2000

Nonparametric Bayes Estimation of Contamination Levels using Observations from the Residual Distribution

Paul H. Kvam
University of Richmond, pkvam@richmond.edu

Ram C. Tiwari

Jyoti N. Zalkikar

Nonparametric Bayes Estimation of Contamination Levels using Observations from the Residual Distribution

Ram C. Tiwari

Department of Mathematics

University of North Carolina, Charlotte

Jyoti N. Zalkikar

Department of Statistics

Florida International University

Paul H. Kvam

School of Industrial and Systems Engineering

Georgia Institute of Technology

Abstract

A nonparametric Bayes estimator of the survival function is derived for right censored data where additional observations from the residual distribution are available. The estimation is motivated by data on contamination concentrations for chromium from one of the EPA's toxic waste sites. The residual sample can be produced by hot spot sampling, where only samples above a given threshold value are collected. The Dirichlet process is used to formulate prior information about the chromium contamination, and we compare the Bayes estimator of the mean concentration level to other estimators currently considered by the EPA and other sources. The Bayes estimator generally outperforms the other estimators under various cost functions. The limiting distribution is the nonparametric maximum likelihood estimator, which is identical to the Kaplan-Meier estimator for concentration values observed below the residual sample threshold. Robustness of the Bayes estimate is examined with respect to misspecification of the prior and its sensitivity to the censoring distribution.

KEY WORDS: Censoring; Cost function; Dirichlet process; Robustness; Skewed distribution.

1 Introduction

Due to the high cost of sampling, practitioners have an increasing need for statistical inference of non i.i.d. data, including censored or truncated observations from the underlying population. This is a serious concern for the U.S. Environmental Protection Agency's (EPA's) Superfund program with regard to the clean up of hazardous toxic waste sites across the country. Due to the expense of environmental sampling procedures, usually no more than a small sample is taken from a site, and censored or truncated data can provide valuable supplemental information to the data analysis.

The underlying distribution for the concentration of contaminants in many of these toxic waste soils is positively skewed. With many right-skewed distributions (e.g., exponential [6]), it is probable that the sample mean $\bar{x}$ will underestimate the true population mean; i.e., $P(\bar{X} < \mu) \geq 1/2$. Consequently, the resulting health risks will be underestimated. In such cases, special attempts are made to capture high values of concentration by sampling additionally from the "hot spots", which are defined as high chemical concentrations representing the upper quantiles of the population. EPA's Superfund program data consist of about 10% of the measurements chosen from the hot spots. Such samples might be generated by using devices that detect only the contaminants that exceed a fixed threshold value. The hot spot data set was called a *purposive sample* in the workshop held by the Office of Emergency and Remedial Response in 1990 [16]. We refer to such data as the observations from the residual distribution. Chen and Jernigan [5] and Chen [3, 4] analyzed the EPA data with

special emphasis given to the skewness of the data, but did not model the hot spot data as observations from the residual distribution.

Let F be the original distribution with support on $R^+ = (0, \infty)$, and let t_0 be a positive known time. With the EPA data, the interval (t_0, ∞) represents chromium levels in the hot spots. The residual distribution is given by its survival function

$$S(t \mid t_0) = \frac{S(t + t_0)}{S(t_0)}, \quad t > 0, \tag{1.1}$$

where for any distribution F, $S \equiv 1 - F$. Naturally, the underlying distribution is not necessarily identifiable if the observations are only available from the residual distribution. To draw inference about F, extra sample information, such as a random sample from F is required.

Residual data appear in many settings, including manufacturing and reliability. For example, if a product manufacturer chooses to test the manufactured item for a limited time before making it available to the consumer, the consumer observes only the item's residual lifetime. This is commonly known as "burn-in". Barlow and Proschan [1], Shaked and Shantikumar [19], among others, emphasized the important role played by residual lifetimes in the analysis of system reliability and aging characteristics.

Related situations also exist in medical studies. For example, the survival data on AIDS patients might include some patients who are known to survive time t_0 beyond the initial stage of the disease and only their remaining lifetimes are observed. These observations constitute a residual lifetime sample. Analysis of AIDS data involving residual lifetimes with right censoring is discussed in Gross and Lai [12]. Left truncated data, combined with right censored lifetimes also have been used in the epidemiological studies of diabetes [11].

Statistical literature contains numerous inferences based on combining information from related samples. Vardi [25] obtained the nonparametric maximum likelihood estimator (NPMLE) using a random sample from F combined with additional observations from the distribution $G(t) \propto \int_0^t w(x)dF(x)$, where $w(.)$ is a known nonnegative bias function. A sample from the residual distribution (1.1) can be regarded as a biased sample from G with $w(x) = I(x > t_0)$, where $I(A)$ is the indicator function of an event A. For ease of notation, the residual distribution is rewritten as

$$\tilde{S}(t) = \frac{S(t)}{S(t_0)}, \quad t > t_0, \tag{1.2}$$

which differs from (1.1) by a location shift.

Recently, Kvam, Singh and Tiwari [15] considered the problem of estimating the underlying distribution function F using a "conventional" sample of randomly right censored lifetimes in addition to independent observations from the residual distribution (1.2), which were also right censored. Thus for the specific choice of the bias function $w(x) = I(x > t_0)$, Kvam et al. [15] extended the work of Vardi [25] to the case of censored data. Their NPMLE of $S(.)$ is identical to the product limit (PL) estimator [13] for all $t < t_0$, but differs for values of $t > t_0$.

In this paper, we consider the estimation of $S(.)$ in a Bayes framework using the Dirichlet process prior [7] for F with parameter $\alpha(.)$, a finite nonnull measure on R^+. The estimator is derived under squared error loss. In Section 2, a result about the Dirichlet process (*Theorem 1*) is established that in turn is used to derive the Bayes estimator of S for the case of uncensored data drawn from the original distribution F as well some from its residual distribution (1.1). In Section 3, the Bayes method is applied to EPA data in order to estimate the mean

of the asymmetric contamination distribution. The Bayes estimator is compared to more standard estimators using various criteria suited for an environmental remediation problem. The performance of the estimators are based on data from one of the EPA's toxic waste sites. The Bayes estimator of F for censored data is derived (using *Theorem 1)* in Section 4, and we examine its sensitivity to different levels of random right censoring. The estimator's Bayes robustness [2] is investigated in Section 5. The effect of the Dirichlet prior distribution on the Bayes estimator is illustrated using the EPA data along with Monte Carlo simulations.

2 Bayes Estimation of the Survival Function

Let F be a Dirichlet process on $(R^+, B(R^+))$ with parameter α, denoted by $F \sim D(\alpha)$, where $B(R^+)$ is the Borel σ -field of subsets of R^+. The role of a Dirichlet process as a prior for the unknown underlying distribution F in solving various nonparametric problems in Bayes setup has been elucidated in the fundamental paper of Ferguson [7]. For a comprehensive review see Ferguson, Phadia and Tiwari [8].

Define $\tilde{\alpha}(A) = \alpha(A \cap (t_0, \infty))$, for $A \in B(R^+)$. Throughout we use the notation $X \amalg Y$ to denote that random elements X and Y are independent. The following result is required for the derivation of the Bayes estimator of S. We will also be using F and $1 - S$ or S interchangeably. The proof is given in the Appendix.

Theorem 1. Let $F \sim D(\alpha)$. *Then* $1 - \tilde{S} \sim D(\tilde{\alpha})$. *Furthermore* $\{S(t) : t \leq t_0\}$ *and* $\left\{\tilde{S}(t) : t > t_0\right\}$ *are independent processes.*

Under squared error loss and prior $D(\alpha)$, the prior guess (the Bayes estimator based on

no sample) of $S(t)$ for $t \leq t_0$, and $\tilde{S}(t)$ for $t > t_0$ are given by

$$\begin{aligned} S_0(t) &= E_{D(\alpha)}\{S(t)\} \\ &= \frac{\alpha(t,\infty)}{\alpha(R^+)}, \quad t \leq t_0 \end{aligned}$$

and

$$\begin{aligned} \tilde{S}_0(t) &= E_{D(\tilde{\alpha})}\{\tilde{S}(t)\} \\ &= \frac{\tilde{\alpha}(t,\infty)}{\tilde{\alpha}(R^+)}, \quad t > t_0 \end{aligned}$$

respectively. Invoking *Theorem 1*, the prior guess for $S(t)$ with $t > t_0$,

$$\begin{aligned} S_0(t) &= E_{D(\alpha)}\{S(t_0)\tilde{S}(t)\} \\ &= E_{D(\alpha)}\{S(t_0)\}E_{D(\tilde{\alpha})}\{\tilde{S}(t)\} \\ &= \frac{\alpha(t_0,\infty)}{\alpha(R^+)}\frac{\tilde{\alpha}(t,\infty)}{\tilde{\alpha}(R^+)} \\ &= S_0(t_0)\tilde{S}_0(t). \end{aligned}$$

Suppose that $X_1, ..., X_n; X_{n+1}, ..., X_{n+m}$ represent $(m+n)$ independent observations of which $X_1, ..., X_n$ given S are identically distributed as $(1-S)$, and $X_{n+1}, ..., X_{n+m}$ given (S, t_0) are identically distributed as $(1-\tilde{S})$. Thus the data consists of a sample $(X_1, ..., X_n)$ of size n from the original distribution F and an additional sample $(X_{n+1}, ..., X_{n+m})$ of size m from the residual distribution $(1-\tilde{S})$. Let $n_0 = \Sigma_{i=1}^{n} I(X_i > t_0)$ be the number of observations from F that are greater than t_0.

For $t \leq t_0$, the Bayes estimator of $S(t)$ under squared error loss and the sample $X_1, ..., X_n$; $X_{n+1}, ..., X_{n+m}$ depends only on $X_1, ..., X_n$. It is well known [7] that the posterior distribution of F given a sample $(X_1, ..., X_n)$ from F is also a Dirichlet process with parameter $\alpha(t)$ updated to $\alpha(t) + \Sigma_{i=1}^{n} I(X_i \leq t)$. Hence the Bayes estimator of $S(t)$ for $t \leq t_0$ is given by

$$\begin{aligned} S_b(t) &= E_{D(\alpha)}\{S(t) \mid X_1, ..., X_n\} \\ &= p_n S_0(t) + (1 - p_n) S_n(t), \end{aligned} \tag{2.3}$$

where $p_n = \alpha(R^+)/(\alpha(R^+) + n)$, and $nS_n(t) = \sum_{i=1}^{n} I(X_i > t)$.

For $t > t_0$, it follows from *Theorem 1* that the Bayes estimator of $S(t)$ is given by

$$E_{D(\alpha)}\{S(t_0) \mid X_1, ..., X_n\} E_{D(\tilde{\alpha})}\{\tilde{S}(t) \mid \tilde{X}_1, ..., \tilde{X}_{n_0}, X_{n+1}, ..., X_{n+m}\},$$

where $\tilde{X}_1, ..., \tilde{X}_{n_0}$ are the observations among $X_1, ..., X_n$ from F that are greater than t_0. Note that among these n observations, only $\tilde{X}_1, ..., \tilde{X}_{n_0}$ provide information for $\tilde{S}(t)$. Let $q_{m+n_0} = \tilde{\alpha}(R^+)/(\tilde{\alpha}(R^+) + n_0 + m)$, and define $S_{m+n_0}(t)$ such that

$$(m + n_0) S_{m+n_0}(t) = \sum_{i=1}^{n} I(\tilde{X}_i > t) + \sum_{i=n+1}^{n+m} I(X_i > t) = \sum_{i=1}^{n+m} I(X_i > t > t_0).$$

Then,

$$\hat{S}(t) = \begin{cases} S_b(t) & t \le t_0 \\ S_b(t_0)\left\{q_{m+n_0}\tilde{S}_0(t) + (1 - q_{m+n_0}) S_{m+n_0}(t)\right\} & t > t_0 \end{cases} \tag{2.4}$$

If $t_0 = 0$, then $n_0 = n$ and with no additional observations from the residual distribution, (i.e., $m = 0$) the estimator $\hat{S}(t)$ reduces to $S_b(t)$. The parameter $\alpha(R^+)$ is a measure of belief in the prior guess of F [7]. Clearly, in the limit as $\alpha(R^+) \to 0$, we have $\tilde{\alpha}(R^+) \to 0$ as well, and the Bayes estimator of $S(t)$, given by (2.3) and (2.4), converges weakly [18] to $S_n(t)$ for $t \le t_0$, and to $S_n(t_0) S_{m+n_0}(t)$ for $t > t_0$. The limiting Bayes estimator of $S(t)$ coincides with the maximum likelihood estimator of $S(t)$ obtained by Kvam et al. [15].

3 Estimating Mean Concentration of Contaminants

For remediation decisions regarding toxic waste sites, the EPA uses particular performance measures tailored to environmental and health risks corresponding to underestimating contamination levels as well as financial risks corresponding to overestimating levels. The contamination level mean is of primary interest. By applying the sample mean with right-skewed data, there is a high probability that the true mean contamination level will be underestimated. Consequently, the public's risk to health can be significantly underestimated as well. To alleviate this problem, the EPA collects about 10% of the data from the residual distribution and uses the upper point of a 95% normal theory confidence interval ($UCL = \bar{X} + 1.96\sigma_{\bar{X}}$) as an estimator, where $\bar{X}$ is the sample mean and $\sigma_{\bar{X}}$ is the standard deviation of $\bar{X}$. However, due to skewness of the underlying distribution, the UCL estimator has a significant probability of falling above the true mean by more than two standard deviations. It is mentioned by the EPA [9] that such large over-estimation can result in unnecessary costs such as money spent in cleaning a site. Furthermore, excessive money spent on one site postpones the remediation of the next site.

Several estimators have been proposed for the mean of an asymmetric distribution including the generalized Bayes estimator for lognormal models [17], the transformed mean obtained through use of a Box-Cox transformation [23, 20], the once-Winsorized mean [10], the penalized mean [5] and most recently the modified penalized mean [4]. Since F has support on R^+, the mean μ of F is given by $\mu = \int_0^\infty S(t)dt$, and the Bayes estimator of μ under squared error loss [7] is

$$
\begin{aligned}
\hat{\mu} &= E_{D(\alpha)}\left\{\int_0^\infty S(t)dt|X_1,...,X_n;X_{n+1},...,X_{n+m}\right\} \\
&= E_{D(\alpha)}\left\{\int_{(0,t_0]} S(t)dt|X_1,...,X_n\right\} \\
&\quad +E_{D(\tilde{\alpha})}\left\{S(t_0)\int_{(t_0,\infty)}\tilde{S}(t)|\tilde{X}_1,...,\tilde{X}_{n_0},X_{n+1},...,X_{n+m}\right\} \\
&= \int_{(0,t_0]} S_b(t)dt + \int_{(t_0,\infty)}\hat{S}(t)dt. \qquad (3.5)
\end{aligned}
$$

If $t_0 = 0$ and $m = 0$, the estimator $\hat{\mu}$ reduces to

$$
\begin{aligned}
\hat{\mu} &= \int_0^\infty S_b(t)dt \\
&= p_n\int_0^\infty S_0(t)dt + (1-p_n)\int_0^\infty S_n(t)dt \\
&= p_n\mu_0 + (1-p_n)\bar{X},
\end{aligned}
$$

where μ_0 is the prior mean.

We are constructing an estimator of the mean for which modest amounts of data (e.g., $n \leq 20$) are combined with smaller samples from hot spots. In addition, limited subjective prior information may be available to further characterize the contamination concentration. The data on chromium concentrations from one of EPA's toxic waste sites consists of 623 observations, assumed here to be the actual population in order to evaluate the performance of $\hat{\mu}$. A Monte Carlo analysis is performed by repeatedly drawing samples from these 623 observations (with replacement) and comparing the resulting estimates with the mean of the population. We denote S_{epa} as the empirical estimator based on these 623 values.

Concentration values range from 0.15 mg/kg to 103.975 mg/kg with a mean of 4.945 mg/kg, and the standard deviation of 10.002 mg/kg. The log-normal model is extensively

used in statistical analysis of environmental data [26]. For this reason, we apply the log-normal distribution as the prior with the Dirichlet process. The population of 623 values, however, does not appear to be distributed as log-normal (see Figure 1). This can be shown analytically using a variety of tests for goodness of fit. We compensated the goodness of fit test (based on the Kolmogorov-Smirnov statistic) by acknowledging that the observations listed at 0.15 mg/kg are actually left censored. However, severe lack of fit also exists in the upper tails of the distribution.

We are interested in properties of the Bayes estimators that are constructed using tacitly assumed prior distributions. We are also interested in the robustness of the Bayes estimators in case the prior distribution is misspecified. We model the prior parameter $\alpha(t)$ using a log-normal distribution with prior parameters $\mu = 0.8$ and $\sigma^2 = 1.6275$. In this case, the mean and variance closely resemble the true underlying population, so any misspecification of the Dirichlet prior is due to the choice of the log-normal distribution for S_0. We take $\alpha(R^+) = 1$ so that the prior guess of $F(t)$ is $\alpha(-\infty, t)$.

To examine the relative performance of the nonparametric Bayes estimator, we compute $\hat{\mu}$ and compare it to two common estimators of the mean: the mean $\bar{X}$ of the random sample of size n, and the biased mean $\bar{X}^*$ of the combined sample of size $n+m$. We sample 10% of the data from the residual distribution (hot spots) using the upper 95^{th} percentile ($t_0 = 23.2453$) as the threshold value.

Three performance measures are used to contrast the estimators: the mean squared error (*mse*) along with the modified cost function (*mcf*) and the EPA's asymmetric cost function

(*acf*), which are defined below. The *mcf*, which is also not symmetric, is defined as

$$mcf = 1 - \Pr(\mu_x - 0.5\sigma_x n^{-\frac{1}{2}} \leq \hat{\theta} \leq \mu_x + 2\sigma_x n^{-\frac{1}{2}}),$$

and differs slightly from the cost function [4] given by

$$cf = 1 - \Pr(\mu_x \leq \hat{\theta} \leq \mu_x + 2\sigma_x n^{-\frac{1}{2}}),$$

where $\hat{\theta}$ is an estimator of the population mean μ_x,and σ_x is the population standard deviation. Under *cf*, the biased mean is a useful estimator due to its conservative nature [4]. However, *mcf* is preferred over *cf* because it does not over-penalize for a small negative error. The asymmetric cost function applied by Flatman and Englund [9] is defined as

$$acf = \begin{cases} c_1|\mu - \hat{\theta}| & \hat{\theta} \leq \mu \\ c_2|\mu - \hat{\theta}| & \hat{\theta} > \mu. \end{cases} \tag{3.6}$$

As they noted, both underestimation and overestimation of the population mean can lead to critical loss. Overestimation of the population mean, which leads to unneeded remediation, can also lead to fewer future remediation in other critical areas of contamination. Differences in the loss are characterized through the positive constants (c_1, c_2). The *mcf* and *acf* criteria are a more suitable performance measures than the *cf* because they better balance these two potential losses.

Tables 1 and 2 show the summary of the resulting estimates for the case in which (n, m) is set at levels (9,1), (18,2), and (45,5). For Table 1, $\alpha(R^+) = 1$ is used, and for Table 2, $\alpha(R^+) = 20$. For the *acf* criteria in (3.6), we assign (c_1, c_2)=$(1, 2)$, assuming more serious consequences exist for underestimating the mean contamination level. Clearly, $\hat{\mu}$ has the

best overall performance among the three procedures. The improvement is more dramatic with larger samples.

In the simulation results summarized in Table 1, little prior information supports the Bayes estimator, so the results are very close to the MLE results in [15]. For the sample of size $n + m = 20$, for example, the *mse* of the MLE is 4.8187, not far from *mse* for the Bayes estimate of the mean. The choice of prior distribution S_0 can strongly effect the Bayes estimate if the prior weight is larger; this can be seen in the summary results of Table 2. We will further investigate this issue in Section 5.

4 Bayes Estimation of S(t) with Censored Data

In various industrial as well as medical experiments, it is not always possible to obtain purely uncensored lifetimes. Technological advances in these fields combined with limited experimental budgets have made random right censoring increasingly common. In this section we consider the problem of estimating the underlying distribution function in the Bayes framework using a conventional sample of randomly right censored lifetimes in addition to independent items generated from the residual distribution in (1.2), which might also be right censored.

Let $X_1, ..., X_n$; $X_{n+1}, ..., X_{n+m}$ be random variables defined as before. Let $Y_1, ..., Y_n$; $Y_{n+1}, ..., Y_{n+m}$ be independent random variables such that $Y_1, ..., Y_n$ are independent identically distributed (*i.i.d.*) H_1 on R^+ and $Y_{n+1}, ..., Y_{n+m}$ be *i.i.d.* H_2 on (t_0, ∞). We assume that $\{X_i\} \amalg \{Y_i\}$. The observed right censored data consists of $(Z_i, \delta_i), i = 1, ..., n+m$, where $Z_i = \min(X_i, Y_i)$ and $\delta_i = I(X_i \leq Y_i)$, $i = 1, ..., n+m$. The Bayes estimator of $S(t)$ for $t \leq t_0$

is given by

$$\hat{S}_{c,n}(t) = E_{D(\alpha)}\{S(t) \mid (Z_i, \delta_i), i = 1, \ldots, n\}.$$

Proceeding along the lines of Susarla and Van Ryzin [21], it can be shown that for $t \leq t_0$,

$$\hat{S}_{c,n}(t) = \frac{\alpha(t,\infty) + nS_{c,n}(t)}{\alpha(R^+) + n} \prod_{i=1}^{n} \left(\frac{\alpha(Z_i,\infty) + nS_{c,n}(Z_i) + 1}{\alpha(Z_i,\infty) + nS_{c,n}(Z_i)} \right)^{I(Z_i \leq t, \delta_i = 0)},$$

where $nS_{c,n}(t) = \sum_{i=1}^{n} I(Z_i > t)$. For $t > t_0$, using *Theorem* 1, the Bayes estimator of $S(t)$ is given by

$$\begin{aligned} \hat{S}_{c,n}(t) &= E_{D(\alpha)}\{S(t) \mid (Z_i, \delta_i), i = 1, \ldots, n+m\} \\ &= E_{D(\alpha)}\{S(t_0) \mid (Z_i, \delta_i), i = 1, \ldots, n\} \times E_{D(\tilde{\alpha})}\{\tilde{S}(t) \mid (Z_i, \delta_i), i = 1, \ldots, n+m\}. \end{aligned}$$

Let $\tilde{Z}_1, \ldots, \tilde{Z}_{n_0}$ denote the observations from among $\{Z_1, \ldots, Z_n\}$ that are greater than t_0 and let $\tilde{\delta}_i$, $i = 1, \ldots, n_0$, be the concomitant values of δ_i associated with $\tilde{Z}_i, i = 1, ..., n_0$. Also, relabel (Z_i, δ_i) , $i = n+1, \ldots, n+m$ as $\left(\tilde{Z}_i, \tilde{\delta}_i\right)$, $i = n_0 + 1, \ldots, m + n_0$. Define

$$(m + n_0)\, \tilde{S}_{c,m+n_0}(t) = \sum_{i=1}^{m+n_0} I\left(\tilde{Z}_i > t\right).$$

Then,

$$\begin{aligned} \hat{S}_{c,n}(t) &= \hat{S}_{c,n}(t_0)\, E_{D(\tilde{\alpha})}\{\tilde{S}(t) \mid \left(\tilde{Z}_i, \tilde{\delta}_i\right), i = 1, \ldots, m + n_0\} \\ &= \hat{S}_{c,n}(t_0)\, \tilde{S}_{c,n}(t), \end{aligned}$$

where

$$\tilde{S}_{c,n}(t) = \frac{\tilde{\alpha}(t,\infty) + (m+n_0)\,\tilde{S}_{c,m+n_0}(t)}{\tilde{\alpha}(R^+) + (m+n_0)} \times \prod_{i=1}^{m+n_0} \left(\frac{\tilde{\alpha}\left(\tilde{Z}_i,\infty\right) + (m+n_0)\,\tilde{S}_{c,m+n_0}\left(\tilde{Z}_i\right) + 1}{\tilde{\alpha}\left(\tilde{Z}_i,\infty\right) + (m+n_0)\,\tilde{S}_{c,m+n_0}(\tilde{Z}_i)} \right)^{I\left(\tilde{Z}_i \le t, \tilde{\delta}_i = 0\right)}.$$

In the limit as $\alpha(R^+) \to 0$, the limiting Bayes estimator of S is given by

$$\begin{aligned} \hat{S}_{c,n}(t) &= \prod_{i=1}^{n} \left(\frac{nS_{c,n}(Z_i)}{nS_{c,n}(Z_i)+1} \right)^{\delta_i}, t \le t_0, \\ &= S_{c,n}(t_0) \prod_{i=1}^{m+n_0} \left(\frac{(m+n_0)\,\tilde{S}_{c,m+n_0}\left(\tilde{Z}_i\right)}{(m+n_0)\,\tilde{S}_{c,m+n_0}\left(\tilde{Z}_i\right)+1} \right)^{\tilde{\delta}_i}, t > t_0. \end{aligned}$$

This is identical to the NPMLE of S derived by Kvam et al.[15]. Note that the limiting Bayes estimator of S, for $t \le t_0$, is the usual PL estimator of S, but for $t > t_0$ it is the rescaled PL estimator, where the scaling factor is less than 1 (and equal to 1 when $t_0 = 0$).

The effect of censoring on the nonparametric Bayes estimator is difficult to assess. We investigate the effect of random right censoring by drawing repeated contamination measurements from the EPA sample and coupling each observation with a randomly generated censoring time. We modeled censoring using the exponential distribution, and left the hot spot samples uncensored (which seems to be a realistic environmental sampling scenario), thus $H_1(t;\lambda) = 1 - e^{-t/\lambda}$ and $H_2(t) = 0$ for $t > 0$. Figure 2 displays the risk of the Bayes estimate $\hat{S}$ with respect to S_{epa} using squared error loss. The risk function, defined

$$R(\hat{S}, S_{epa}; \lambda) = \int_{-\infty}^{\infty} (\hat{S} - S_{epa})^2 dS_{epa}, \tag{4.7}$$

is a function of the mean λ for the censoring distribution. Naturally, the error increases as the mean of the censoring distribution decreases (and more observations become censored).

In this interval of $0 \leq \lambda \leq 10$, the decrease in risk is approximately proportional to the increase in the proportion of the population that becomes censored:

λ	1	2	3	4	5	6	7	8	9	10
$P(censoring)$	.388	.505	.570	.617	.652	.680	.703	.722	.740	.754

Censoring probabilities are computed using simulations. Again, we selected $n = 18$ regular observations with possible right censoring, along with m=2 uncensored hot spot samples, and assigned t_0 to be the upper 95^{th} percentile of the EPA data.

5 Discussion

In this paper, we derived a nonparametric Bayes estimator of the survival function when a conventional random sample was supplemented with observations from the residual distribution. The estimator was motivated by the EPA's problem of estimating contamination levels when *i.i.d.* samples are combined with hot spot samples from upper percentiles of the contamination distribution. In Section 3, the Bayes estimator is compared to standard estimators of the mean contamination level using various loss functions. Specifically, we examine bias, mean-squared error, and two other loss functions, including the EPA's asymmetric cost function [9]. The gains in using the Bayes estimate are clearly apparent in this case, especially in terms of *mse* and *acf*. Both *mcf* and *acf* allow for a small amount of under-estimation and help to demonstrate that subjective penalties for under-estimation and over-estimation can be easily parameterized.

As noted in Section 3, the choice of prior distribution can strongly affect the Bayes estimate if the prior weight is significant. In the Bayes framework, the prior $\alpha(.)$ is assumed to be known. If errors due to prior misspecification (as discussed by Berger [2]) are inconsequential, the estimator shows *Bayes robustness*. We demonstrate through simulation that the Bayes estimator using additional information from the residual distribution exhibits substantial Bayes robustness with respect to misspecification of the prior (S_0). Using the same sampling scheme from Section 3, where (n,m)=(18,2) and the hot spot consisted of the upper 5^{th} percentile of the population, three families of prior distributions were considered: the Lognormal, Normal and Weibull. Each distribution was assigned the same mean and variance as the underlying population, so the measure of robustness was based on other properties of the prior. Three different prior weights were used to contrast the prior distributions: $\alpha(R^+)$=(1,10,20).

Robustness is measured with $R(\hat{S}, S_{epa})$, as defined in (4.7), now assuming no censoring occurs. The results, listed in Table 3, indicate that the choice of prior (among these three considered) has little effect on the risk. Actually, none of the three distributions models the population of EPA data particularly well. Although the lognormal is the intuitive choice for a prior, its performance is below that of the other priors in each case. Despite the obvious reasons the normal distribution should not characterize the underlying contamination distribution (e.g., it is symmetric, and $P(X < 0) > 0.30$), it produced a Bayes estimate with slightly smaller risk than the other two distributions.

We further examine robustness as a function of the misspecified prior. Results are based on simulated data with $S(x) = e^{-x}$ representing the true prior distribution. In Figures 3 and

4, the effect of prior mean misspecification is displayed for the case in which $S_0(x) = e^{-\theta x}$, where $\theta \epsilon (0, 3)$ for the *misspecified* Dirichlet prior model. For this treatment, samples of size 10 and 20 are drawn, with 10% of the data selected from the hot spot. The hot spot is characterized by a threshold selected to be $t_0 = 3$, approximately the 0.95 quantile for $S(x)$, the exponential distribution with mean equal to one. With samples of size 10, the risk increases no more than 14% at the point where the prior mean is misspecified to be three times smaller than the actual prior mean. If θ is between 50% and 200% of the true mean, the increase in risk is less than 5%. With samples of size 20, the effect of mean misspecification is further dampened; the risk is less than 8% for all values of $\theta \epsilon (0, 3)$.

To deal with situations wherein the expert's opinion about $\alpha(t)$ is either partially or completely unknown, the empirical Bayes framework [14, 22, 24] is under investigation for this particular estimation problem where additional residual observations are present. The robustness of $\hat{\mu}$ with respect to the choice of the functional form of $\alpha(t)$ as well as with respect to its parameters is also under further study.

6 Appendix

Proof of Theorem 1. It suffices to show that for $k \geq 1$ and some measurable partition $t_0 < t_1 < \cdots < t_{k+1} = \infty$ of (t_0, ∞), the distribution of

$$\left(1 - \tilde{S}(t_1), \tilde{S}(t_1) - \tilde{S}(t_2), \ldots, \tilde{S}(t_k) - \tilde{S}(t_{k+1})\right)$$

is a singular Dirichlet with parameters $(\tilde{\alpha}(t_0, t_1], \tilde{\alpha}(t_1, t_2], \ldots, \tilde{\alpha}(t_k, t_{k+1}))$. For any $t \leq t_0$, let $Z(0, t]$, $Z(t, t_0]$, $Z(t_0, t_1], \ldots, Z(t_k, t_{k+1})$ be independent Gamma random variables (defined on a common probability space) with common scale parameter $\beta > 0$ and shape parameters,

respectively, $\alpha(0,t]$, $\alpha(t,t_0]$,$\alpha(t_0,t_1]$, ..., $\alpha\left(t_k,t_{k+1}\right)$. Note that $Z\left(R^+\right)=Z(0,t]+Z(t,t_0]+Z(t_0,t_1]+\cdots+Z\left(t_k,t_{k+1}\right)$ is a Gamma random variable with scale parameter β and shape parameter $\alpha\left(R^+\right)$, and that $Z(0,t]/Z\left(R^+\right)=F\left(t\right)$, and $Z(t_i,t_{i+1}]/Z\left(R^+\right)=F\left(t_{i+1}\right)-F\left(t_i\right)$ for $i=0,1,\ldots,k+1$. Furthermore, $1-\tilde{S}\left(t_1\right)=Z(t_0,t_1]/Z(t_0,\infty)$, $\tilde{S}\left(t_1\right)-\tilde{S}\left(t_2\right)=Z(t_1,t_2]/Z\left(t_0,\infty\right)$,..., $\tilde{S}\left(t_k\right)-\tilde{S}\left(t_{k+1}\right)=Z(t_i,t_{i+1}]/Z\left(t_0,\infty\right)$. Using a standard property of Gamma distributions,

$$\begin{aligned}\left(1-\tilde{S}\left(t_1\right),\tilde{S}\left(t_1\right)-\tilde{S}\left(t_2\right),\ldots,\tilde{S}\left(t_k\right)-\tilde{S}\left(t_{k+1}\right)\right) &\stackrel{d}{=} \left(\frac{Z(t_0,t_1]}{Z\left(t_0,\infty\right)},\frac{Z(t_1,t_2]}{Z\left(t_0,\infty\right)},\ldots,\frac{Z\left(t_k,t_{k+1}\right)}{Z\left(t_0,\infty\right)}\right)\\ &\sim D\left(\tilde{\alpha}(t_0,t_1],\tilde{\alpha}(t_1,t_2],\ldots,\tilde{\alpha}\left(t_k,t_{k+1}\right)\right).\end{aligned}$$

That is, $1-\tilde{S}\sim D(\tilde{\alpha})$. Furthermore,

$$\left(\frac{Z(t_0,t_1]}{Z\left(t_0,\infty\right)},\frac{Z(t_1,t_2]}{Z\left(t_0,\infty\right)},\ldots,\frac{Z\left(t_k,t_{k+1}\right)}{Z\left(t_0,\infty\right)}\right)\amalg\left(Z(0,t],Z(t,t_0],Z\left(t_0,\infty\right)\right)$$

which implies that

$$\left(\frac{Z(t_0,t_1]}{Z\left(t_0,\infty\right)},\frac{Z(t_1,t_2]}{Z\left(t_0,\infty\right)},\ldots,\frac{Z\left(t_k,t_{k+1}\right)}{Z\left(t_0,\infty\right)}\right)\amalg\frac{Z(0,t]}{Z(0,t]+Z(t,t_0]+Z\left(t_0,\infty\right)}$$

or

$$\left(1-\tilde{S}\left(t_1\right),\tilde{S}\left(t_1\right)-\tilde{S}\left(t_2\right),\ldots,\tilde{S}\left(t_k\right)-\tilde{S}\left(t_{k+1}\right)\right)\amalg F(t).$$

Thus $\{F(t):t\leq t_0\}\amalg\{\tilde{S}(t):t>t_0\}$ or equivalently $\{S(t):t\leq t_0\}\amalg\{\tilde{S}(t):t>t_0\}$.□

References

[1] Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life-Testing.* Holt, Rhinehart and Winston, New York.

[2] Berger, J. O. (1984) "The robust Bayes viewpoint (with discussion)," *Robustness of Bayes analysis*, J. Kadane, ed., 63 - 144, Amsterdam: North Holland.

[3] Chen, Ling (1995), "A Minimum Cost Estimator for the Mean of Positively Skewed Distributions With Applications to Estimation of Exposure to Contaminated Soils," *Environmetrics* **6,** 181-193.

[4] Chen, Ling (1997), "An Improved Penalized Mean for Estimating the Mean Concentration of Contaminants," Manuscript, Florida International University, Miami, Florida.

[5] Chen, Ling and Jernigan, W. Robert (1996), "Conservative, Nonparametric Estimation of Mean Concentration to Contaminants," *Statistica Sinica* **6,** 547-559.

[6] Chen, J. and Rubin, H. (1986), "Bounds for the difference between median and mean of gamma and Poisson distributions," *Probability and Statistics Letters*, **4**, 281- 283.

[7] Ferguson, T. S. (1973), "A Bayes Analysis of Some Nonparametric Problems," *The Annals of Statistics* **1,** 209-230.

[8] Ferguson, T. S., Phadia, E. G., and Tiwari, R. C. (1992), "Bayes Nonparametric Inference," in *Current issues in Statistical inference: Essays in Honor of D. Basu (Eds. M. Ghosh and P. K. Pathak). IMS Lecture Notes & Monograph Series.* **17,** 127-150.

[9] Flatman, G. T., Englund, E. J. (1991), "Asymmetric loss function for Superfund remediation decisions", *ASA Proceedings of Business and Economic Statistics Section*, 204 - 209.

[10] Fuller, W.A. (1991), "Simple Estimators for the Mean of Skewed Populations", *Statistica Sinica*, 1, 137-158.

[11] Green, A. and Hougaard, P. (1984), "Epidemiological studies of diabetes mellitus in Denmark: Mortality and causes of death among insulin-treated diabetic patients", *Diabetologia*, 26, 190-194.

[12] Gross, S. T. and Lai, T. L. (1996), "Nonparametric Estimation and Regression Analysis With Left-truncated and Right-censored Data," *Journal of the American Statistical Association* **91,** 1166-1180.

[13] Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association* **53,** 457-481.

[14] Korwar, R.M. and Hollander, M. (1976), "Empirical Bayes Estimation of a Distribution Function", *The Annals of Statistics* **4,** 581-588.

[15] Kvam, P. H., Singh, H., and Tiwari, R. C. (1999), "Nonparametric Estimation of the Survival Function Based on Censored Data With Additional Observations from the Residual Life Distribution," *Statistica Sinica* **9** 229 - 246.

[16] Minutes of EPA (1990), "Methods to Estimate Long-Term Exposure to Contaminated Soils at Superfund Sites", Feb. 23, pp. 1-22.

[17] Rukhin, A.L. (1986), "Improved Estimation in Lognormal Models", *Journal of the American Statistical Association* **81,** 1046-1049.

[18] Sethuraman, J. and Tiwari, R. C. (1982), "Convergence of Dirichlet Measures and the Interpretation of Their Parameter," in *Statistical Decision Theory and Related Topics III* 2 (*Eds. S.S. Gupta and J.O. Berger*) Academic Press, New York.

[19] Shaked, M. and Shanthikumar, J. G. (1994), *Stochastic Orders and their Applications.* Academic Press, London.

[20] Shumway, R.H., Azari, A.S., and Johnson, P. (1989), "Estimating Mean Concentrations Under Transformation for Environmental Data With Detection Limits", *Technometrics* **31,** 347-356.

[21] Susarla, V. and Van Ryzin, J. (1976), "Nonparametric Bayes Estimation of Survival Curves from Incomplete Observations," *Journal of the American Statistical Association* **71,** 897-902.

[22] Susarla, V. and Van Ryzin, J. (1978), "Empirical Bayes Estimation of a Distribution (Survival) Function from Right Censored Observations", *The Annals of Statistics* **6**, 740-754.

[23] Taylor, J.M.G. (1985), "Measures of Location of Skew Distributions Obtained Through Box-Cox Transformations", *Journal of the American Statistical Association* **80,** 427-432.

[24] Tiwari, R.C. and Zalkikar, J.N. (1985), "Empirical Bayes Estimation of Functionals of Unknown Probability Measures", *Communications in Statistics- Theory and Methods* **14**, 2963-2996.

[25] Vardi, Y. (1985), "Empirical Distributions in Selection Bias Models," *The Annals of Statistics* **13,** 178-203.

[26] White, P. (1992), "Assessment of soil contact exposures at sites with toxic contamination", U.S. Environmental Protection Agency Report, Office of Exposure Assessment, Group RD 689.

Table 1. Summary of Performance for $\hat{\mu}$, $\bar{X}$, and $\bar{X}^*$ based on data resampled from chromium concentrations with 10% resampled from hot spots. Prior weight is $\alpha(R^+) = 1$.

$n+m$		$\hat{\mu}$	$\bar{X}$	$\bar{X}^*$
10	$\lvert bias \rvert$	0.0245	0.0113	3.2428
	mse	7.9808	10.7186	23.265
	acf	3.2761	3.6881	6.8814
	mcf	0.3758	0.2367	0.4204
20	$\lvert bias \rvert$	0.0110	0.0071	3.2551
	mse	4.4528	5.5317	17.1781
	acf	2.4759	2.7286	6.6031
	mcf	0.3711	0.4124	0.2776
50	$\lvert bias \rvert$	0.0031	0.0037	3.2632
	mse	1.8776	2.2413	13.286
	acf	1.6289	1.7751	6.5296
	mcf	0.3562	0.3966	0.5678

Table 2. Summary of Performance for $\hat{\mu}$, $\bar{X}$, and $\bar{X}^*$ based on data resampled from chromium concentrations with 10% resampled from hot spots. Prior weight is $\alpha(R^+) = 20$.

$n+m$		$\hat{\mu}$	$\bar{X}$	$\bar{X}^*$
10	$\lvert bias \rvert$	0.1888	0.0171	3.2138
	mse	1.0793	10.8609	22.879
	acf	1.1514	3.6747	6.8148
	mcf	0.0210	0.4206	0.1886
20	$\lvert bias \rvert$	0.0798	0.0357	3.2974
	mse	1.2349	5.7759	17.700
	acf	1.2810	2.7851	6.6924
	mcf	0.1683	0.4149	0.2887
50	$\lvert bias \rvert$	0.0346	0.0024	3.2619
	mse	0.9655	2.2111	13.275
	acf	1.1594	1.7681	6.5273
	mcf	0.2693	0.3906	0.5687

Table 3. Risk (based on squared error loss) for various prior distributions, using n = 18, m = 2 and prior weights $\alpha(R^+) = (1, 10, 20)$.

distribution	$\alpha(R^+)$	$R(\hat{S}, S_{epa})$
Weibull	1	0.29
	10	0.30
	20	0.30
Normal	1	0.29
	10	0.26
	20	0.25
Lognormal	1	0.30
	10	0.33
	20	0.35

Figure 1: Cumulative distribution function of the best fitting lognormal (bottom) vs. empirical distribution function (top)

Figure 2: Effect of censoring on $R(\hat{S}, S_{epa}; \lambda)$. Simulation based on n = 18, m = 2, $H_1(t; \lambda) = 1 - e^{-t/\lambda}$ and $H_2(t) = 0$, $t > 0$. Top curve represents $\alpha(R^+)$=10, and lower curve is $\alpha(R^+)$=1.

Figure 3: Effect of mean misspecification on prior with n = 9, m = 1.

Figure 4: Effect of mean misspecification on prior with n = 18, m = 2.