Hindawi Publishing Corporation
Science and Technology of Nuclear Installations
Volume 2013, Article ID 705878, 18 pages
http://dx.doi.org/10.1155/2013/705878

Research Article

Estimating Alarm Thresholds for Process Monitoring Data under Different Assumptions about the Data Generating Mechanism

Tom Burr,[1] Michael S. Hamada,[1] John Howell,[2] Misha Skurikhin,[1] Larry Ticknor,[1] and Brian Weaver[1]

[1] *Statistical Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[2] *Mechanical Engineering Department, University of Glasgow, Glasgow G12 8QQ, UK*

Correspondence should be addressed to Tom Burr; tburr@lanl.gov

Received 7 December 2012; Revised 10 May 2013; Accepted 15 May 2013

Academic Editor: Michael F. Simpson

Process monitoring (PM) for nuclear safeguards sometimes requires estimation of thresholds corresponding to small false alarm rates. Threshold estimation dates to the 1920s with the Shewhart control chart; however, because possible new roles for PM are being evaluated in nuclear safeguards, it is timely to consider modern model selection options in the context of threshold estimation. One of the possible new PM roles involves PM residuals, where a residual is defined as residual = data − prediction. This paper reviews alarm threshold estimation, introduces model selection options, and considers a range of assumptions regarding the data-generating mechanism for PM residuals. Two PM examples from nuclear safeguards are included to motivate the need for alarm threshold estimation. The first example involves mixtures of probability distributions that arise in solution monitoring, which is a common type of PM. The second example involves periodic partial cleanout of in-process inventory, leading to challenging structure in the time series of PM residuals.

1. Introduction

Nuclear material accounting (NMA) is a component of nuclear safeguards, which are designed to deter or detect diversion of special nuclear material (SNM) from the fuel cycle to a weapons program. NMA consists of periodic, low frequency, comparisons of measured SNM inputs to measured SNM outputs, with adjustments for measured changes in inventory. Specifically, the residuals in NMA are the material balances defined as $\mathrm{MB} = T_{\mathrm{in}} + I_{\mathrm{begin}} - T_{\mathrm{out}} - I_{\mathrm{end}}$, where T is a transfer and I is an inventory.

Process monitoring (PM) is a relatively recent safeguards component. Although usually collected very frequently, PM data are often only an indirect measurement of the SNM and are typically used as a qualitative measure to supplement NMA or to support indirect estimation of difficult-to-measure inventory for NMA [1–3]. However, possible new roles for PM are being evaluated in nuclear safeguards. One of the possible new PM roles involves PM residuals, where a residual is defined as residual = data − prediction. One challenge in combining NMA and PM data is that PM residuals often have a probability distribution that cannot be adequately modeled by a normal (Gaussian) distribution but instead have an unknown distribution that must be inferred from training data.

We assume throughout that typical behavior of PM residuals, as defined by the probability distribution of the PM residuals, must be estimated using training data that is assumed to be free of loss (by diversion or innocent loss). Because of this assumption, it is helpful to consider settings with many applications other than safeguards that arise in standard statistical process control. In standard statistical process control settings, a quantitative attribute such as a manufactured part's dimension is measured and monitored.

For part i, let the true part dimension be T_i and let the measured part dimension be $M_i = T_i + R_i$, where R_i is a random measurement error. Assuming the manufacturing process is "in control," Phase I of statistical process control refers to the training period on anomaly-free data that is used to characterize the distribution of M_i, which varies because both T_i and R_i vary among parts. Phase I is followed by Phase II which refers to ongoing testing or monitoring for departure from Phase I behavior that has been statistically characterized.

A common test for departure from Phase I behavior is to estimate an alarm threshold such as is done in the basic Shewhart control chart [4], where continuous data is often assumed to have approximately a normal distribution and pass-fail data is assumed to follow a homogeneous Bernoulli distribution. Although threshold estimation with the Shewhart control chart (which alarms if the maximum observed data value exceeds the alarm limit) dates back to the 1920s, it is timely to consider modern model selection options in the context of threshold estimation. This paper reviews alarm threshold estimation, introduces model selection options to support threshold estimation, and considers a range of assumptions regarding the data-generation mechanism. Two examples from nuclear safeguards are included to motivate the need for alarm threshold estimation. The first example involves mixtures of probability distributions that arise in solution monitoring, which is a common type of PM. The second example involves periodic partial cleanout of in-process inventory, leading to challenging structure in the time series of PM residuals.

The paper is organized as follows. Section 2 provides additional background and a brief literature review. Section 3 describes the specific cases considered and gives numerical examples. The cases are defined by assumptions made about the data-generating mechanism for the monitored quantities, which in our context are the PM residuals. Section 4 gives the two PM examples from safeguards. Section 5 is a summary.

2. Background and Literature Review

Phase I training as used in many quality control applications often has the luxury of very large sample size, such as 10^6 or more observations from a manufacturing step [4–9]. In the context of monitoring PM residuals, we seek to require as little Phase I training data as possible before monitoring for typical behavior begins. Therefore, the quality control literature that is most relevant for PM needs is that concerned with Phase I training data size requirements [4–27]. As one example, [7] considers the effect on estimated tail probabilities of estimation error in estimated parameters of assumed data distributions. As another example, [20] considers extreme tail probability estimation while making minimal assumptions about the distributional form of the tail behavior. References [4–27] are among relatively few quality control publications that have investigated the amount of Phase I training data required for accurate estimation of alarm limits to achieve a desired low false alarm probability α in Shewhart or other control charts. Estimation error can be expressed as error in the alarm limit or as error in the estimated false alarm probability.

Often in safeguards it is necessary to control the per-period false alarm rate. For example, if there are $n = 100$ observations per year and the application requires a false alarm probability of $\alpha = 0.01$ per year, then the Shewhart alarm rule considers the distribution of the maximum of x_1, $x_2, \ldots, x_n$. However, for simplicity here, we consider the alarm rate per data point rather than per period (see Section 3.1.3).

This paper focuses on the error in the estimated false alarm probability for several types of assumed data generating mechanisms, including single-family parametric models such as the normal and log-normal and mixtures of single-family parametric models. Specifically, we start by assuming that the individual data points $x_1, x_2, \ldots, x_n$ are independently and identically (iid) distributed as $N(\mu, \sigma^2)$, the normal distribution with mean μ and variance σ^2. Then the only inference task is to estimate μ and σ in order to estimate the alarm threshold T so that the probability of a data point being at or above T is some small false alarm probability α such as 0.01. That is, we want to estimate T so that $p = P(x_i \geq T) = \alpha$. We use the symbol α when a small probability p refers to the false alarm probability during Phase II monitoring. Alternatively, to estimate T, we might assume almost nothing about the distribution of the data points x_i and use a nonparametric alternative such as a weighted average of the sorted data values which are also known as the sample quantiles. For other assumptions about the data, the inference task will change, as we demonstrate in Section 3.

One conclusion of this paper is that a rough guide for the required training data size n for accurate quantile estimation is that $n \geq 100$. Suppose $n = 100$ and we want to estimate the 0.999 quantile, so $\alpha = 0.001$. Let $x_{(1)}, x_{(2)}, \ldots, x_{(100)}$ denote the sorted values. A reasonable estimate is some type of weighted average $a_1 x_{(99)} + a_2 x_{(100)}$ of the two largest values, where $a_1 + a_2 = 1$. There must be some type of modeling to select the weights a_1 and a_2. One type of modeling is described in paragraph three of this section, in which a parametric form $N(\mu, \sigma^2)$ for x_i is assumed. Other types of modeling are described in Section 3.

3. Cases Considered Regarding the Data-Generation Mechanism

This section examines the amount of training data required for accurate estimation of alarm limits for a range of assumptions regarding the data generation mechanism. The main question that we address is as follows: what is the behavior of the estimation error (relative and absolute) in $\hat{p}$, the estimate of the probability of a data point being above the threshold T as a function of sample size n under various assumptions about the data generation mechanism and under various estimation approaches.

For this question regarding estimation error in $\hat{p}$, we consider the following Cases (a)–(f). In each case we estimate a threshold T for a desired α. We summarize the behavior of $\hat{p}$ as an estimate of p in terms of root mean squared error

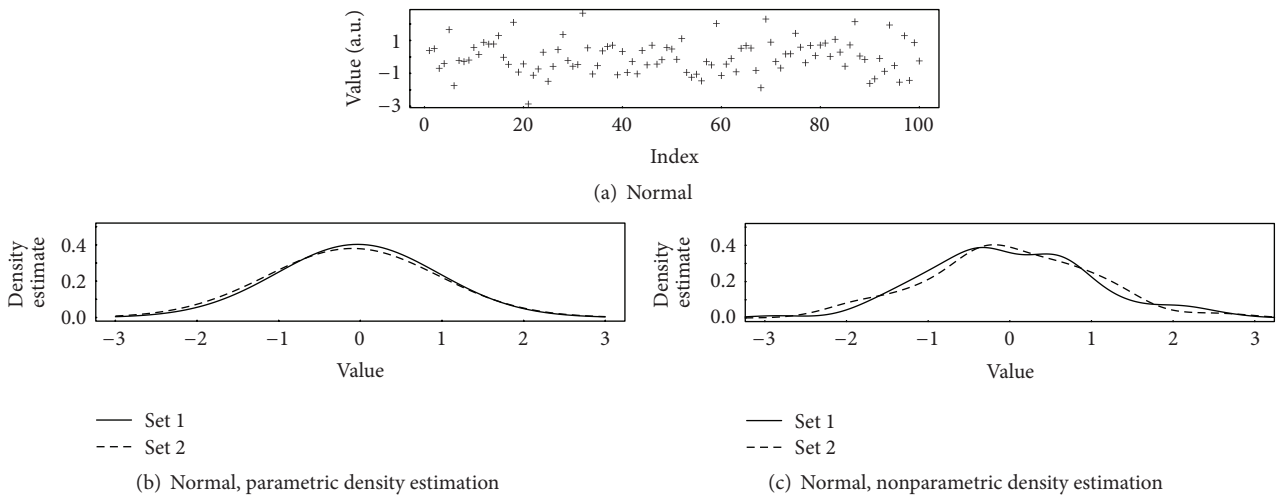


FIGURE 1: (a) 100 transformed observations $y_i = (x_i - \overline{x})/\widehat{\sigma}$, in arbitrary units (au), with x_i simulated as $x_i \sim$ iid $N(\mu, \sigma^2)$; (b) the estimated probability density using parametric density estimation from two sets of 100 simulated and transformed observations as in (a); (c) same as (b) but using nonparametric density estimation.

(RMSE), with MSE = $E\{(\widehat{p} - p)^2\}$, where E denotes expected value with respect to the distribution of the x_i, and RMSE = $\sqrt{E\{(\widehat{p} - p)^2\}}$.

(a) Assume the x_i are from a single parametric model. First, the x_i are generated from a $N(\mu, \sigma^2)$ distribution, and we assume the x_i are iid $N(\mu, \sigma^2)$ and estimate μ and σ^2. Next, we generate the x_i as iid from distributions other than the normal, but we incorrectly assume normality and estimate the parameters of the assumed distribution in order to calculate T for a desired α. If instead we assume (correctly) the same distribution as that used to generate the data, then we use the correct distribution to calculate an estimate of T with estimated parameters and then estimate p.

(b) Assume the x_i are a mixture of a known number of normals and we estimate the mixture means and variances and relative frequencies as a way to estimate p.

(c) Assume the x_i are a mixture of an unknown number of normals, and as in (b) we estimate the mixture means and variances and relative frequencies as a way to estimate p.

(d) Assume the x_i are a mixture of an unknown number of unknown distributions.

(e) Assume the x_i are iid from some known distribution to be discovered using model selection.

(f) Assume nothing about the distribution of x_i. Evaluate density estimation [28] and nonparametric quantile estimation [29].

3.1. Case (a): Parametric Modeling

3.1.1. Normal Data. Figure 1(a) plots a time series of $n = 100$ transformed observations $y_i = (x_i - \overline{x})/\widehat{\sigma}$. Here, x_i is simulated as $x_i \sim$ iid $N(\mu, \sigma^2)$, $\overline{x} = \widehat{\mu}$ is the usual sample mean, and $\widehat{\sigma} = \sqrt{\sum_{i=1}^{n}((x_i - \overline{x})^2/(n-1))}$ is the usual sample standard deviation. The "~" is standard notation for "is distributed as." Figures 1(b) and 1(c) plot the estimated probability distribution for the same 100 y_i values as in Figure 1(a), and also for a second set of 100 y_i values to check for consistency between two sets of 100 simulated values. Figure 1(b) uses parametric density estimation while Figure 1(c) uses nonparametric density estimation. In Figure 1(b), $\overline{x} = \widehat{\mu}$ to estimate μ and $\widehat{\sigma}$ is used to estimate σ, so the two sets of 100 y_i values lead to quite similar density estimates based on the normal probability density $N(\widehat{\mu}_1, \widehat{\sigma}_1^2)$ in set 1 and $N(\widehat{\mu}_2, \widehat{\sigma}_2^2)$ in set 2. In Figure 1(c) we use nonparametric density estimation which is a type of smoothed histogram that does not assume we know the true probability distribution, so the two sets of estimated densities are more different than in Figure 1(b) (see Section 3.6.1).

In short, if we know the true parametric model and only need to estimate its parameters, then the RMSE will be relatively small even for small sample size n. Of course, one rarely knows the true parametric model, which is why we consider Case (b) in Section 3.2–Case (f) in Section 3.6 but include Case (a) in Section 3.1 for comparison and for comparison to other literature such as [9].

3.1.2. Example of Nonnormal Data. As an example of nonnormal data, Figure 2 is the same as Figure 1, but is for $x_i \sim$ iid gamma (shape = 1, rate = 0.1). Notice that for $n = 100$ observations, there is nonnegligible estimation error in the nonparametric estimate of the probability distribution (Figures 1(c) and 2(c)). However, as we will show, a rough rule of thumb is that $n = 100$ observations is adequate for estimation needs in PM, such as reasonably small RMSE in $\widehat{p}$ as an estimate of p. The rule of thumb is motivated by finding in our examples that either: (a) there is a very slow decrease

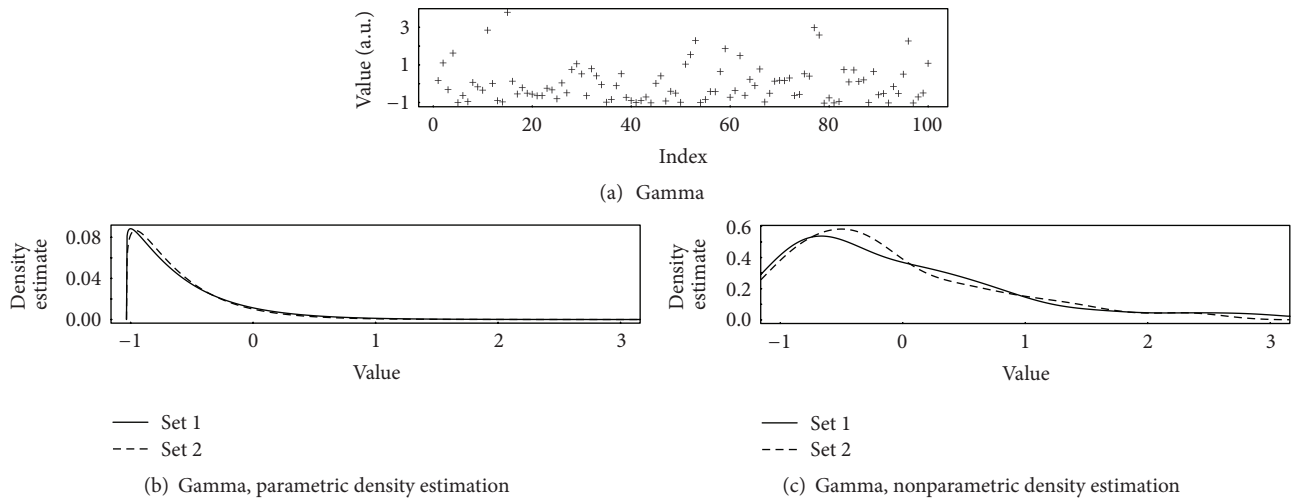


FIGURE 2: The same as Figure 1, but 100 transformed observations $y_i = (x_i - \overline{x})/\widehat{\sigma}$ from the gamma (shape = 1, rate = 0.1) distribution.

in the RMSE as n increases beyond 100, so increasing PM training data requirements beyond approximately $n = 100$ observations is probably not necessary, or (b) the RMSE is very small for some value of n near 100 or less.

3.1.3. The True Alarm Probability Compared to the Estimated Alarm Probability for Normal and Nonnormal Data. For normal data, Figure 3(a) plots the true alarm probability (see the next paragraph) versus the sample size using a nominal false alarm probability (FAP) of 0.001 to estimate the threshold T. We selected 0.001 as a small but reasonable FAP per PM data stream, anticipating that a per-year FAP over all NMA and PM streams should be 0.05 or less. In all our examples, we use either 0.001 or 0.025 as examples of small FAPs. As mentioned in Section 2, the desired FAP per PM stream will depend on the number of PM streams. And, the sampling frequency in a given PM stream will determine the FAP per sampling observation to maintain the desired per-year FAP. For example, if a PM stream has independent 10 samples per year, and a 0.01 FAP is allowed for that PM stream, then the per-sample FAP should be approximately 0.001.

Figure 3(a) was produced using simulation in R [30] as follows. As in Figure 1(a), generate data as $x_i \sim$ iid $N(\mu, \sigma^2)$. From these data estimate μ using $\widehat{\mu} = \overline{x}$ (the sample mean), and estimate σ^2 using $\widehat{\sigma}^2 = \sum_{i=1}^{n}((x_i - \overline{x})^2/(n-1))$ (the sample variance). Substitute $\widehat{\mu}$ for μ and $\widehat{\sigma}$ for σ in the normal probability cumulative distribution function to estimate the alarm threshold T corresponding to the 0.999 quantile. Specifically, $T_{0.001} = \mu + 3.09\sigma$ so $\widehat{T}_{0.001} = \widehat{\mu} + 3.09\widehat{\sigma}$. Notice in Figure 3(a) that the true alarm probability is considerably larger than the nominal (0.001) alarm probability marked by a horizontal line until approximately $n = 20$ or slightly larger. The "true" alarm probability was estimated with negligible estimation error by using 10^6 simulations. Throughout this paper we distinguish the true alarm probability (which is estimated with negligible estimation error by using many simulations) from the estimated alarm probability (whose estimation error is a key quantity that we study).

For nonnormal data in Figures 3(b)–3(d), x_i is generated as in Figure 3(a), but as iid from the lognormal, gamma and $t(2\text{df})$ distributions. In all four Figures 3(a)–3(d), we estimate the parameters of the normal distribution in order to estimate T for a desired α. That is, we assume (incorrectly for Figures 3(b)–3(d)) that the x_i are distributed as iid $N(\mu, \sigma^2)$ to illustrate the need for Case (b) in Section 3.2–Case (f) in Section 3.6. Therefore, we again use $\widehat{T}_{0.001} = \widehat{\mu} + 3.09\widehat{\sigma}$ in Figures 3(b)–3(d) for the lognormal generated from $\exp(x)$ with $x \sim N(\mu = 0, \sigma = 1)$ (so the mean of the lognormal is 1.65 and the variance is 4.67), the gamma (shape = 1, rate = 0.1), and the $t(2)$ distributions, respectively. Notice that for all three distributions, the true alarm rate goes below the nominal rate of 0.001 for large sample sizes. In other cases not shown (e.g., for the gamma (shape = 1, rate = 2) distribution), the true alarm rate is larger than the nominal rate for all sample sizes. If instead we correctly assume the same distribution as that used to generate the data, then we estimate parameters from the correct distribution to estimate p. For comparison, we return to this ideal situation in which we know the correct distribution in Section 3.4.

In addition to the true alarm probability, the estimation error in the alarm rate as measured, for example, by the RMSE is also of interest. The RMSE combines both bias (defined as the difference between the true alarm probability and the long-run average of the estimated probability) and variance in the well-known expression RMSE = bias^2 + variance [28].

Figure 4 plots the RMSE versus sample size for the same four distributions as in Figure 3(a), again assuming the true distribution is normal as described above, which is incorrect except for in Figure 4(a). The RMSE_{sim} was calculated across $n_{\text{sim}} = 10^4$ simulations using $\text{RMSE}_{\text{sim}} = \sqrt{\sum_{i=1}^{n_{\text{sim}}}(\widehat{p} - p)^2}$, where p is the true tail probability and $\widehat{p}$ is defined by using, for example, $\widehat{T} = \widehat{\mu} + 3.09\widehat{\sigma}$ as in Figure 3. Note that RMSE_{sim} approaches 0 as n increases (Figures 3(a) and 4(a)) as one

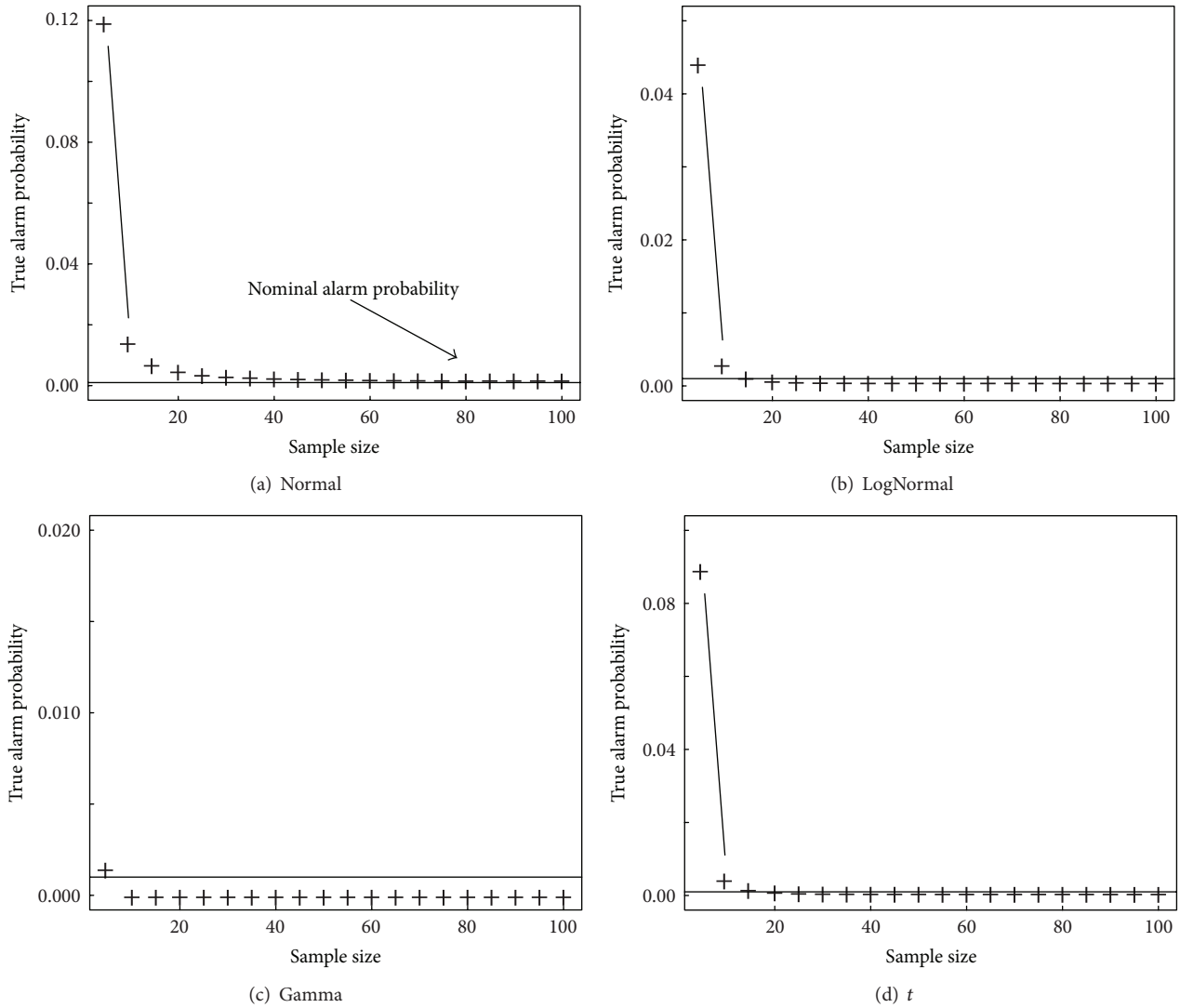


FIGURE 3: True alarm probability versus sample size (nominal alarm probability is 0.001) for (a) normal, (b) lognormal, (c) gamma (1, 0.1), and (d) $t(2)$. The true alarm probabilities were estimated by simulation of 10^4 observations in R, and results are repeatable to within ±0.0001.

would expect. Figure 4 includes a horizontal line at 10% of 0.001 (0.0001) for visual comparison.

Notice in Figure 4 that the RMSE does decrease as n increases, even when we incorrectly assume a normal distribution (Cases (b), (c), and (d)). Future work will investigate the tradeoff between estimator bias and variance in the RMSE in the context of assuming slightly wrong underlying distributions. That is, the RMSE could be acceptably low in Figures 4(b)–4(d) despite wrongly assuming that the true distribution is normal. However, the obvious bias in $\widehat{p}$ as an estimate of p does not vanish as n increases (see Figure 3), so it is unlikely to be acceptable to blindly assume PM data streams have a normal distribution. Therefore, we also consider Case (b) in Section 3.2–Case (f) in Section 3.6.

3.2. Case (b): Assume the x_i Are a Mixture of a Known Number of Normal Distributions and Estimate the Mixture Means and Variances and Relative Frequencies as a Way to Estimate p. In Case (b) we assume the x_i are a mixture of a known number of normal distributions, but we must infer which x_i belong to which mixture component. One tool to infer group membership is model based clustering as implemented in the `Mclust` function in R [30, 31]. Using `Mclust` to infer group membership, we estimated the RMSE versus sample size for a nominal alarm probability of $p = 0.001$ for the case of overlapping groups (see Figure 5(a)) and two well-separated groups (see Figure 5(b)). Figure 5 was generated by applying density estimation using the density function in R to 10^4 simulated values from the overlapping-group case and from the well-separated group case. Figure 6 shows the RMSE in the case of well-separated and overlapping groups. For comparison, notice from Figure 6 that the RMSE is smaller using `Mclust` than using a nonparametric option (based on the sample quantiles as described in Section 3.7), and that the RMSE is nearly the same whether the groups are overlapping or well separated.

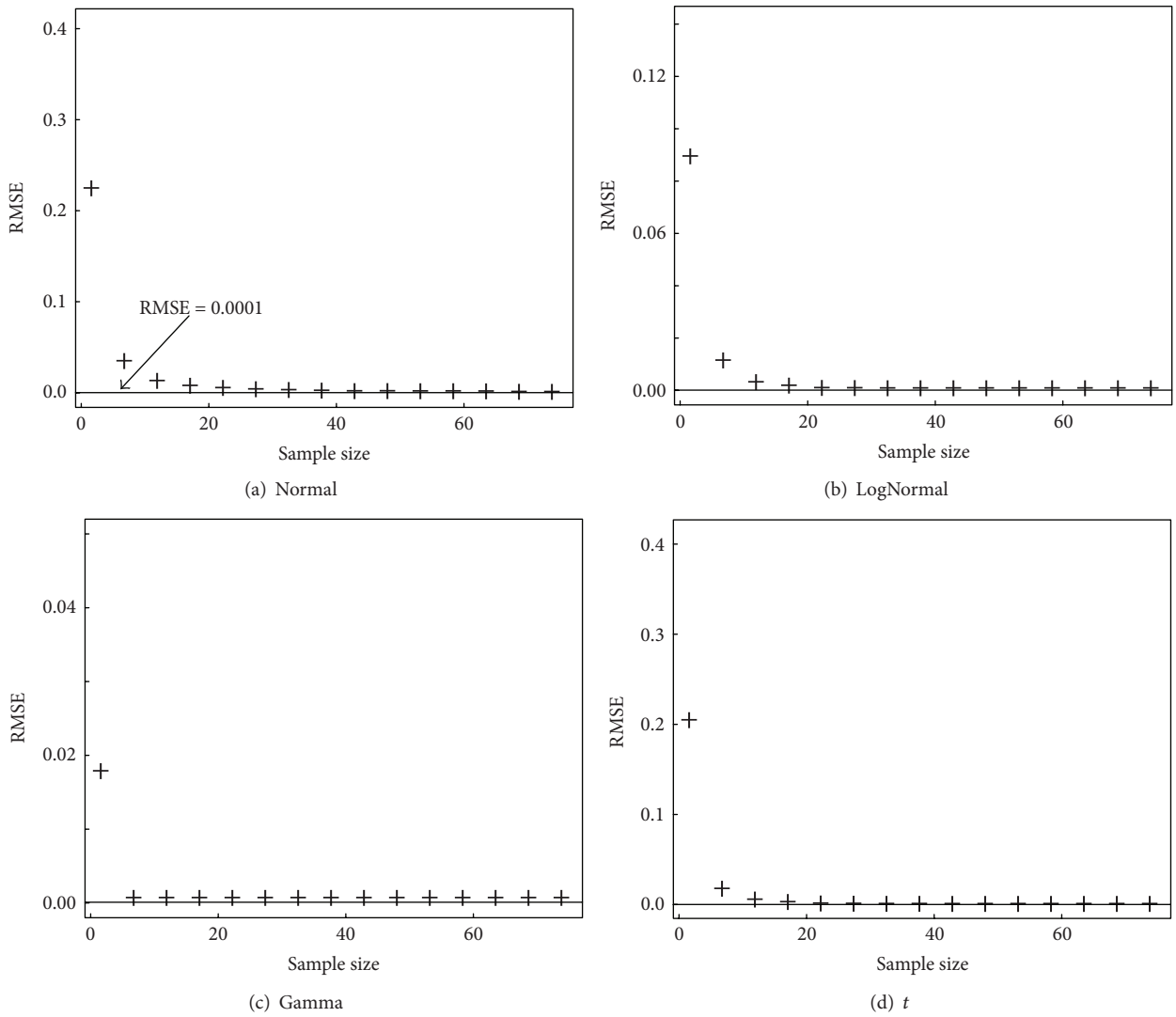
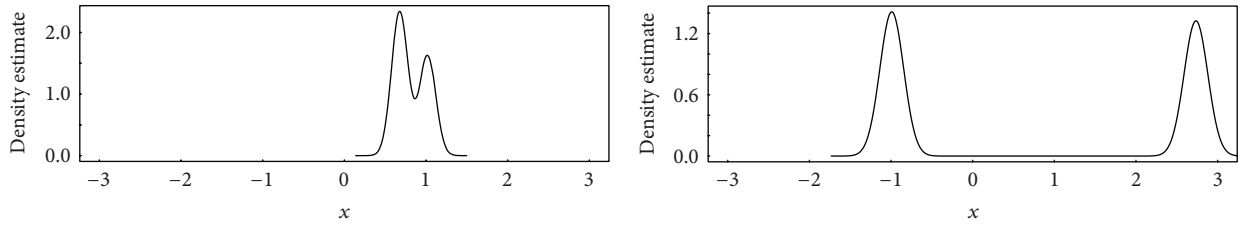


Figure 4: log (RMSE) versus sample size (nominal alarm probability is 0.001) for (a) normal, (b) lognormal, (c) gamma (1, 0.01), and (d) $t(2)$ for two estimation options labeled 1 and 2, respectively. The true RMSEs were estimated by simulation of 10^4 observations in R, and results are repeatable to within ±0.0001.

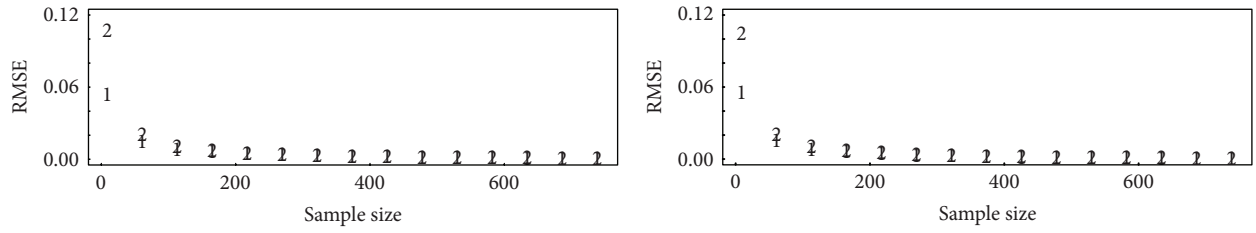
To explain our approach for estimating alarm thresholds using `Mclust`, we can simply consider the case where the means μ_i differ among components, but the standard deviations σ_i are the same in each component, denoted σ. The mean and standard deviation of the mixture are then $\mu_{\text{mix}} = \sum_{i=1}^{N_{\text{comp}}} \pi_i \mu_i$ and $\sigma^2_{\text{mix}} = \sigma^2 + \sum_{i=1}^{N_{\text{comp}}} \pi_i(\mu_i^2 - \mu^2_{\text{mix}})$, where π_i is the relative frequency of component i. Our main interest is in the probabilities of exceeding specified thresholds, such as a multiple k of the standard deviation, where k is usually in the range of approximately 2 to 4. It can be shown by straightforward calculation that when x_i are iid from a mixture of normal distributions, when testing only for large positive outliers as we do in all our examples, then $P(x - \mu_{\text{mix}} > k\sigma_{\text{mix}}) = \sum_{i=1}^{N_{\text{comp}}} \pi_{\text{i}}(1 - \varphi((\mu_{\text{mix}} - \mu_{\text{i}})/\sigma + k(\sigma_{\text{mix}}/\sigma)))$, where φ is the standard normal density. This expression for $P(x - \mu_{\text{mix}} > k\sigma_{\text{mix}})$ is used to estimate the threshold T using N_{comp} and using estimates provided by `Mclust` of the relative frequency π_i, the means μ_i, and standard deviations σ_i of each component.

For many mixtures, these tail probabilities are smaller than those of the corresponding reference distribution, which is a single-component normal having the same standard deviation as the mixture, σ_{mix}. Therefore, higher probabilities of mean-centered values exceeding $k\sigma$ are not necessarily expected. However, for other mixtures, particularly those having very unequal π_i, the tails are fatter (giving larger probabilities to extreme values) than the reference normal. For example, consider the random variable x arising from a mixture consisting of three components with $\pi_1 = 0.0833$, $\pi_2 = 0.833$, and $\pi_3 = 0.0833$; $\mu_1 = -3$, $\mu_2 = 0$, and $\mu_3 = 5$; $\sigma = 1$. For this random variable x we have $P(x - \mu_{\text{mix}} > k\sigma_{\text{mix}}) = 0.088$, 0.013, and 0.0001 for $k = 2, 3$, and 4, respectively. The corresponding probabilities for the single-component reference normal are 0.046, 0.003, and 0.00006, which are



(a) Overlapping groups (b) Well-separated groups

Figure 5: Overlapping (a) and well-separated (b) groups.



(1) Fit mixture
(2) Nonparametric

(a) Overlapping groups (b) Well-separated groups

Figure 6: The RMSE versus sample size for (a) overlapping groups and (b) well-separated groups for a known number of normals using the "fit mixture" option and using the quantile-based nonparametric option described in Section 3.7.

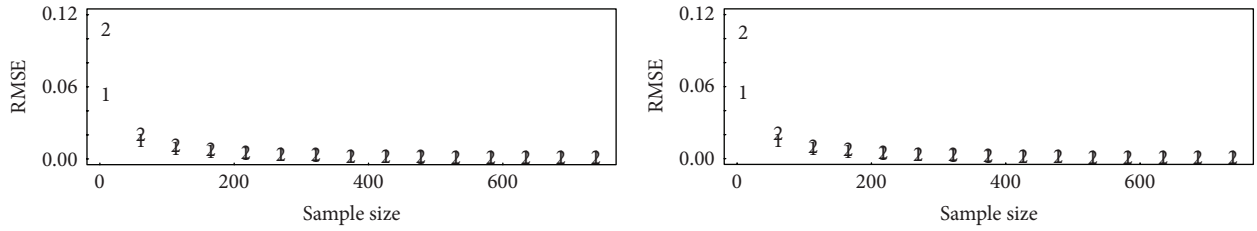
significantly smaller, indicating that this particular mixture has fatter-than-normal tails. On the other hand, consider the random variable X arising from a symmetric mixture consisting of three components having $\pi_1 = 0.25$, $\pi_2 = 0.5$, and $\pi_3 = 0.25$ and $\mu_1 = -3$, $\mu_2 = 0$, and $\mu_3 = 3$ with $\sigma = 1$. Then $P(x - \mu_{\text{mix}} > k\sigma_{\text{mix}}) = 0.023$, 0.000014, and 4.4×10^{-11} for $k = 2, 3$, and 4, respectively, indicating that this mixture has thinner-than-normal tails.

3.3. Case (c): Assume the x_i Are a Mixture of an Unknown Number of Normal Distributions, and as in (b), Estimate the Mixture Means and Variances and Relative Frequencies as a Way to Estimate p . In Case (c) we must estimate the number of components, unlike Case (b). We assume the x_i are a mixture of an unknown number of normal distributions, so we must infer which x_i belong to which mixture component and how many mixture components are present. As mentioned in Section 3.2, one tool to infer group membership is model-based clustering as implemented in the `Mclust` function in `R` [31]. As in Case (b), using `Mclust`, we estimated the RMSE versus sample size for a nominal alarm probability of $p = 0.001$.

As in Figures 6 and 7 plots the RMSE versus sample size for overlapping (Figure 7(a)) and well-separated (Figure 7(b)) groups. Using the Bayesian information criterion option in `Mclust` to choose the number of groups, the estimated probability of inferring the correct number of groups (2) when the candidate number of groups is any number from 1 to 10 is small to moderate (0.3 to 0.5) for overlapping groups and large (0.8 or higher) for well-separated groups. In comparing Figures 6 to 7, we note that for the examples considered, the RMSE is approximately the same whether we know there are 2 groups (Figure 6) or whether we estimate the number of groups (Figure 7).

3.4. Case (d): Assume the x_i Are iid from an Unknown Distribution (not a Mixture but a Single-Component Distribution) to Be Discovered Using Model Selection. First, assume we know the correct distribution and use the same four distributions (normal, lognormal, gamma (1,.1), $t(2)$) as in Figure 3. In this case, using `fitdistr` in `R` (which uses maximum likelihood fitting) to estimate the parameters of the known distribution, the bias $\widehat{p}$ is negligible for any of the four distributions. That is, if we are fortunate enough to correctly estimate or know the true distribution rather than blindly assume a normal, then the bias and RMSE in $\widehat{p}$ are approximately the same as shown in the "generate normal, assume normal" case shown in Figures 3(a) and 4(a).

Next, and more relevant for applications, assume the generating distribution is unknown, but one could use features of the data to select a distribution. Data features to choose a distribution could be the observed sample quantiles, or a quantitative assessment of a quantile-quantile plot that plots expected quantiles assuming a candidate data distribution versus the observed quantiles, or the raw data using model selection options such as the Bayesian information criterion (BIC). Here the BIC is defined as $\text{BIC} = 2\log(\text{ML}) - k\ln(n)$, [28, 31] where ML is the maximum value of the likelihood, k is the number of model parameters, and n is the sample size. Models having large BIC values are preferred. We note that



(1) Fit mixture
(2) Nonparametric

(a) Overlapping groups

(b) Well-separated groups

Figure 7: The RMSE versus sample size for (a) well-separated and (b) overlapping groups for a mixture of an unknown number of normals (same case as Figure 6, except the number of groups is unknown so it must be estimated).

the BIC is sometimes defined as −1 times the BIC definition we use, in which case models having small BIC values are preferred. We used the BIC as provided by `Mclust` in Case (c), and we use the BIC in the next example and in other examples to follow.

Figure 8 plots the RMSE versus sample size for the following experiment. For each of 1000 simulations, let the true likelihood be randomly selected with equal probability to be normal, t, lognormal, or gamma. Use the BIC to infer the likelihood and use estimated parameters of the chosen (inferred) likelihood to estimate, for example, the 0.0975 quantile, corresponding to a $p = 0.025$ tail area probability. Figure 8 shows that in this small experiment, large sample sizes are required in order for BIC-based model selection to outperform option 2, which blindly assumes the normal likelihood, or option 3 which blindly assumes a t distribution.

A second set of 1000 simulations shows that these RMSE results are repeatable to within 10% relative. In addition, it is of interest to assess how often the BIC approach leads to correct model selection in 1000 simulations, which is illustrated in Table 1. The three entries in each cell are the estimated probabilities of inferring the indicated distribution in the column when the true distribution is given in the row. The three cell entries correspond to sample sizes of $n = 25, 50$, and 1000, respectively. For example, when the true likelihood is the normal distribution, there is high probability, 0.86, 0.90, and 0.99, of correctly inferring the normal distribution for $n = 25$, 50, and 1000, respectively.

The RMSE results in Figure 8 are for the specific small experiment described. If a different collection of candidate likelihoods are used, we suggest using simulation to assess whether the BIC-based approach to choose a distribution is likely to lead to smaller RMSE than blindly assuming a particular likelihood such as the normal or t.

3.5. Case (e): Assume the x_i Are a Mixture of an Unknown Number of Unknown Distributions. In case (e) we assume the x_i are a mixture of an unknown number of unknown distributions, so we must infer which x_i belong to which mixture component and how many mixture components are present. One tool to infer group membership is model based clustering as implemented in the `mixtools` package in R [32].

Using `npEM` (nonparametric estimation maximization) in `mixtools`, we estimated the number of components in three cases (each case has 100 observations): a single-component normal, a mixture of two overlapping and equal proportion component (50 observations in each component) normal distributions as in Figure 6(b), and a mixture of two well-separated normal distributions (50 observations in each components) as in Figure 6(a). Figure 9 compares the BIC values from `npEM` (which does not assume a distributional form for the component) to the BIC values from `Mclust` (which assumes that each component has a normal distribution) for the three cases. Because the components are normal distributions in all three cases, we expect `Mclust` results to be better than `npEM` results. However, we also expected `npEM` results to do reasonably well even when the underlying distributions are all normal. Notice however in Figure 9(d) (for the case of two overlapping normal distributions) that `npEM` predicts 9 or 10 components rather than 2 components.

In repeated experiments such as this, `npEM` performs very erratically in the case of overlapping components. Apparently, using density estimation (see Section 3.6) in the manner that `npEM` does is not effective for the case of overlapping normal distributions. Of course `Mclust` is tuned to work best when the component distributions are normal, so we repeated the above experiment in which the true number of components is 1, 2 overlapping, and 2 well separated, but each component was lognormal. The estimated number of components was 2, 2, and 4, respectively for `npEM` and was 2, 3, and 9, respectively for Mclust, so neither `Mclust` nor `npEM` performed well, but `Mclust` did worse than `npEM`. The poor performance of `Mclust` is not surprising because of the lognormal distribution for each component, but `npEM` does not assume any particular distribution so its poor performance is disappointing. These experiments indicate that mixture fitting is difficult [33], and that `npEM` performs erratically for all sample sizes unless the groups are distinct and well separated.

Using either `npEM` or `Mclust` to infer the number of groups, we have a choice regarding how to estimate quantiles using the inferred groupings. For example, we could fit a distribution to each inferred component and follow options

Table 1: The estimated probability rounded to the nearest 0.01 of inferring the likelihood model using BIC. The inferred likelihood is in columns and the true likelihood is in rows, for sample sizes $n = 25$, 50, and 1000. The true likelihood is either normal, $t(2)$, lognormal, or gamma(1, 0.1) in the experiment.

True	Inferred			
	Normal	$t(2)$	Lognormal	Gamma(1, 0.1)
Normal	0.86, 0.90, 0.99	0.02, 0.07, 0	0, 0, 0	0.11, 0.03, 0
$t(2)$	0.81, 0.83, 0.98	0.06, 0.10, 0.02	0, 0, 0	0.13, 0.07, 0
Lognormal	0, 0, 0	0, 0, 0	0.64, 0.83, 1.0	0.36, 0.17, 0
Gamma(1, 0.1)	0, 0, 0	0, 0, 0	0.37, 0.24, 0.01	0.63, 0.76, 0.99

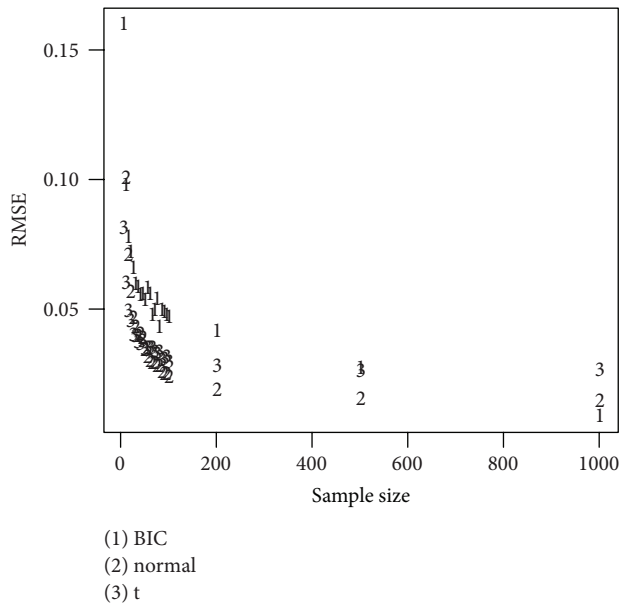


Figure 8: The RMSE versus sample size using the BIC to select the model, or blindly assuming the normal or t distribution.

described in Section 3.1.3 to estimate the tail behavior within each component. Because of the extremely slow run time and erratic performance of `npEM`, we do not experiment further with this option in this paper.

3.6. Case (f): Assume Almost Nothing about the Distribution of x_i Except That It Has Finite Moments of All Orders. In Case (f) we assume almost nothing about the distribution of x_i except that it has finite moments of all orders and consider a nonparametric ("distribution free") approach to quantile estimation. We note that the term "nonparametric," although well established in statistical literature, is somewhat misleading. The term "nonparametric" refers in this paper to the fact that the approach works for any distribution that has finite moments of all orders. All such distributions have parameters such as the mean and variance, but we follow convention and use the term "nonparametric."

Accurate nonparametric estimation of quantiles, particularly extreme quantiles, requires large n. Therefore, it is reasonable to consider whether there other options besides brute force nonparametric (sample quantiles) to estimate T. This subsection describes an option based on nonparametric density estimation and on empirical likelihood. A tail-behavior modeling option such as that in [20] will also be investigated in future work.

3.6.1. Density Estimation. The function `density` in `R` uses a kernel density estimation approach [28]. Most readers are familiar with histograms, which are crude density estimators. Improved density estimators essentially are smoothed histograms (as in Figures 1(c) and 2(c)). Typically, a density estimator at value x is given by $\widehat{f}(x) = (1/n)\sum_{i=1}^{n} K(x, x_i, h)$, where K is a symmetric "kernel" function such as the normal density function $K(x, \mu, \sigma) = (1/\sigma\sqrt{2\pi})\exp\{-(x-\mu)^2/2\sigma^2\}$ so $\widehat{f}(x) = (1/nh\sqrt{2\pi})\sum_{i=1}^{n}\exp\{-(x-x_i)^2/2h^2\}$. The estimate $\widehat{f}(x)$ can be used to estimate p for a candidate value of a quantile q simply by using $\widehat{p} = \int_{-\infty}^{q}\widehat{f}(x)dx$. The main technical challenge with kernel density estimation is choosing an effective bandwidth h [28] and cross validation as used in the function `density` in `R` is reasonably effective for bandwidth selection.

3.6.2. Empirical Likelihood. Empirical likelihood methods use likelihood methods but do not require a parametric family for the data. In the context of quantile estimation, smoothed versions of the empirical cumulative distribution function (which puts probability $1/n$ on each of n observations) are used with or without the sorted data $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. The versions that use the sorted data are extensions of the options available in the `quantile` function in `R`. We found that all 9 options in the `quantile` function give very similar RMSE results, and that all 9 options use weighted averages of the sample quantiles as described briefly in Section 2 and also in Section 3.7.

Motivated by empirical likelihood, we added a 10th option for nonparametric quantile estimate that uses a weighted average of all of $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ rather than a weighted average of the two sorted values $x_{(i)}, x_{(i+1)}$ that bracket the desired pth quantile such that $x_{(i)}/n \le p \le (x_{(i+1)}/n)x_{(i)}$. All 10 options give very similar results; however, if there is interest in providing a confidence interval for $\widehat{p}$, then [29] claims good accuracy (the nominal confidence interval behavior is close to the actual confidence interval behavior) with empirical likelihood [29].

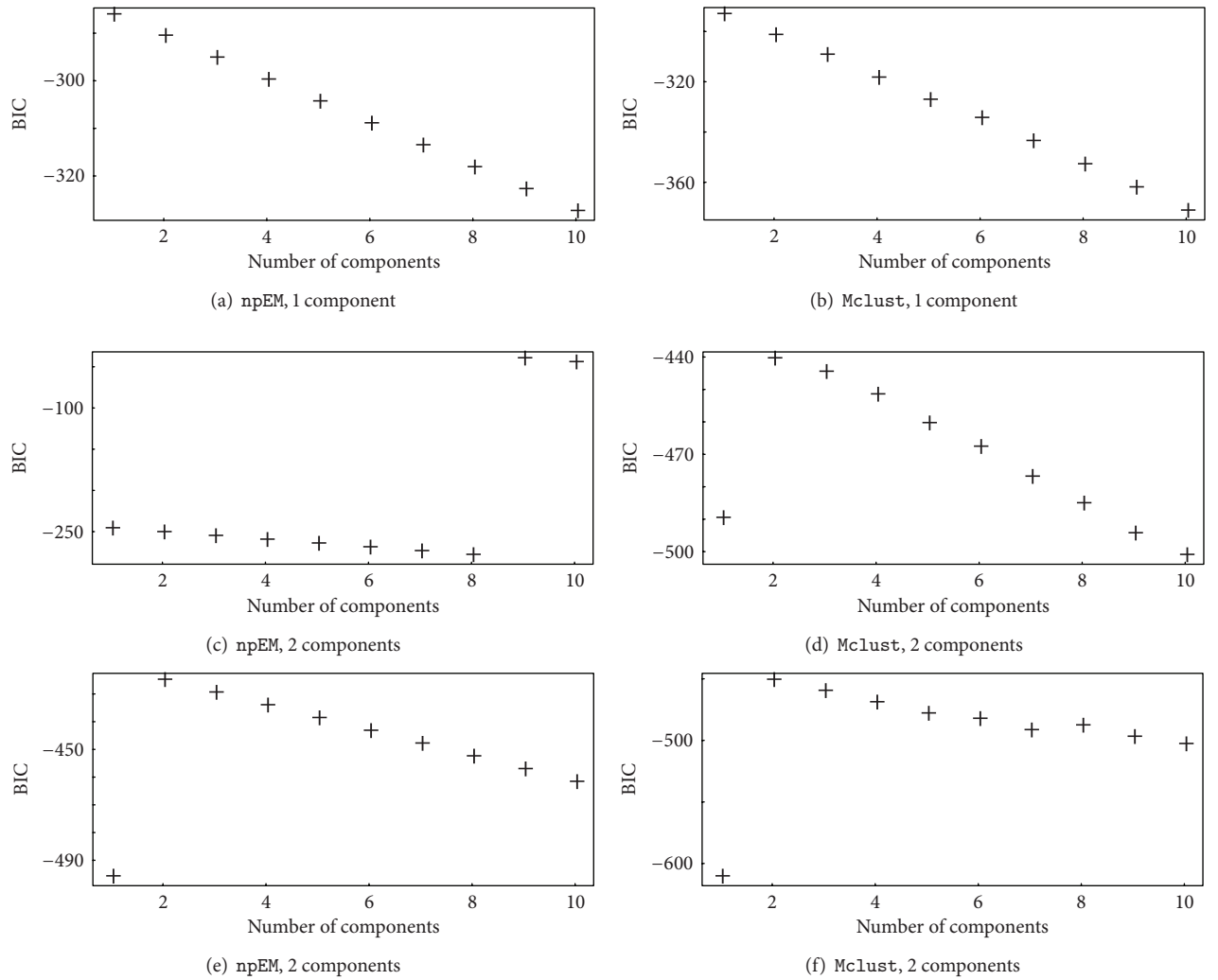
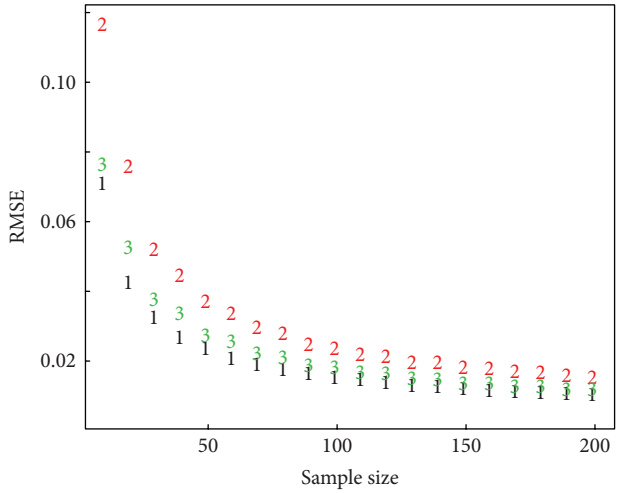


Figure 9: Comparison of npEM to Mclust for fitting: (a) single-component normal; (b) mixture of two overlapping normals as in Figure 6(b); (c) mixture of two well-separated normals as in Figure 6(a). In the application of both npEM and Mclust, the BIC is used to select the number of components, with the maximum BIC value corresponding to the chosen number of components.

3.7. Comparing Three Quantile Estimation Options for the 0.95, 0.99, 0.995, and 0.999 Quantiles. In this section we compare three of the presented quantile estimation options for four small false alarm probabilities (0.05, 0.01, 0.005, and 0.001). The three options are as follows: (1) assume a single-component normal (Section 3.1), (2) use a weighted average of the sample quantiles (Section 3.7.1 below), and (3) use density estimation (Section 3.6.1). For the sake of brevity here, we omit other options such as mixture fitting.

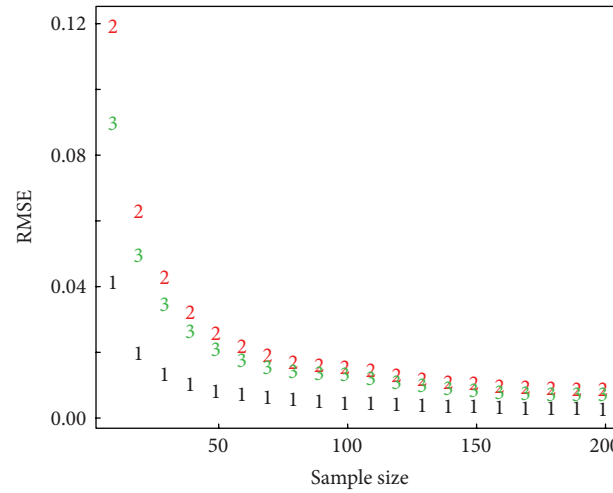
3.7.1. Using the Sample Quantiles. Section 2 described a nonparametric approach that uses the sample quantiles, which is robust to distributional assumptions but less efficient than option 1 if the true distribution is normal. To estimate the RMSE of $\widehat{p}$ for option 2 we used the `quantile` function to estimate the 0.999 quantile of the original simulated data. To estimate the true p corresponding to $\widehat{T}$, which is how often a data value would be above the estimated 0.999 quantile, we simulated 10^6 observations and tallied the number of times the simulated data exceeded the estimated quantile. Alternatively, to estimate p we could use the known true distribution in cases such as the $N(\mu, \sigma^2)$ for which integration is simple. The RMSE was then estimated as before, using 10^4 simulations for each evaluated sample size. There are many ways to estimate the 0.999 quantile and the `quantile` function in `R` implements the nine options described in [21] to estimate quantiles from data without explicitly assuming a parametric distribution. We experimented with all nine options available in the `quantile` function. In addition [29] considers weighted averages of the sample quantiles (see Section 3.6.2).

For option 2, we found almost no difference in average RMSE values among the nine `quantile` function options we tried (such as ordinary sample quantiles or linearly interpolating between sample quantiles) and report results here for option 4 in `quantile`, which linearly interpolates between sample quantiles.



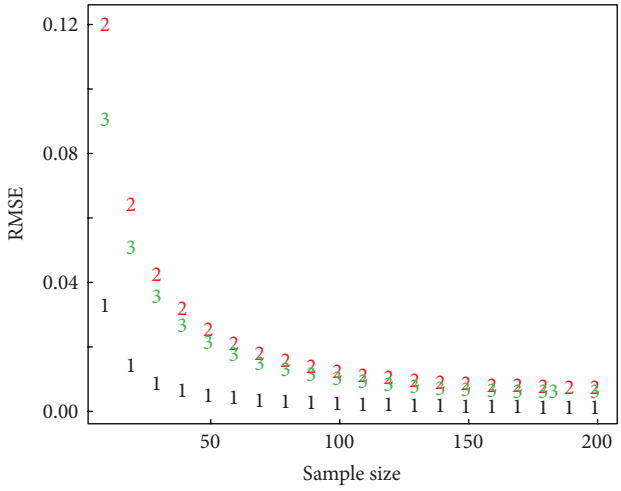
(1) Assume normal
(2) Quantile function
(3) Density estimation

(a) Estimate the 0.95 quantile



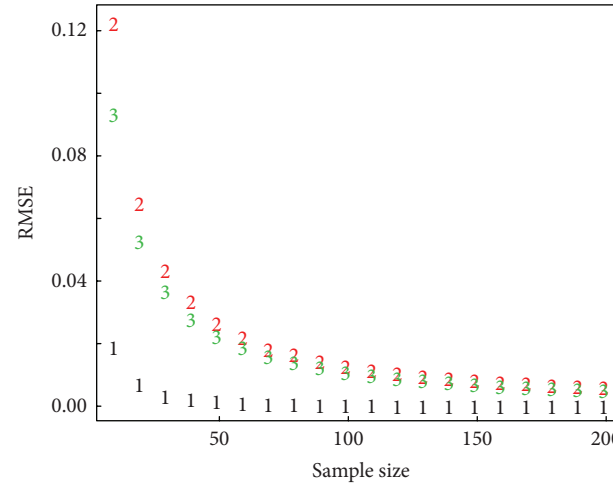
(1) Assume normal
(2) Quantile function
(3) Density estimation

(b) Estimate the 0.99 quantile



(1) Assume normal
(2) Quantile function
(3) Density estimation

(c) Estimate the 0.995 quantile



(1) Assume normal
(2) Quantile function
(3) Density estimation

(d) Estimate the 0.999 quantile

FIGURE 10: RMSE versus sample size assuming a normal distribution for three estimation options labeled 1–3. The true RMSEs were estimated by simulation of 10^4 observations in R, and results are repeatable to within ±0.001. Option 1 correctly assumes a normal distribution. Option 2 uses the `quantile` function in R. Option 3 is density estimation as discussed in Section 3.6.1.

3.7.2. RMSE Results for Options 1–3 in This Section. Figure 10 plots the RMSE in 10^4 realizations for sample sizes ranging from 5 to 200 for the case in which the true distribution is a single-component normal for false alarm probabilities of (a) 0.05, (b), 0.01, (c) 0.005, and (d) 0.001. We know that the "assume a single-component normal" is the best possible method, and we know that density estimation is nonparametric and therefore performs fairly well for a wide range of underlying true distributions. Therefore, for the case in Figure 10, we expect the RMSE for most other methods to lie between the RMSE of option 1 and option 3.

Figure 11 is the same as Figure 10, except the true distribution is a $t(2)$ distribution, so option 1 would not perform well, so we assumed (correctly) that the true distribution was known to be a $t(2)$ distribution. Notice in Figure 11 that we did not attempt to use the BIC to select a distribution (but see Case (d) in Section 3.4 where it appears that selecting a single-component distribution can be reasonably effective).

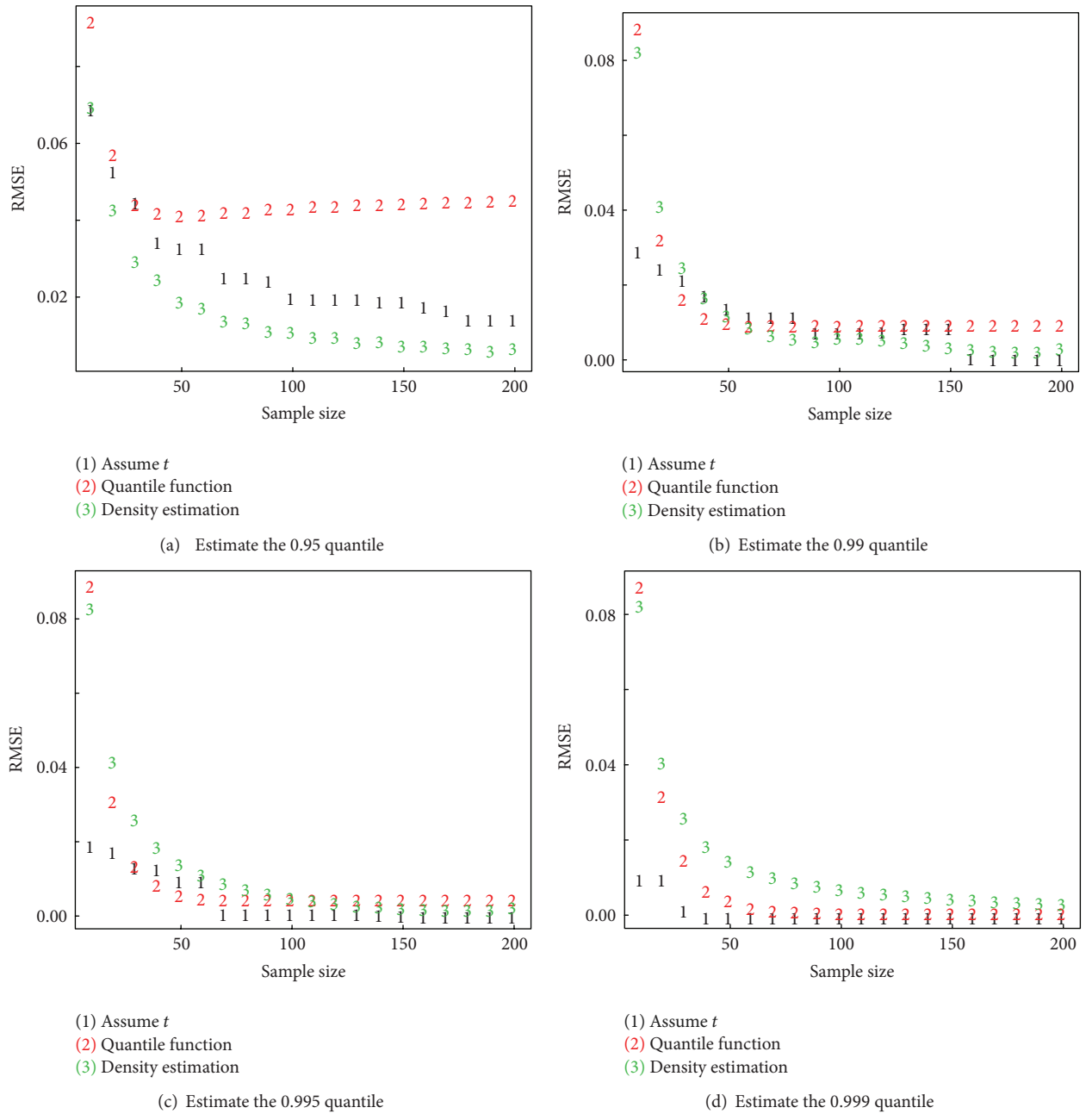


FIGURE 11: RMSE versus sample size assuming a normal distribution for three estimation options labeled 1–3. The true RMSEs were estimated by simulation of 10^4 observations in R, and results are repeatable to within ±0.001. Option 1 correctly assumes a $t(2)$ distribution. Option 2 uses the `quantile` function in R. Option 3 is density estimation as discussed in Section 3.6.1. This is the same as Figure 10, but for a $t(2)$ distribution rather than a normal distribution.

3.8. Extensions. Extensions needed beyond those previously discussed include (1) evaluate our ability to estimate alarm limits for non-iid data such as Page's statistic (which could be applied to iid data, but would still not be iid due to the sequential nature of Page's statistic), (2) extend (1) to the multivariate setting, and (3) consider nonstationary data or "concept drift."

Regarding extension (1), Page's test [3, 27] is defined as $P_t = \max(0, P_{t-1}+x_t-k)$. In monitoring PM and/or NMA data streams, Page's test has been found to be simple and effective, and Page's test alarms if $P_t > h$ at any time t during the evaluation for some threshold h. Reference [34] and others in quality control outside of safeguards advocate the use of two Page's tests (one test for abrupt, one test for protracted). For good abrupt loss or diversion detection, use large k and very small h. For good protracted loss or diversion detection, use smaller k and larger h [34].

Regarding extension (2), note that we have considered only scalar data x, but multivariate versions of Page's test have been applied in safeguards [3]. Estimating multivariate

quantiles is more challenging, but we anticipate that multivariate density estimation is a feasible candidate for up to 5 or 10 dimensions.

Regarding extension (3), a real concern with some PM residual streams is that their behavior could change over time. In the sparse literature on such nonstationary behavior, "concept drift," includes the time lag from the past to the current observation and works with blocks of near-stationary residuals.

4. Case Studies from Nuclear Safeguards

In traditional safeguards, periodic nuclear materials accounting (NMA) measurements confirm the presence of special nuclear material (SNM) in accountability units to within relatively small measurement error. Process monitoring (PM) is used to confirm the absence of undeclared flows that could divert SNM for illicit use. Despite occasional attempts to quantify the diversion detection capability of PM, nearly all quantified statements regarding safeguards effectiveness involve NMA, with PM used as a added qualitative measure or to support very frequent NMA, which is called near real time accounting (NRTA).

To assess the extent to which PM can provide quantitative assessment in effectiveness evaluation is one of ten technical challenges in the anticipated increased use of PM data that were discussed during the "2011 Consultancy Meeting on Proliferation Resistance Aspects of Process Management and Process Monitoring/Operating Data" held at the International Atomic Energy Agency. This paper describes traditional roles for PM in support of NRTA and also describes possible front-line roles for PM. If PM data is to be used more quantitatively than it currently is, then historical training data is required in order to estimate PM data behavior under normal operating conditions. Normal operating conditions typically exhibit process variation, so PM data analysis can require relatively long periods of diversion-free training data.

The goal of this case study is to support the goal of using PM data in a more quantitative manner than it currently is. One obstacle to quantitative use of PM data is the need to estimate alarm limits using training data that is free from facility misuse.

In the context of nuclear safeguards, [3] describes how both traditional nuclear material accounting (NMA) data and process monitoring (PM) data analyses lead to time series of residuals that can be monitored, as in statistical process control settings. Unlike standard statistical process control, NMA and PM residuals are usually on different time scales, are serially and cross-datastream correlated, and exhibit departure from standard statistical distributions such as the normal distribution [3]. By "cross-datastream," we mean, for example, that a time series of NMA residuals could be cross correlated with a time series of PM residuals. An example is a waste stream measurement that is used as part of the material balance in NMA and is also used in PM [1–3].

In the context of quantitatively combining PM and NMA subsystems for an improved overall system, some PM data streams [1–3] and/or NMA data streams could be recorded at very high frequency, requiring a very low false alarm rate (extreme quantile) such as 0.0001. For comparison to nonparametric quantile estimation in cases having a few tens or hundreds of observations, we also consider more moderate false alarm rates such as 0.05 or 0.025.

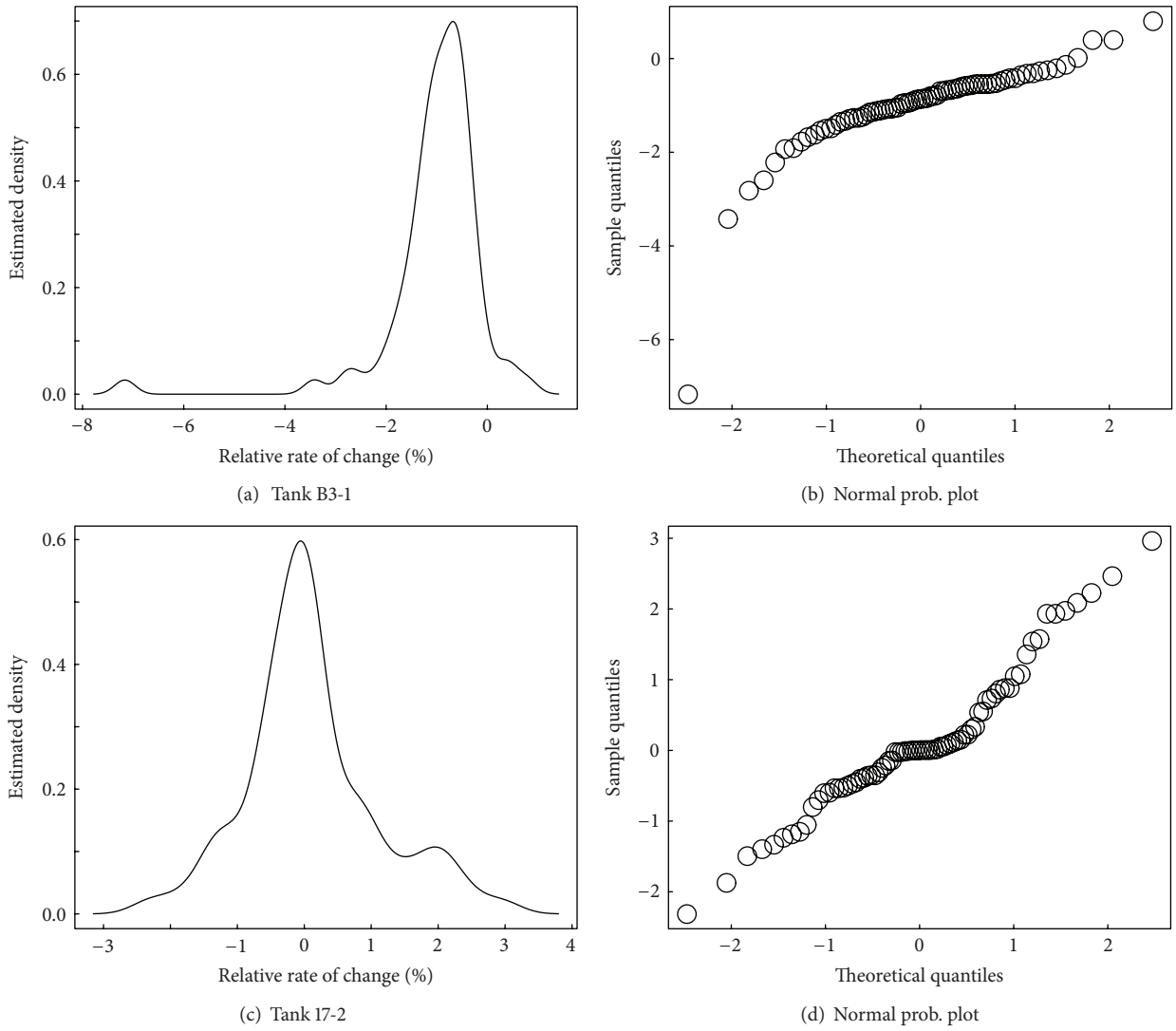
In most applications of PM, some type of training period during which we assume there are no diversions is required in order to learn normal behavior. The goal is to assess training data needs for various PM data types. This section considers two examples. These two case studies examine the amount of training data required for accurate estimation of alarm limits for a range of assumptions regarding the data generation mechanism.

4.1. Example 1: Mixture Fitting for Solution Monitoring Data. Initial studies on solution monitoring data indicate various nonnormal behavior in residuals that arise from monitoring tanks during nontransfer ("wait") modes and also during transfer modes [33, 35, 36]. And, one study considers the impact of nonnormal behavior on loss detection probabilities [37].

Figure 12 plots the estimated probability density (a smoothed histogram) for residuals during 73 wait modes for U storage tank named B3-1 and during 74 tank wait modes for storage tank named 17-2 at Savannah River National Laboratory. The residual is the tank level at the end of the wait mode minus the tank level at the beginning of the wait mode. There is qualitative evidence for mixture behavior and Figure 13 provides quantitative assessment using the BIC as in Section 3 [34, 35, 38, 39]. The normal probability plots in Figure 12 provide additional qualitative evidence for nonnormal behavior. For most mixtures [33] found that approximately 100 training observations are required for adequate estimation of mixture components.

To illustrate the impact of modeling assumptions on estimated tail probabilities, we scale the 73 residuals from tank B3-1 by dividing by the observed standard deviation and estimate the probability the scaled residual exceeds the sample mean by 2. For a standard normal random variable, the estimate is 0.023. If we fit a mixture with 3 components as suggested by the BIC in Figure 13 (differences in BIC of 10 or more are strong evidence for favoring one model over another), the mixture-based estimate is 0.0065, which is considerably smaller than 0.023. Similarly, we scale the 74 residuals from tank 17-3 by dividing by the observed standard deviation deviation and estimate the probability the scaled residual exceeds the sample mean by 2. If we fit a mixture with 2 components (the BIC in Figure 13 suggests that 1 component is adequate so this calculation is purely for illustration), the mixture-based estimate is 0.04, which is considerably larger than 0.023.

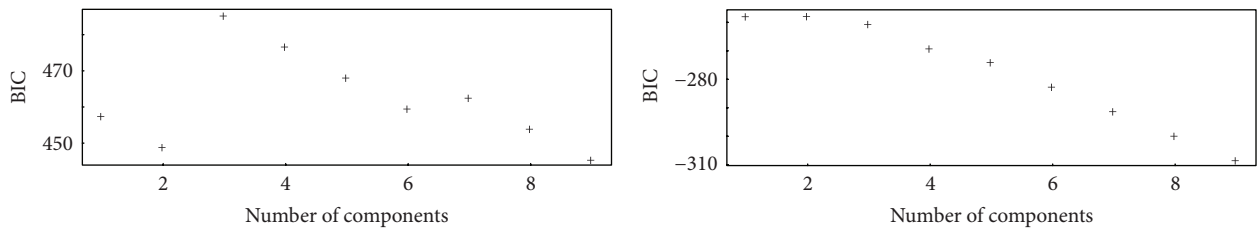
Examples with mixtures in Section 3 and in [33] suggest that approximately 100 or more observations are required in order to have a reasonably high probability of inferring the correct number of components. In solution monitoring, each component has a physical explanation, such as a period of



(a) Tank B3-1 (b) Normal prob. plot

(c) Tank 17-2 (d) Normal prob. plot

Figure 12: Qualitative evidence of mixtures during (a) 73 (from tank B3-1) and (c) 74 (from tank 17-2) tank wait modes.



(a) Tank B3-1 (b) Tank 17-2

Figure 13: The BIC values versus candidate number of components for the 73 wait mode residuals in tank B3-1 and the 74 wait-mode residuals in tank 17-2.

evaporation leading to slight loss during the "wait" mode or condensation leading to slight gain during the "wait" mode.

4.2. Example 2. As an example of batch-to-batch cross talk, in a Pu oxide powder-handling facility, it is common to weigh each can of oxide as it enters [40] and exits a glove box operation. Waste generated during the glove box operation is periodically recovered using a partial or full cleanout [40], and the material not recovered there is distinguished as either "hidden" inventory and "holdup." Hidden inventory remains

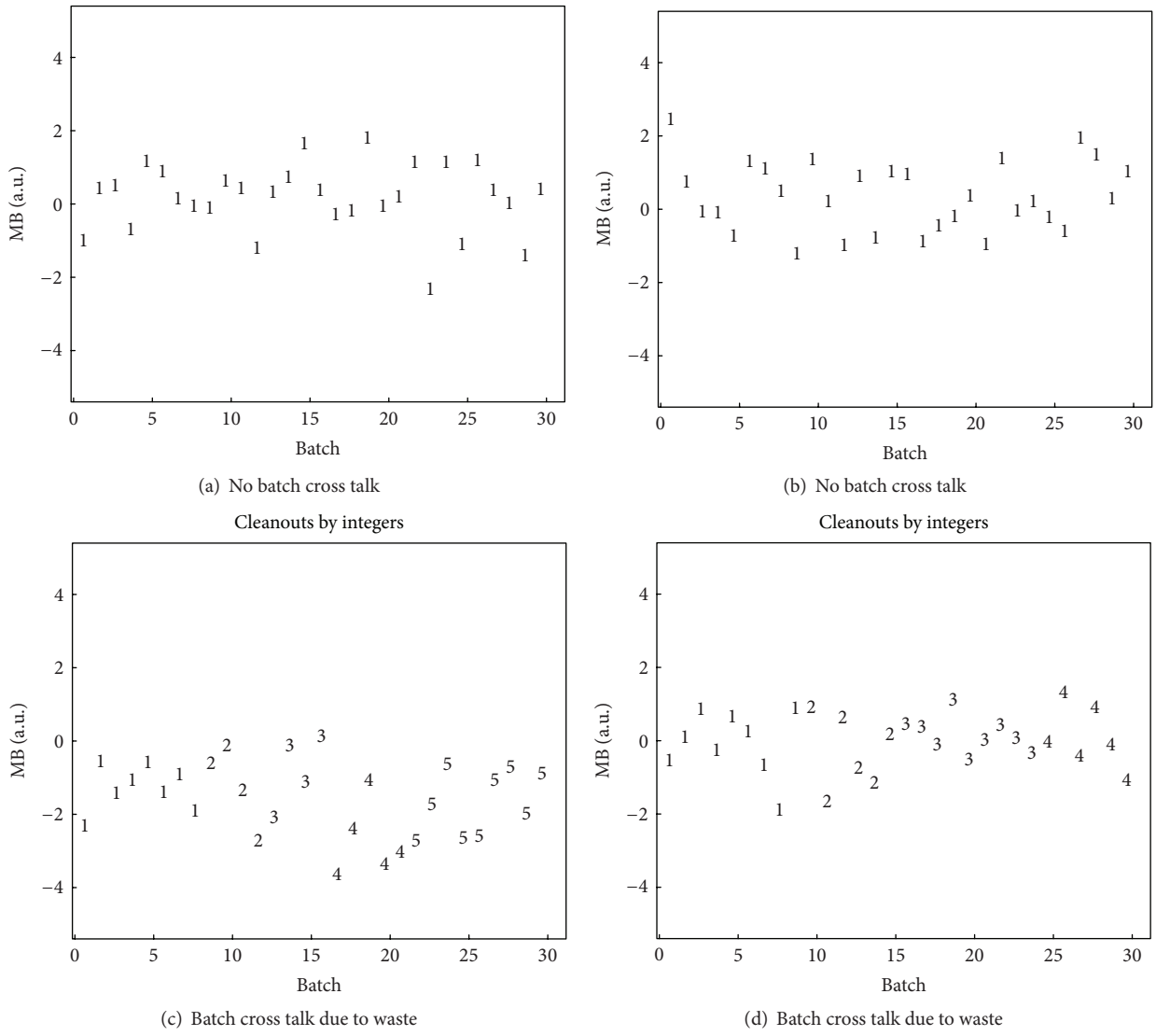


Figure 14: Simulated examples of MB sequences in arbitrary units (a.u.). In (a) and (b) there is no batch-to-batch cross talk and in (c) and (d) there is batch-to-batch cross talk. Plots (c) and (d) have batch-to-batch cross talk due to periodic cleanout of glove box waste.

even after thorough cleanout and is not accessible even to indirect measurement while holdup can partly remain after cleanout but is accessible to indirect measurement.

In this example, the periodically recovered Pu powder is allocated to an estimate of holdup for each batch occurring during the period between glove box cleanouts. For example, suppose 100 mg of Pu powder is recovered after 3 batches of processing Pu oxide cans in a glove box. Then 100/3 = 33.3 mg of Pu powder is reassigned to each of batch 1, batch 2, and batch 3. Figure 14 is an example, with two realizations in a situation with zero holdup, and two realizations in a situation as just described, but with some variation in how many batches of cans are processed before the glove box is cleaned out, with batch-to-batch cross talk arising from periodic cleanout of holdup.

Section 3 described quantile estimation for desired small tail probabilities for several cases, including assuming the data x is normally distributed and assuming the distribution of x is a mixture of distributions. Process variation arising from varying amounts of waste generated per batch will generally lead to batch MBs having an unknown distribution. Therefore, either mixture fitting (because mixtures of normal distributions are known to provide an effective approximation to many distributions) or BIC-based likelihood selection can be considered as a way to estimate desired quantiles.

The residuals in nuclear material accounting (NMA) are the material balances defined as $\mathrm{MB} = T_{\mathrm{in}} + I_{\mathrm{begin}} - T_{\mathrm{out}} - I_{\mathrm{end}}$, where T is a transfer and I is an inventory. In the absence of process variation such as irregular amounts of SNM deposited to holdup per period, and periodic cleanout of the holdup, then the MB will have approximately a normal distribution (because of the central limit effect that arises from combining many measurements in the MB calculation). However, facilities have sometimes observed nonnormal

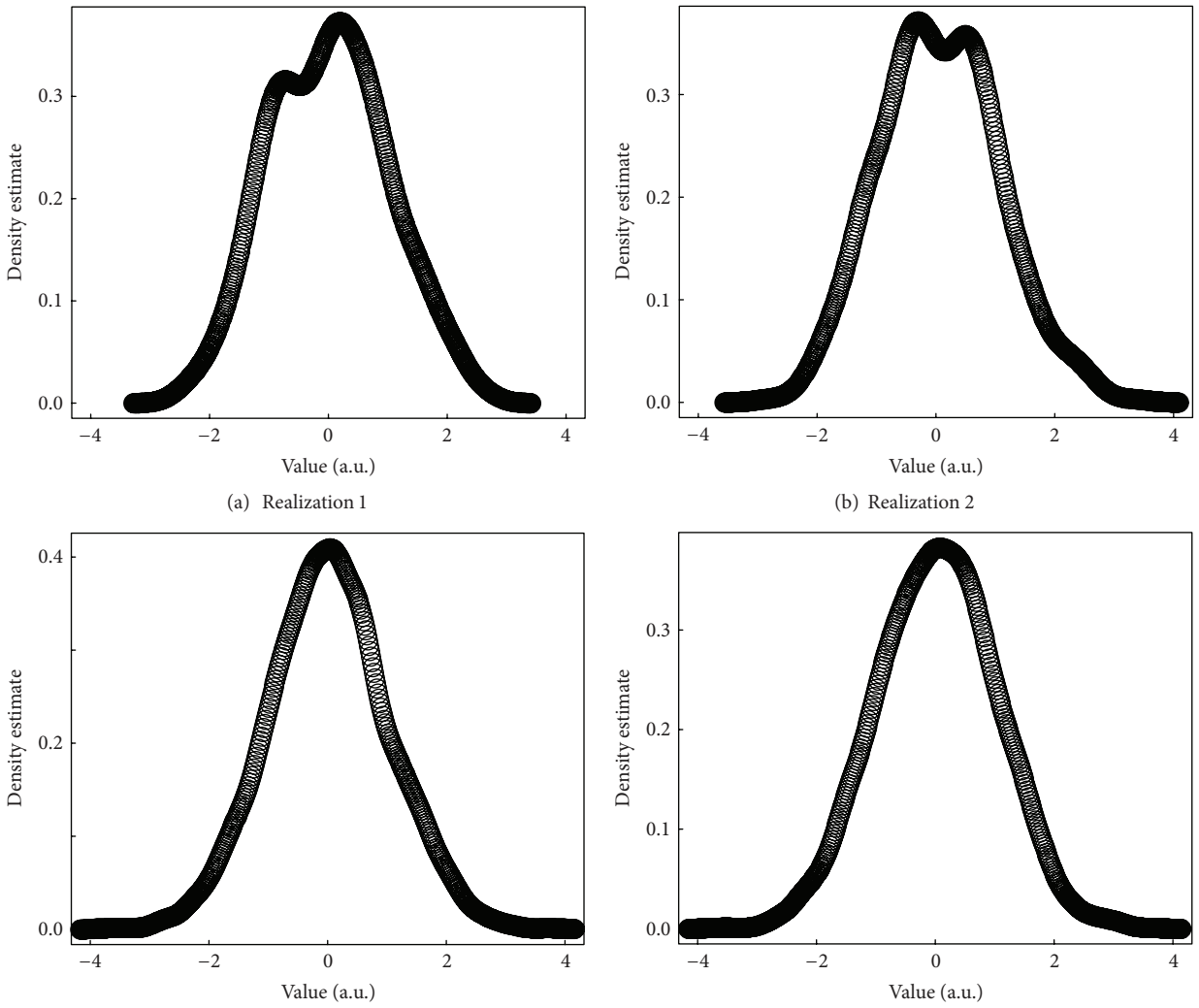


FIGURE 15: Realizations of 600 values in arbitrary units (a.u.). In (a) there are 600 MBs with cleanout between every batch of 3 MBs, with the amount to holdup each period having a normal distribution. Subplot (b) is the same as (a), but with the amount to holdup each period having a uniform distribution. Subplots (c) and (d) are for 600 realizations from a normal distribution for comparison.

MBs, particularly if holdup can fluctuate wildly, as assumed in Example 2.

Figure 15 plots nonparametric density estimates from 600 simulated batches, with cleanout between every set of 3 batches, so there are 200 cleanouts. In Figure 15, the assumed throughput is 100 units, inventory is 100 units, and amount deposited to the glove box is 5 units per batch. The assumed measurement error standard deviations are 0.5% relative random and systematic, with 10% process variation in the average amount deposited to holdup of 5 units. The recovered powder is measured with 1% relative random and systematic error standard deviation. Figure 15(a) assumes the amount deposited to holdup each batch is $H \sim N(5,.5)$. Figure 15(b) assumes the amount deposited to holdup each batch is $H \sim$ Uniform with a mean of 5 and width corresponding to a standard deviation of 0.5. For comparison, Figures 15(c) and 15(d) are each for 600 simulated normal random variables. There is evidence for mixture behavior in Figures 15(a) and 15(b) and also evidence for thinner-than-normal tails in Figures 15(a) and 15(b).

For a more quantitative assessment of whether the resulting distribution of the 600 MBs is approximately normally distributed, one can estimate the 0.025 and 0.975 quantiles for a normal distribution by using $\widehat{\mu} \pm 1.96\widehat{\sigma}$, where $\widehat{\mu}$ is the sample mean and $\widehat{\sigma}$ is the sample standard deviation. Alternatively, one can simply use the observed quantiles of the 600 observations to estimate the 0.025 and 0.975 quantiles (or use any of the options described in Section 3). For Figure 15(b), the (0.025, 0.975) quantiles are estimated as (−3.72, 6.04) using the observed quanttiles or (−3.44, 5.82) using $\widehat{\mu} \pm 1.96\widehat{\sigma}$. The differences $-3.44 - (-3.72) = 0.28$ and $(6.04 - 5.82) = 0.22$ are both much too large to occur by chance (which we confirmed by simulation in R), so there is strong evidence that it is not adequate to assume a normal

distribution. For Figure 15(b), the (0.025, 0.975) quantiles are estimated as (−2.43, 3.06) using the observed quantiles and (−2.46, 3.16) using $\widehat{\mu} \pm 1.96\widehat{\sigma}$. In this case, the differences $-2.46 - (-2.43) = -0.03$ and $3.16 - 3.06 = 0.10$ are not too large to have occurred by chance; however, the differences are in the direction of evidence for a thinner-than-normal distribution.

This batch-to-batch cross talk illustrates the possibility of nonnormal MBs when MBs are computed for each batch. Batch MBs are currently regarded as PM residuals rather than NMA residuals. In either case, sequences of batch MBs are very likely to require cautious analyses, with attention to alarm threshold estimation as in Section 3.

To end this Examples section, we mention that although pyro-reprocessing options are only in the development stages, similar batch-to-batch cross talk is expected, for example, to arise from partial cleanouts of the electrorefiner (ER) [40–47]. PM residual streams associated with the ER are therefore likely to exhibit batch-to-batch "cross-talk" that complicates safeguards, largely due to Uranium and U/TRU (transUranium) behavior in the ER and other process equipment. That is, apparent losses in one batch can appear as a gain in another batch as in our example above.

5. Summary

We presented options to estimate an alarm threshold corresponding to a small false alarm probability p for a range of assumptions regarding the data-generating mechanism. Because analytical evaluation is very difficult, depending on the case, we recommend simulation studies such as presented here to estimate the root-mean-squared estimation error in $\widehat{p}$ to estimate a false alarm probability for candidate threshold estimation options.

In some cases, parametric distributions such as the normal or lognormal or a mixture of normals can provide a reasonable approximation upon which to base alarm threshold estimation. Not surprisingly, the more one correctly assumes about the underlying data-generation mechanism, the smaller the required sample size for accurate estimate of p. As a rough rule of thumb, approximately 100 observations are required for reasonably effective estimation of p. The rule of thumb is motivated by finding in our examples that either (a) there is a very slow decrease in the RMSE as n increases beyond 100, so increasing PM training data requirements beyond approximately $n = 100$ observations is probably not necessary, or (b) the RMSE is very small for some value of n near 100 or less. Of course there are exceptions to any such rule. For example, we considered mixture distributions for which none of the components was extremely rare. If one or more mixture components is rare (such as less than 5% of the overall distribution), then larger sample sizes are needed.

Two process monitoring case studies from nuclear safeguards were presented. The case studies support a safeguards systems option that combines PM and NMA residuals on equal footing [31]. The option requires estimating quantiles in PM residuals corresponding to user-specified small tail probabilities per residual stream, such as 0.001 or 0.025, in order to maintain a small (such as 0.05) per-year system-wide false alarm probability.

Acknowledgments

The authors thank the US National Nuclear Security Administration, office NA22 and the Materials Protection, Control and Accounting (MPACT) program under Nuclear Energy programs.

References

[1] T. Burr, A. Bakel, S. Bryan et al., "Roles for process monitoring in nuclear safeguards at aqueous reprocessing plants," *Journal of Nuclear Materials Management*, vol. 40, no. 2, pp. 42–53, 2012.

[2] T. Burr, M. S. Hamada, J. Howell, M. Skurikhin, L. Ticknor, and B. Weaver, "Data requirements for learning alarm rules for process monitoring," in *Proceedings of the 9th ANS/INMM conference on facilities operation and safeguards interface*, Savannah, Ga, USA, 2012.

[3] T. Burr, M. S. Hamada, M. Skurikhin, and B. Weaver, "Pattern recognition options to combine process monitoring and material accounting data in nuclear safeguards," *Statistics Research Letters*, vol. 1, no. 1, 2012.

[4] C. Roes, *Shewhart-Type Charts in Statistical Process Control [Ph.D. thesis]*, University of Amsterdam, 1995.

[5] W. Albers and W. C. M. Kallenberg, "Self-adapting control charts," *Statistica Neerlandica*, vol. 60, no. 3, pp. 292–308, 2006.

[6] W. Albers and W. Kallenberg, "Alternative Shewhart-type charts for grouped observations," *Metron*, vol. 64, pp. 357–375, 2006.

[7] W. Albers and W. C. M. Kallenberg, "New corrections for old control charts," *Quality Engineering*, vol. 17, no. 3, pp. 467–473, 2005.

[8] W. Albers and W. Kallenberg, "Improved data driven control charts," *International Journal Pure and Applied Mathematics*, vol. 37, pp. 423–439, 2005.

[9] W. Albers, W. C. M. Kallenberg, and S. Nurdiati, "Parametric control charts," *Journal of Statistical Planning and Inference*, vol. 124, no. 1, pp. 159–184, 2004.

[10] W. Albers and W. Kallenberg, "Are estimated control charts in control?" Technical Report 1569, Faculty of Mathematical Sciences, 2001.

[11] W. Albers, W. Kallenberg, and S. Nurdiati, "Parametric control charts," Technical Report 1623, Faculty of Mathematical Sciences, University of Twente, 2002.

[12] W. Albers and W. Kallenberg, "Estimation in Shewhart control charts," Memorandum 1559, University of Twente, 2000.

[13] S. Chakraborti, "Run length, average run length and false alarm rate of shewhart X-bar chart: exact derivations by conditioning," *Communications Statistics in Simulation and Computation*, vol. 29, no. 1, pp. 61–81, 2000.

[14] L. K. Chan, K. P. Hapuarachchi, and B. D. Macpherson, "Robustness of XQ and R charts," *IEEE Transactions on Reliability*, vol. 37, no. 1, pp. 117–123, 1988.

[15] G. Chen, "The mean and standard deviation of the run length distribution of X-charts when control limits are estimated," *Statistica Sinica*, vol. 7, no. 3, pp. 789–798, 1997.

[16] B. Colosimo and E. Castillo, *Bayesian Process Monitoring, Control, and Optimization*, Chapman and Hall, Boca Raton, Florida, USA, 2007.

[17] L. de Haan and A. K. Sinha, "Estimating the probability of a rare event," *Annals of Statistics*, vol. 27, no. 2, pp. 732–759, 1999.

[18] A. Dekkers and L. de Haan, "On the estimation of the extreme-value index and large quantile estimation," *Annals of Statistics*, vol. 17, pp. 1795–1832, 1989.

[19] B. Ghosh, M. Reynolds Jr., and Y. Hui, "Shewhart XQ-charts with estimated process variance," *Communications in Statistics*, vol. 10, pp. 1797–1822, 1981.

[20] P. Hall and I. Weissman, "On the estimation of extreme tail probabilities," *Annals of Statistics*, vol. 25, no. 3, pp. 1311–1326, 1997.

[21] R. J. Hyndman and Y. Fan, "Sample quantiles in statistical packages," *The American Statistician*, vol. 50, no. 4, pp. 361–365, 1996.

[22] W. A. Jensen, J. B. Birch, and W. H. Woodall, "High breakdown estimation methods for phase I multivariate control charts," *Quality and Reliability Engineering International*, vol. 23, no. 5, pp. 615–629, 2007.

[23] G. R. Mercado, M. D. Conerly, and M. B. Perry, "Phase I control chart based on a kernel estimator of the quantile function," *Quality and Reliability Engineering International*, vol. 27, no. 8, pp. 1131–1144, 2011.

[24] G. Nedumaran and J. J. Pignatiello Jr., "On estimating X̄control chart limits," *Journal of Quality Technology*, vol. 33, no. 2, pp. 206–212, 2001.

[25] E. A. Pappanastos and B. M. Adams, "Alternative designs of the Hodges-Lehmann control chart," *Journal of Quality Technology*, vol. 28, no. 2, pp. 213–223, 1996.

[26] C. Quesenberry, "The effect of sample size on estimated limits for XQ and X control charts," *Journal of Quality Technology*, vol. 25, pp. 237–247, 1993.

[27] W. H. Woodall and D. C. Montgomery, "Research issues and ideas in statistical process control," *Journal of Quality Technology*, vol. 31, no. 4, pp. 376–386, 1999.

[28] T. Hastie, R. Tibshirini, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001.

[29] W. Zhou and B.-Y. Jing, "Adjusted empirical likelihood method for quantiles," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 4, pp. 689–703, 2003.

[30] R Development Core Team. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2004, http://www.r-project.org/ .

[31] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[32] T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young, "Mixtools: an R package for analyzing finite mixture models," *Journal of Statistical Software*, vol. 32, no. 6, pp. 1–29, 2009.

[33] T. Burr and M. S. Hamada, "Estimating alarm thresholds and the number of components in mixture distributions," *Nuclear Instruments and Methods in Physics Research A*, vol. 685, pp. 55–61, 2012.

[34] B. Jones and J. Wark, "Near real time materials accountancy system for THORP, ESARDA Bulletin," 1991, http://esarda2.jrc.it/db_proceeding/mfile/P_1991_avignon_117.

[35] T. Burr, M. Suzuki, J. Howell, M. S. Hamada, and C. E. Longo, "Signal estimation and change detection in tank data for nuclear safeguards," *Nuclear Instruments and Methods in Physics Research A*, vol. 640, no. 1, pp. 200–212, 2011.

[36] M. Suzuki, M. Hori, S. Nagaoka, and T. Kimura, "Study on loss detection algorithms using tank monitoring data," *Journal of Nuclear Science and Technology*, vol. 46, no. 2, pp. 184–192, 2009.

[37] T. Burr, M. S. Hamada, J. Howell, and M. Suzuki, "Loss detection results on simulated tank data modified by realistic effects," *Journal of Nuclear Science and Technology*, vol. 49, no. 2, pp. 209–221, 2012.

[38] J. L. Wadsworth and J. A. Tawn, "Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling," *Journal of the Royal Statistical Society B*, vol. 74, no. 3, pp. 543–567, 2012.

[39] M. Black and R. Hickey, "Maintaining the performance of a learned classifier under concept drift," *Intelligent Data Analysis*, vol. 3, pp. 453–474, 1999.

[40] C. Xerri, J. Beckers, M. Boella et al., "Control of nuclear material holdup in MOX fuel fabrication plants in Europe," *ESARDA Bulletin*, vol. 31, pp. 69–75, 2002.

[41] R. O. Hoover, S. Phongikaroon, M. F. Simpson, S. X. Li, and T.-S. Yoo, "Development of computational models for the mark-IV electrorefiner-Effect of uranium, plutonium, and zirconium dissolution at the fuel basket-salt interface," *Nuclear Technology*, vol. 171, no. 3, pp. 276–284, 2010.

[42] M. Simpson and S. Herrmann, "Modeling the pyrochemical reduction of spent UO_2 fuel in a pilot-scale reactor," Tech. Rep. INL/CON-06-11597, 2006.

[43] M. A. Williamson and J. L. Willit, "Pyroprocessing flowsheets for recycling used nuclear fuel," *Nuclear Engineering and Technology*, vol. 43, no. 4, pp. 329–334, 2011.

[44] D. Vaden, R. Benedict, K. Goff, R. Bucher, and A. Yacout, "Material accountancy in an electrochemical fuel conditioning facilty, Argonne National Laboratory report-9606116," 1996, http://www.iaea.org/inis/collection/NCLCollectionStore/_Pub-lic/27/063/27063600.pdf .

[45] J. Zhang, "EChem modeling: a kinetic model for electrorefining based on diffusion control," Los Alamos National Laboratory Report, 2011.

[46] D. Vaden, "Fuel conditioning facility electrorefiner process model," *Separation Science and Technology*, vol. 41, no. 10, pp. 1985–2001, 2006.

[47] G. Cowell, J. Howell, and T. Burr, "Some implications of applying NRTA to a MOX facility with significant temporary stores," in *Proceedings of the 51st Annual Meeting of the Institute of Nuclear Materials Management*, 2010.