to provide a methodology that is more flexible in modeling weight fraction vectors than that found in Section 3. To demonstrate this possibility we apply the mixture model idea to the Lwin data. For sake of computational simplicity we set

$$\eta = \eta_1 = \eta_2,\ \kappa = \kappa_1 = \kappa_2 \text{ and } \kappa' = \kappa'_1 = \kappa'_2. \tag{12}$$

This condition simplifies the mean cumulative weight fraction and covariance functions in the following manner.

$$CW(s) = \frac{(1-\rho)I(C_0, s, \eta, \mu_{s1}, \sigma_{s1}) + \rho I(C_0, s, \eta, \mu_{s2}, \sigma_{s2})}{(1-\rho)I(C_0, C_k, \eta, \mu_{s1}, \sigma_{s1}) + \rho I(C_0, C_k, \eta, \mu_{s2}, \sigma_{s2})} \tag{13}$$

and expression (11) becomes

$$\gamma_i = \left(\frac{\kappa'}{\kappa}\right) \frac{(1-\rho)I(C_{i-1}, C_i, 2\eta, \mu_{s1}, \sigma_{s1}) + \rho I(C_{i-1}, C_i, 2\eta, \mu_{s2}, \sigma_{s2})}{(1-\rho)I(C_{i-1}, C_i, \eta, \mu_{s1}, \sigma_{s1}) + \rho I(C_{i-1}, C_i, \eta, \mu_{s2}, \sigma_{s2})}$$

and letting

$$\gamma_i^* = \left(\frac{\kappa}{\kappa'}\right) \gamma_i = \frac{(1-\rho)I(C_{i-1}, C_i, 2\eta, \mu_{s1}, \sigma_{s1}) + \rho I(C_{i-1}, C_i, 2\eta, \mu_{s2}, \sigma_{s2})}{(1-\rho)I(C_{i-1}, C_i, \eta, \mu_{s1}, \sigma_{s1}) + \rho I(C_{i-1}, C_i, \eta, \mu_{s2}, \sigma_{s2})} \tag{14}$$

with $\tau = \kappa'/m\kappa$, the covariance forms (3) and (6) carry over exactly to the mixture context (once the more complicated (14) replaces (4)).

Therefore under (12), the mixture model for a single vector $\boldsymbol{q}$ has parameters $\rho, \mu_{s1}, \sigma_{s1}, \mu_{s2}$, $\sigma_{s2}, \eta,$ and τ, the first 6 of which determine the mean vector (and up to the scale factor τ) the covariance matrix for the multivariate normal distributions. To do a one-sample Bayes analysis, one must place priors on these 7 parameters.

To preserve identifiability we employ a truncated bivariate normal prior distribution for

(μ_{s1}^*, μ_{s2}^*)

$$\begin{pmatrix} \mu_{s1}^* \\ \mu_{s2}^* \end{pmatrix} \overset{iid}{\sim} \text{truncated-BVN} \left(\begin{pmatrix} m_{10} \\ m_{20} \end{pmatrix}, \mathbf{diag}(s_{10}^2, s_{20}^2) \right)$$

where the truncation is to the part of $\mathbb{R}^2$ where $\mu_{s1}^* < \mu_{s2}^*$, independent of $\sigma_{si}^2 \overset{iid}{\sim}$ inv-$\Gamma(0.1, 0.1)$, for $i = 1, 2$, independent of $\rho \sim$ Beta$(1, 1)$ and of $\eta \sim$ U$(0, 3)$, and $\tau \sim$ Exp(0.01). We set $m_{10} = m_{20} = 0$ and $s_{10}^2 = s_{20}^2 = 10$. As in the non-mixture case, inverse Gamma distributions with small scale and shape parameters are typical approximately non-informative priors for variance parameters. Additionally *a piori* little is known about ρ, hence a uniform distribution on the interval (0,1) seems reasonable.

Table 2: Posterior inferences for mixture model parameters for the Lwin data.

Parameter	Median	95% Credible Interval	Parameter	Median	95% Credible Interval
μ_{s1}^*	-0.06	$(-0.29, 0.11)$	η	1.16	(0.48, 1.95)
σ_{s1}^2	0.69	(0.11, 1.21)	τ	0.001	(0.000, 0.002)
μ_{s2}^*	1.46	(1.11, 1.61)	ρ	0.15	(0.03, 0.34)
σ_{s2}^2	0.16	(0.05, 0.38)			

Results from the mixture model applied to the Lwin data can be found in Table 2 and Figure 3. Notice that the inferences regarding $CW(s)$ and $\boldsymbol{p}_{\text{new}}$ found in Figure 3 are very similar to those in Figure 2, save the fact that the variability associated with the cumulative weight fractions at large particle sizes is better captured in the former. Even though ρ was introduced primarily to provide more flexibility in modeling, it also provides some potentially valuable information. That is, it appears plausible that the material represented by the Lwin data is comprised of a mixture of particle types whose composition is roughly 15% for Type

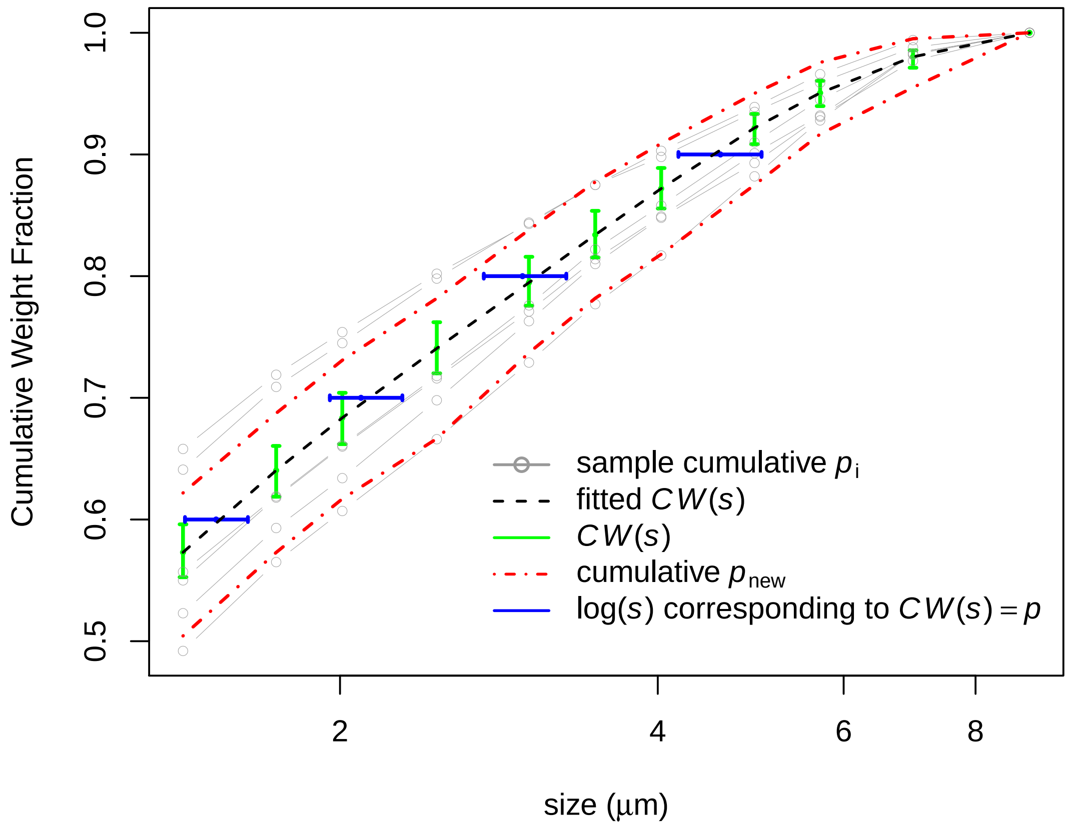

Figure 3: Results of Lwin data analysis when using a mixture model. The dash-dotted lines correspond to 95% error bands for $\boldsymbol{p}_{\text{new}}$. The vertical bars correspond to 95% point-wise credible intervals for $CW(s)$. The horizontal bars are 95% credible intervals for the value of $\log(s)$ corresponding to a particular $p = CW(s)$.

1 to 85% for Type 2 (by particle count).

5 Bayes Analyses of Sieving Studies With Hierarchical Structure

We have addressed the case where from the same supply of material (say, a jth batch), n weight fraction vectors $\boldsymbol{p}_{j1}, \boldsymbol{p}_{j2}, \ldots, \boldsymbol{p}_{jn}$ are gathered. Here we consider the possibility of sampling n specimens from each of several different batches indexed by j, themselves all coming from a fixed process.

Let $\boldsymbol{p}_{ji}$ describe the ith replicate taken from the jth batch of material. In order to accommodate samples from different batches, we might modify the model in Section 3 by allowing the model parameters μ_s^* and σ_s^2 to vary batch-to-batch around the characteristics of the process that generated the different batches.

That is, one could suppose that across batches j, $\mu_{sj}^* \overset{iid}{\sim} \mathrm{N}(\mu_{s\text{process}}^*, \lambda^2)$ independent of $\sigma_{sj}^2 \overset{iid}{\sim}$ inv-$\Gamma(\alpha, \beta)$. One might further adopt independent priors with $\mu_{s\text{process}}^* \sim \mathrm{N}(\mu_0, \sigma_{\text{process}0}^2)$, $\lambda^2 \sim$ inv-$\Gamma(a_0, b_0)$, and $\alpha, \beta \sim \mathrm{Exp}(c_0)$. And, in light of the potential interpretation of η and τ as parameters relating particle weight to size, it is plausible to assume all batches have the same parameters η and τ and use the (also independent) priors $\eta \sim \mathrm{U}(0, 3)$ and $\tau \sim \mathrm{Exp}(\lambda_0)$.

As in the one-sample case, using an MCMC algorithm we can hope to obtain approximate posterior distributions of the model parameters and parametric functions. Also, posterior predictive distributions for values of an unobserved batch cumulative weight fraction function or an observed vector of weight fractions from a new batch are possible.

5.1 An Example

Sieving data from plastic-bonded explosive (PBX) 9501 powder collected at the Mason & Hanger-Silas Mason Co., Inc., Pantex Plant previously studied by Huckett and Wendelberger (2002) were obtained from scientists at Los Alamos National Laboratory. These represent powder samples obtained from 6 different batches of material, 2 specimens from each batch. Each sample was passed through a series of 21 sieves, producing measured weight fractions for $k = 22$ size intervals with $C_0 = 0$ and $C_{22} = \infty$. These data are represented in Figure 4 and are available in tabular form in the supplemental materials.

We learned from the subject matter scientists that the material represented in Figure 4 was in fact produced by (purposely) separately manufacturing "small" particles and "large" particles of the same basic substance and then mixing small and large particle lots. This reality makes lognormal modeling of particle sizes for the PBX powder implausible, and strongly suggests use of a hierarchical version of the mixture modeling of Section 4. So before finishing our analysis of these PBX data, we consider some generalities for hierarchical modeling in a mixture context, based on the development of Section 4.

5.2 Hierarchical Analysis in a Mixture Setting

We consider the special case of hierarchical modeling for mixture lots, where one assumes that there is a single fixed (but unknown) relationship between size and weight operating for both particle types and across all lots. That is, once again we consider the parameters in display (12) to all be fixed unknown parameters that do not vary with lot. For the Lwin data this condition was employed for sake of simplicity, here it is warranted by the process which created the material. On the other hand, assume that the parameters μ_{s1}, σ_{s1}, μ_{s2} and σ_{s2} are all lot-specific, varying according to "process distributions." As with the Lwin data set

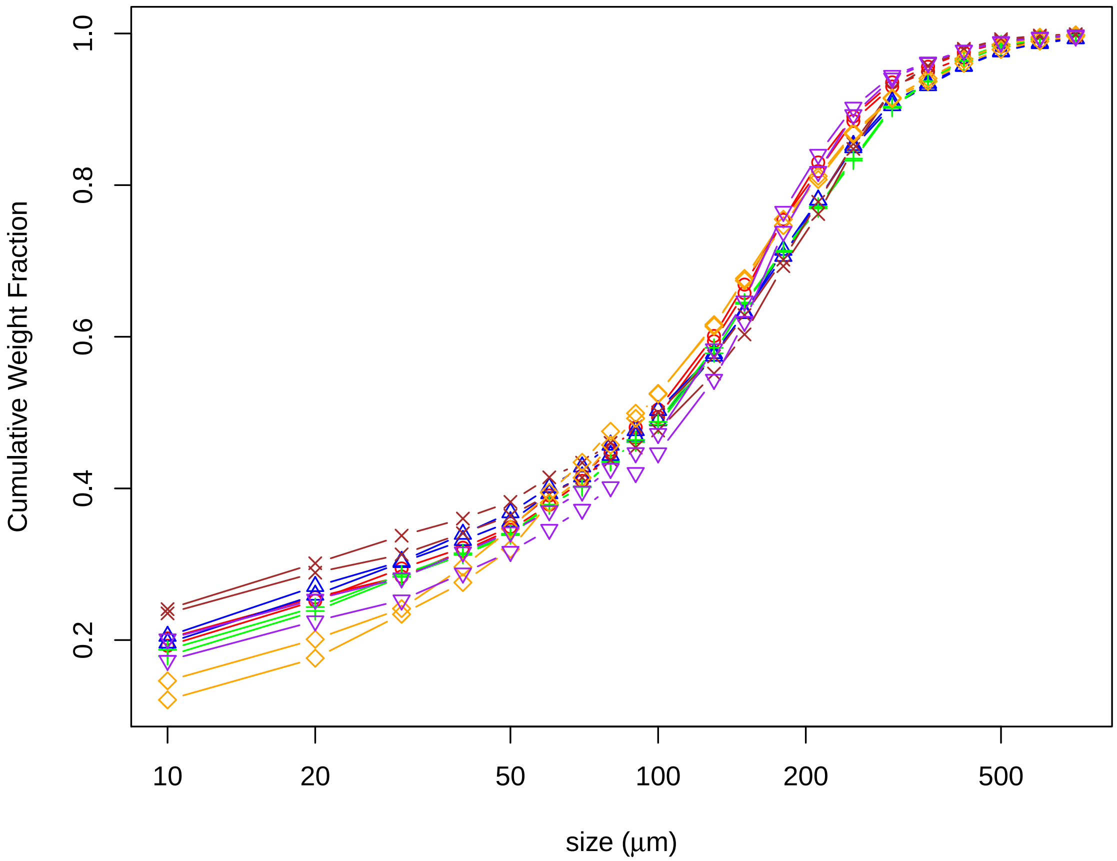

Figure 4: 12 vectors of cumulative weight fractions from the PBX data.

(12) facilitates the use of the mixture cumulative weight covariance functions provided in display (13) and (14).

In the present mixture context, a generalization of the structure suggested in Section 4 is this. Continue to abbreviate as $\mu_s^* = \mu_s + \eta\sigma_s^2$. Without the mixture assumption, this value locates the "center" of the cumulative weight, and as such is more directly interpretable than is μ_s. One might assume that across batches

$$\begin{pmatrix} \mu^*_{s1j} \\ \mu^*_{s2j} \end{pmatrix} \overset{iid}{\sim} \text{truncated-BVN}\left(\begin{pmatrix} \mu^*_{s\text{process1}} \\ \mu^*_{s\text{process2}} \end{pmatrix}, \mathbf{diag}(\lambda_1^2, \lambda_2^2) \right)$$

where the truncation is to the part of $\mathbb{R}^2$ where $\mu_{s1j} < \mu_{s2j}$, independent of $\sigma^2_{s1j} \overset{iid}{\sim}$ inv-$\Gamma(\alpha_1, \beta_1)$, independent of $\sigma^2_{s2j} \overset{iid}{\sim}$ inv-$\Gamma(\alpha_2, \beta_2)$, independent of $\rho \overset{iid}{\sim}$ Beta(α, β).

Further one might adopt independent priors

$$\begin{pmatrix} \mu^*_{s\text{process1}} \\ \mu^*_{s\text{process2}} \end{pmatrix} \overset{iid}{\sim} \text{truncated-BVN}\left(\begin{pmatrix} \mu_{10} \\ \mu_{20} \end{pmatrix}, \mathbf{diag}(\sigma^2_{\text{process10}}, \sigma^2_{\text{process20}}) \right),$$

$\lambda_i^2 \sim$ inv-$\Gamma(a_0, b_0)$, $\alpha, \alpha_i, \beta, \beta_i \sim \text{Exp}(c_0)$, for $i = 1, 2$; $\eta \sim \text{U}(0, 3)$ and $\tau \sim \text{Exp}(\lambda_0)$, where all values with a 0 subscript are user-supplied constants. Then one can use a Gibbs or MH-within-Gibbs MCMC algorithm to get approximate posterior distributions of the model parameters, interesting parametric functions, and predictions of future "observations" of various types.

The values selected as prior distribution parameters for our application were $a_0 = 0.1$, $b_0 = 0.1$, $c_0 = 1000$ and $\sigma^2_{\text{process10}} = 10$ and$\sigma^2_{\text{process20}} = 1$ and values $\mu_{10} = 2.5$, $\mu_{20} = 5.0$, and $\lambda_0 = 0.01$. The values employed for $\sigma^2_{\text{process10}}$ and $\sigma^2_{\text{process20}}$ were intended to provide a relatively "flat" or non-informative prior for plausible locations of the batch-specific cumulative weight functions. The means (μ_{10}, μ_{20}) were chosen to be $\log s$ values with average cumulative weight fractions around 0.25 and 0.75. Values of a_0 and b_0 provide a common approximately non-informative prior for variance parameters.

Table 3: Posterior inferences for mixture model parameters for the PBX Data.

Parameter	Median	95% Credible Interval	Parameter	Median	95% Credible Interval
$\mu^*_{s\text{process}1}$	2.223	(1.778, 2.671)	η	0.005	(0.000, 0.017)
$\mu_{\sigma_{s1}}$	1.872	(1.187, 2.554)	τ	0.002	(0.001, 0.003)
$\mu^*_{s\text{process}2}$	5.041	(4.711, 5.249)	μ_ρ	0.653	(0.585, 0.714)
$\mu_{\sigma_{s2}}$	0.378	(0.320, 0.442)			

Using these constants, Bayes analyses based on logs of weight fraction ratios were made. Posterior inferences for model parameters are in Table 3 (where, for $i = 1, 2$, $\mu_{\sigma_{si}} = \beta_i/(\alpha_i - 1)$, and $\mu_\rho = \alpha/(\alpha + \beta)$). Notice in particular that posterior inferences for $\mu_{\sigma 1}$ and $\mu_{\sigma 2}$ are clearly different and provide quantitative confirmation that the PBX material is composed of a mixture of two particle types. It seems that (on a log scale) small particles are less consistent in size that are big ones. We note as well that the very small value for η indicated by this analysis probably closes the possibility of a sensible physical interpretation of this model parameter in terms of shape properties of PBX particles. The parameter is simply a partial identifier for a useful member of a plausible parametric family of multivariate normal models for weight fractions or log ratios of the same.

Figure 5 shows the original 12 cumulative weight fraction functions, 6 fitted batch-specific functions $CW(s)_j$, and 95% prediction limits for both values of $CW(s)_{\text{new}}$ and cumulative $\boldsymbol{p}_{\text{new}}$. $CW(s)_j$ are obtained by plugging posterior medians into form (13), $CW(s)_{\text{new}}$ are the batch cumulative weight fractions as functions of size for a new batch taken from the process, and cumulative $\boldsymbol{p}_{\text{new}}$ are empirical cumulative weight fractions for a single specimen from a new batch (in some sense an observable version of $CW(s)_{\text{new}}$). The fact that there is not *much* difference between these intervals is consistent with the fact that the major part of

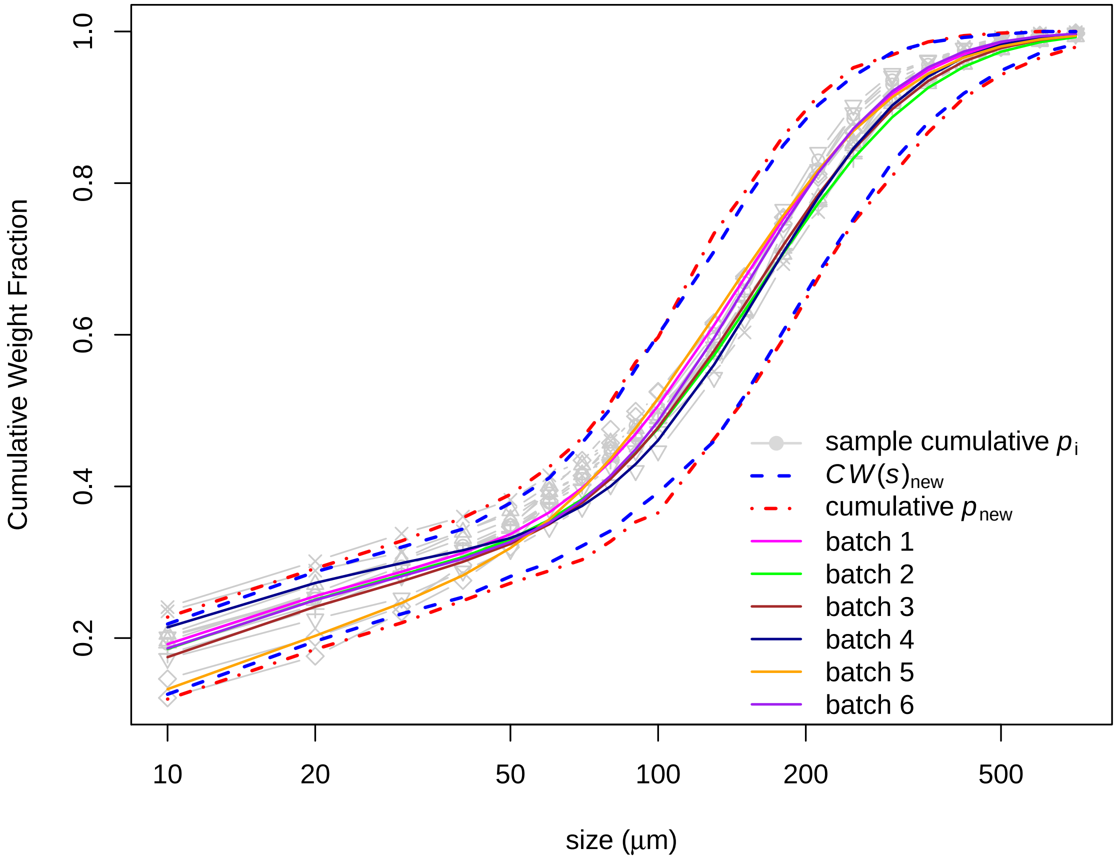

Figure 5: Fit for the PBX data.

the variation seen in the data of Figure 4 is between-batch variation, not variation between specimens for a given batch.

The graphic suggests that the models used here are flexible enough to adequately fit the existing lot cumulative weight fraction functions and to quantify process variation at small log particle sizes. However, it seems that process variation at large particle sizes is overstated by the present fitting. Therefore, future research could involve model generalizations that allow a bit more flexibility in modeling variability. This could be carried out for example, by allowing σ_s^2 to depend on $\log(s)$. This shortcoming admitted, it does seem like the analyses provide relatively transparent and straightforward methods of inference and prediction that are of subject matter importance.

6 Conclusions

The Bayes analyses presented in this article have the virtues of employing a very popular and standard form for $CW(s)$ and generalizations thereof, and defensible patterned forms for covariance matrices of observables. They further enable natural hierarchical and mixture modeling of weight fraction vectors. They provide rational/principled ways of making inferences on parametric functions that have subject matter meaning, and provide defensible predictions. Our graphics point out some aspects of less than perfect model fit in the particular examples employed here and the benefit of incorporating the added flexibility afforded by the mixture generalization. A reasonable extension for further research would be to consider mixtures of $k > 2$ types of particles. This should provide an even more flexible methodology relative to the $k = 2$ particle mixture. However, the flexibility would come at a potentially large computation cost and the complications regarding identifiability will greatly increase. We remark also that the current covariance structures model only variation in specimen weight fractions due to sampling particles. Another generalization that we have considered is to allow for measurement noise in the determination of weights. Ultimately however, what has been presented here is the fundamental basis of a new flexible and effective methodology for the analysis of particle sieving studies.

7 Supplementary Materials

Materials available on-line are

1) SupplementaryMaterial.pdf, a document containing a listing of the data sets and more details regarding the MCMC algorithms,

2) LwinAnalysis, a zipped folder contaning `C` and `R` code that enables the analysis of the Lwin data example, and

3) PBXanalysis, a zipped folder containing `C` and `R` code that enables the analysis of the PBX data example.

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Boca Raton: Chapman and Hall.

Allen, T. (2003), *Powder sampling and particle size determination*, Amsterdam: Elsevier B.V.

Dalby, R. N. and Byron, P. R. (1988), "Comparison of Output Particle Size Distributions from Pressurized Aerosols Formulated as Solutions or Suspensions," *Pharmaceutical Research*, 5, 36–39.

Duncan, A. J. (1962), "Bulk Sampling: Problems and Lines of Attack," *Technometrics*, 4, 319–344.

Gy, P. M. (1992), *Sampling of Heterogeneous and Dynamic Material Systems: Theories of Heterogeneity, Sampling, and Homogenizing*, Amesterdam: Elsevier.

Huckett, J. C. and Wendelberger, J. R. (2002), "Comparison of particle size data for PBX9501, LA-UR-5710," Tech. rep., Los Alamos National Laboratory.

Leyva, N. (2006), "Statistical Inference for Particle Systems from Sieving Studies," Ph.D. thesis, Department of Statistics, Iowa State University.

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

Maricq, M. M., Podsiadlik, D. H., and Chase, R. E. (1999), “Gasoline Vehicle Particle Size Distributions: Comparison of Steady Ste, FTP, and UQ06 Measurements,” *Environmental Science and Technology*, 33, 2007–2015.

Pitard, F. F. (1993), *Pierre Gy's Sampling Theory and Sampling Practice: Heterogeneity, Sampling Correctness, and Statistical Process Control*, Boca Raton: CRC Press, 2nd ed.

Scheaffer, R. (1969), “Sampling Mixtures of Multi-sized Particles: An Application of Renewal Theory,” *Technometrics*, 11, 285–298.

Smith, P. L. (2001), *A Primer for Sampling Solids, Liquids, and Gases: Based on the Seven Sampling Errors of Pierre Gy*, Philadelphia: ASA-SIAM.

Smyth, H. D. D. and Hickey, A. J. (2003), “Multimodal Particle Size Distributions Emitted from HFA-134a Solution Pressurized Meter-Dose Inhalers,” *AAPS PharmaSciTech*, 4.

Sommer, K. (1986), *Sampling of Powders and Bulk Materials*, Berlin, Heidelberg: Springer-Verlag.

Van der Bilt, A., Abbink, J. H., Mowlana, F., and Heath, M. R. (1993), “A Comparison Between Data Analysis Methods Concerning Particle Size Distributions Obtained by Mastication in Man,” *Archives of Oral Biology*, 38, 163–167.