Statistics and Its Interface Volume 7 (2014) 455–463

Bayesian nonparametric density estimation for doubly-truncated data

Yuhui Chen and Timothy Hanson[*]

A Bayesian nonparametric density estimator is presented for doubly-truncated data. The estimator is based on a Pólya tree prior, and readily extended to truncated regression. The approach nicely blends a standard parametric normal fit with the nonparametric maximum likelihood estimator. Since the density is directly modeled, a standard likelihood approach applies; inference is efficiently obtained through an adaptive Markov chain Monte Carlo and no manual tuning is required. The estimator is broadly illustrated on simulated data, the quasar luminosity data of Efron and Petrosian (1999), times of cancer diagnosis considered in Moreira and Uña-Álvarez (2012), and the AIDS induction time data of Lagakos, Barraj, and De Gruttola (1988).

Keywords and phrases: Pólya tree, Regression, Truncation.

1. INTRODUCTION

Doubly-truncated data arise across many disciplines, and are usually the result of cross-sectional or retrospective sampling within a time 'window' imposed by a database, or where data are only seen above or below some threshhold. Examples from epidemiology include the time to HIV infection via blood transfusion (Bilker and Wang, 1996), and diagnosis times of childhood cancer in Portugal (Moreira and Uña-Álvarez, 2012). An urban planning example, the life spans of buildings in and around Taiwan is investigated by Chi, Tsai, and Hu (2004). An example from astronomy, the luminosity of quasars, is provided by Efron and Petrosian (1999). The analysis of heights of military personnel (where there are minimum height requirements for entry) is given by A'Hearn (2004).

This paper focuses on the Bayesian nonparametric estimation of densities in the presence of doubly-truncated data. Truncated data occurs when a random variable x_i is only observed to occur because it happened in an interval (a_i, b_i). For example, if x_i is time at which prostrate cancer is diagnosed, but the database only has information concerning individuals who have retired, then we can only observe $x_i|x_i > r_i$ years, where r_i is when subject i retired.

The Bayesian approach that we develop essentially blends the frequentist nonparametric maximum likelihood estimate (NPMLE) with a parametric estimate, so we first consider a brief description of the NPMLE based on observed data $x_1, \ldots, x_n$, originally formulated by Turnbull (1976). Consider the nonparametric 'model' $F = \sum_{i=1}^n f_i \delta_{x_i}$ where δ_x is Dirac measure on x, x_i are the observed data points truncated to $R_i = (a_i, b_i)$, and $\mathbf{f} = (f_1, \ldots, f_n)'$ contains probabilities that sum to one with a probability f_i on x_i. Let $x \sim F$. Then for each interval R_i,

$$F_i = P(x \in R_i) = \sum_{j:x_j \in R_i} P(x = x_j) = \sum_{j:x_j \in R_i} f_j.$$

In matrix form this is

$$\mathbf{F} = \mathbf{J}\mathbf{f}. \tag{1}$$

Where $\mathbf{J}$ is the $n \times n$ matrix defined as $\mathbf{J}_{ij} = 1$ if $x_j \in R_i$, otherwise 0. Now consider the likelihood, $\mathcal{L}(F) = \prod_{i=1}^n f_i/F_i$. Taking the partial derivatives of the log-likelihood and setting equal to zero gives the likelihood equations $1/f_j = \sum_{i=1}^n J_{ij}/F_i$, or in matrix form

$$\frac{1}{\mathbf{f}} = \mathbf{J}' \frac{1}{\mathbf{F}}, \tag{2}$$

where $1/\mathbf{f}$ is the vector $(1/f_1, \ldots, 1/f_n)$ and $1/\mathbf{F}$ is defined similarly; see Efron and Petrosian (1999). The NPMLE $\hat{\mathbf{f}}$ is obtained by starting with an intial set of values for $\mathbf{f}$ and iterating between (1) and (2), rescaling after (2) so that $\sum_{j=1}^n f_j = 1$. Note that this provides a purely nonparametric estimator that is conditional on the observed truncation intervals. A full likelihood can be formed that also includes the truncation times; see Shen (2010) and Moreira and Uña-Álvarez (2012). In this paper, we follow Turnbull (1976) and consider the conditional NPMLE.

Efron and Petrosian (1999) give an improvement over Turnbull's (1976) iterative method (just described), involving the discrete hazard; they showed that the NPMLE $\hat{\mathbf{f}} = (\hat{f}_1, \ldots, \hat{f}_n)$ has the hazard function $\hat{\mathbf{h}} = (\hat{h}_1, \ldots, \hat{h}_n)$ satisfying:

$$\frac{1}{\hat{h}_j} = N_j + \sum_{i=1}^n J_{ij}\hat{Q}_i, \quad j = 1, 2, \ldots, n \tag{3}$$

where $N_j = \sum_{i=1}^n I(a_i \le x_j \le b_i)$ and $\hat{Q}_i = \hat{S}(b_i)/\hat{F}_i$ with $\hat{S}(b_i) = \sum_{m=1}^n \hat{f}_m I(x_m > b_i)$ and $\hat{F}_i = \sum_{m=1}^n \hat{f}_m J_{im} =$

[*] Corresponding author.

$\sum_{m=1}^{n} \hat{f}_m I(a_i \leq x_m \leq b_i)$. Shen (2010) later generalized Efron and Petrosian's (1999) method for doubly truncated data by an alternative derivation of $\hat{\mathbf{h}}$ and investigated the consistency and the weak convergence of the NPMLE.

Wang (1989) provided a semiparametric approach for estimation on truncated data. By recognizing that the joint likelihood of the observed data is the product of the marginal likelihood of data $\mathbf{x} = (x_1, \ldots, x_n)'$ and the conditional likelihood of the truncation intervals given $\mathbf{x}$, and assuming a parametric model for the conditional truncation probabilities, the maximum likelihood estimates of the parameters $\boldsymbol{\theta}$ can be obtained by maximizing this conditional likelihood. The estimated marginal distribution using $\hat{\boldsymbol{\theta}}$ is completed using a nonparametric approach. Moreira and Uña-Álvarez (2012) very recently introduced kernel-based density estimation for doubly truncated data. They gave two approaches by smoothing a semiparametric estimator (as in Wang, 1989) or smoothing the NPMLE (Turnbull, 1976).

The NPMLE behaves nicely when the sample size $n \to \infty$, but for moderate sample sizes, it can have problems; e.g. Woodroofe (1985, p. 168) shows that a frequentist estimator can equal unity at $\min\{x_1, \ldots, x_n\}$ with positive probability. To avoid such difficulties, Gasparini (1996) developed a nonparametric Bayesian approach for estimating F on left-truncated data based on Dirichlet processes; the estimate is the posterior expectation of F assuming the truncation values are fixed constant. Also see Tiwari and Zalkikar (1993). Along these lines, Zhou and Luan (2005) proposed another Bayesian nonparametric method for right censored and left truncated data with Dirichlet process priors; they have implemented their approach in an R function `NPBayesT`. There has been no published Bayesian nonparametric approach to doubly-truncated data that we are aware of.

In this paper we consider a Bayesian nonparametric estimator for a density for continuous variables x_i, based on a finite Pólya tree prior. The estimator blends the merits of both nonparametric and parametric modeling. Merits include increased power and efficiency when the parametric family approximately holds, and also robustness against misspecifying a completely parametric family. A potential problem is that the Pólya tree is 'centered' at a parametric family—we use the Gaussian family—and this may affect inference in small samples when data are highly non-normal. Section 2 reviews the Pólya tree prior centered at the normal family; Section 3 develops an efficient algorithm for obtaining inference using Markov chain Monte Carlo. In Section 4 we examine the estimator on simulated data and compare it to the NPMLE. Section 5 examines the quasar luminosity data first analyzed by Efron and Petrosian (1999). Section 6 compares our density estimator to the kernel-smoothed versions introduced by Moreira and Uña-Álvarez (2012). In Section 7, we extend the truncated density estimation model to truncated regression models and illustrate their use on AIDS incubation time. Section 8 concludes the paper with a brief discussion.

2. NORMAL CENTERED PÓLYA TREE PRIORS

A finite normal centered Pólya tree (PT) prior for a random distribution F with $J \in \mathbb{N}$ levels is characterized by a collection of refined partitions of $\mathbb{R}$ and associated conditional probabilities. We denote $\Phi_{\boldsymbol{\theta}}(\cdot)$ as the normal cumulative distribution function (or measure) with parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$ and the corresponding density function as $\phi_{\boldsymbol{\theta}}(\cdot)$. The random distribution F is centered at $\Phi_{\boldsymbol{\theta}}$ by first defining partition sets at level j as the intervals

$$B_{\boldsymbol{\theta}}^j(k) = \big(\Phi_{\boldsymbol{\theta}}^{-1}\{(k-1)/2^j\}, \Phi_{\boldsymbol{\theta}}^{-1}\{k/2^j\}\big), \tag{4}$$

where $j = 1, \ldots, J$ and $k = 1, \ldots, 2^j$. The set $\Pi_{\boldsymbol{\theta}}^j = \{B_{\boldsymbol{\theta}}^j(k) : k = 1, \ldots, 2^j\}$ partitions $\mathbb{R}$ up to a set of Lebesque measure zero, and furthermore $\Phi_{\boldsymbol{\theta}}\{B_{\boldsymbol{\theta}}^j(1)\} = \Phi_{\boldsymbol{\theta}}\{B_{\boldsymbol{\theta}}^j(2)\} = \cdots = \Phi_{\boldsymbol{\theta}}\{B_{\boldsymbol{\theta}}^j(2^j)\} = 2^{-j}$. Note that a 'parent' set $B_{\boldsymbol{\theta}}^j(k)$ in the partition $\Pi_{\boldsymbol{\theta}}^j$ has two 'offspring' sets $B_{\boldsymbol{\theta}}^{j+1}(2k-1)$ and $B_{\boldsymbol{\theta}}^{j+1}(2k)$ in the partition $\Pi_{\boldsymbol{\theta}}^{j+1}$, i.e. $B_{\boldsymbol{\theta}}^j(k) = B_{\boldsymbol{\theta}}^{j+1}(2k-1) \cup B_{\boldsymbol{\theta}}^{j+1}(2k)$. A Pólya tree assigns a beta prior to each pair of conditional offspring probabilities from a parent set. Define these two conditional probabilities for $B_{\boldsymbol{\theta}}^j(k)$ as

$$\begin{aligned} \mathcal{Y}^{j+1}(2k-1) &= F\big\{B_{\boldsymbol{\theta}}^{j+1}(2k-1) | B_{\boldsymbol{\theta}}^j(k)\big\} \\ \mathcal{Y}^{j+1}(2k) &= F\big\{B_{\boldsymbol{\theta}}^{j+1}(2k) | B_{\boldsymbol{\theta}}^j(k)\big\}. \end{aligned} \tag{5}$$

Note that, necessarily, $\mathcal{Y}^{j+1}(2k-1) + \mathcal{Y}^{j+1}(2k) = 1$ and $B_{\boldsymbol{\theta}}^0(1) = \mathbb{R}$. We consider the standard prior (Lavine, 1992)

$$\big\{\mathcal{Y}^{j+1}(2k-1), \mathcal{Y}^{j+1}(2k)\big\} \overset{ind.}{\sim} \text{beta}\big\{c(j+1)^2, c(j+1)^2\big\}, \tag{6}$$

where $j = 0, \ldots, J-1$ and $k = 1, \ldots, 2^j$. Following Hanson (2006a), F follows $\Phi_{\boldsymbol{\theta}}$ on sets in $\Pi_{\boldsymbol{\theta}}^J$ by requiring

$$F\big\{A | B_{\boldsymbol{\theta}}^J(k)\big\} = \Phi_{\boldsymbol{\theta}}(A) 2^{-J} \text{ for all } A \subset B_{\boldsymbol{\theta}}^J(k). \tag{7}$$

The parameter c, like the precision parameter in the Dirichlet process (Ferguson, 1973, 1974), determines how closely F follows the centering distribution, both in the prior and the posterior. With $c \to \infty$, the centering distribution $\Phi_{\boldsymbol{\theta}}$ will be obtained.

Let $\mathcal{Y} = \{\mathcal{Y}^j(k) : j = 1, \ldots, J; k = 1, \ldots, 2^j\}$ and $\lceil x \rceil$ denote the ceiling function, the smallest integer greater than or equal to x. Define the vector of probabilities on sets in the finest partition $\Pi_{\boldsymbol{\theta}}^J$ as $\mathbf{V}_{\mathcal{Y}} = (V_{\mathcal{Y}}(1), V_{\mathcal{Y}}(2), \ldots, V_{\mathcal{Y}}(2^J))'$, where $V_{\mathcal{Y}}(k) = F\{B_{\boldsymbol{\theta}}^J(k)\} = \prod_{j=1}^J \mathcal{Y}^j(\lceil 2^{j-J} k \rceil)$, the product of J conditional probabilities associated with the set in each partition $\Pi_{\boldsymbol{\theta}}^j$. This product is the result of the multiplication rule for conditional probabilities and the nesting

$$\begin{aligned} B_{\boldsymbol{\theta}}^1\big(\lceil k/2^{J-1} \rceil\big) &\supset B_{\boldsymbol{\theta}}^2\big(\lceil k/2^{J-2} \rceil\big) \\ &\supset \cdots \supset B_{\boldsymbol{\theta}}^{J-1}\big(\lceil k/2 \rceil\big) \supset B_{\boldsymbol{\theta}}^J(k). \end{aligned}$$

Given $\mathcal{Y}$ and $\boldsymbol{\theta}$, the density associated with F is given by:

$$f(x|\mathcal{Y},\boldsymbol{\theta}) = 2^J V_{\mathcal{Y}}\{k_{\boldsymbol{\theta}}(J;x)\}\phi_{\boldsymbol{\theta}}(x) \tag{8}$$

where $k_{\boldsymbol{\theta}}(j;x) = \lceil 2^j \Phi_{\boldsymbol{\theta}}(x)\rceil$ and the corresponding cumulative distribution function is

$$\begin{aligned} F(x|\mathcal{Y},\boldsymbol{\theta}) &= \left[\sum_{k=1}^{k_{\boldsymbol{\theta}}(J;x)-1} V_{\mathcal{Y}}(k)\right] \\ &\quad + V_{\mathcal{Y}}\{k_{\boldsymbol{\theta}}(J;x)\}\big[2^J\Phi_{\boldsymbol{\theta}}(x) - k_{\boldsymbol{\theta}}(J;x)+1\big]. \end{aligned} \tag{9}$$

The prior specified by (4), (5), (6), and (7) is referred to as a finite Pólya tree prior with J levels, centered at the $\Phi_{\boldsymbol{\theta}}$ distribution with precision c, and is denoted

$$F|\boldsymbol{\theta}, c \sim PT_J(c, \Phi_{\boldsymbol{\theta}}). \tag{10}$$

Given $\boldsymbol{\theta}$, the random aspects of F are the elements of $\mathcal{Y}$. The random F is centered at $\Phi_{\boldsymbol{\theta}}$ in the sense that $E\{F(A)\} = \int_A \phi_{\boldsymbol{\theta}}(x)dx$ for all measurable $A \subset \mathbb{R}$.

3. DENSITY ESTIMATION

The data is assumed to be generated in the following manner. Random interval endpoints (a_i, b_i), including $\pm\infty$, arise from a bivariate distribution G. Given (a_i, b_i), x_i is generated according to F restricted to the interval (a_i, b_i). Note that if $(a_i, b_i) = \mathbb{R}$ then x_i is untruncated, i.e. observed as usual. The data is denoted $\mathcal{D} = \{(x_i, a_i, b_i)\}_{i=1}^n$.

Assume the Pólya tree model (10). We are interested in estimating the density $f(\cdot)$ and cumulative distribution function $F(\cdot)$. The likelihood of $(\mathcal{Y}, \boldsymbol{\theta})$ is given by

$$\mathcal{L}(\mathcal{Y},\boldsymbol{\theta}) = \prod_{i=1}^n \frac{f(x_i|\mathcal{Y},\boldsymbol{\theta})}{F(b_i|\mathcal{Y},\boldsymbol{\theta}) - F(a_i|\mathcal{Y},\boldsymbol{\theta})}, \tag{11}$$

where $f(\cdot|\mathcal{Y},\boldsymbol{\theta})$ and $F(\cdot|\mathcal{Y},\boldsymbol{\theta})$ are defined in (8) and (9). Commonly used priors for μ and σ^{-2} for the underlying normal case are $\mu \sim N(m, v)$, independent of $\sigma^{-2} \sim \Gamma(a, b)$, where m, v, a, and b are hyper-parameters. We have found some small amount of prior information for σ^2 to greatly improve Markov chain Monte Carlo mixing. Throughout this paper we set $a = b = 1$; this prior on σ^{-2} implies $P(0.47 < \sigma < 9.97) = 0.98$. The prior for the precision parameter c is $c \sim \Gamma(5, 1)$ (Hanson, 2006), independent of μ and σ^2. The posterior density of $(\mathcal{Y}, \boldsymbol{\theta}, c)$ is proportional to

$$\begin{aligned} \pi(\mathcal{Y},\boldsymbol{\theta},c|\mathcal{D}) &\propto \mathcal{L}(\mathcal{Y},\boldsymbol{\theta}) \times \phi_{m,v}(\mu) \times \Gamma^{-1}\big(\sigma^2|a,b\big) \\ &\quad \times \Gamma(c|5,1) \times \prod_{j=1}^J \prod_{k=1}^{2^{j-1}} \mathrm{beta}\big(\mathcal{Y}^j(2k-1)|cj^2, cj^2\big). \end{aligned}$$

Due to truncation, the full conditional distributions are intractable and Gibbs sampling is out of the question. Through extensive trial and error, a simple, fast, and highly effective approach is to use componentwise adaptive random-walk Metropolis-Hastings (ARWMH) (Haario, Saksman, and Tamminen, 2005) for μ, $\log(\sigma)$, and $\log(c)$, coupled with an adaptive block update of all the logit-transformed conditional probabilities (Haario, Saksman, and Tamminen, 2001). Let $i(j,k) = 2^{j-1} + k - 1$ and $w_{i(j,k)} = \log\{\mathcal{Y}^j(2k-1)/[1 - \mathcal{Y}^j(2k-1)]\}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, 2^{j-1}$. Then $\mathbf{w} = (w_1, \ldots, w_{2^J-1})'$ contains the logit-transformed conditional probabilities; note that $\mathcal{Y}^j(2k) = 1 - \mathcal{Y}^j(2k-1)$. Crude starting values for μ and σ^2 are obtained by $\widehat{\mu}_0 = n^{-1}\sum_{i=1}^n x_i$ and $\widehat{\sigma}_0^2 = n^{-1}\sum_{i=1}^n (x_i - \widehat{\mu})^2$ respectively. c is initialized to $c_0 = 1$ and $\mathbf{w}_0 = \mathbf{0}_d$, a vector of d zeroes, where $d = 2^J - 1$. The initial tuning values of the proposal distributions in ARWMH, with respect to μ, $\log(\sigma)$, $\log(c)$, and $\mathbf{w}$, are $V_0^{\mu} = 1$, $V_0^{\sigma} = 1$, $V_0^c = 1$, and $V_0^{\mathbf{w}}$ defined as a $d \times d$ matrix with diagonal elements 0.05 and all off diagonal elements 0. Repeat steps 1 to 4 below $B + M$ times where B is the number of samples omitted during the burn-in stage, and M is the number of samples kept. In the simulation study section, $B = 30{,}000$ and $M = 30{,}000$ are used and work well to get good Markov chain mixing. Naive standard deviations are first used for m_0 iterations (we chose $m_0 = 20$ for all simulations and data analyses), then adaptive versions are used.

Algorithm for fitting Pólya tree to doubly truncated data

For $m = 1, 2, \ldots, B + M$ do:

1. Update μ via adaptive Metropolis-Hastings:
   (a) $\mu^* \sim N(\mu_{m-1}, V_{m-1}^{\mu})$; $\rho = \frac{\pi(\mathcal{Y}_{m-1},\boldsymbol{\theta}^*,c_{m-1})}{\pi(\mathcal{Y}_{m-1},\boldsymbol{\theta}_{m-1},c_{m-1})}$ where $\boldsymbol{\theta}^* = (\mu^*, \sigma_{m-1})$ and $\boldsymbol{\theta}_{m-1} = (\mu_{m-1}, \sigma_{m-1})$.

   (b) $u \sim U(0,1)$; if $u < \rho$ then set $\mu_m = \mu^*$, otherwise $\mu_m = \mu_{m-1}$.

   (c) After m_0 iterations, update $V_m^{\mu} = \frac{m-2}{m-1}V_{m-1}^{\mu} + \frac{sd}{m-1}((m-1)\bar{\mu}_{m-2}^2 - m\bar{\mu}_{m-1}^2 + \mu_{m-1}^2) + sd\epsilon$ where $\bar{\mu}_m = \sum_{k=1}^m \mu_k/m$ and μ_k is the sampled μ at the kth iteration. Here, we set $sd = 0.02$ and $\epsilon = 0.001$.

2. Update σ via adaptive Metropolis-Hastings:
   (a) $\log(\sigma^*) \sim N(\log(\sigma_{m-1}), V_{m-1}^{\sigma})$; $\rho = \frac{\pi(\mathcal{Y}_{m-1},\boldsymbol{\theta}^*,c_{m-1})}{\pi(\mathcal{Y}_{m-1},\boldsymbol{\theta}_{m-1},c_{m-1})}$ where $\boldsymbol{\theta}^* = (\mu_m, \sigma^*)$ and $\boldsymbol{\theta}_{m-1} = (\mu_m, \sigma_{m-1})$.

   (b) Generate $u \sim U(0,1)$; if $u < \rho$ then set $\sigma_m = \sigma^*$, otherwise $\sigma_m = \sigma_{m-1}$.

   (c) After m_0 iterations, update $V_m^{\sigma} = \frac{m-2}{m-1}V_{m-1}^{\sigma} + \frac{sd}{m-1}((m-1)\overline{\log(\sigma)}_{m-2}^2 - m\overline{\log(\sigma)}_{m-1}^2 + \log(\sigma)_{m-1}^2) + sd\epsilon$, where $\overline{\log(\sigma)}_m = \sum_{k=1}^m \log(\sigma)_k/m$ and $\log(\sigma)_k$ is the sampled $\log(\sigma)$ at the kth iteration. We set $sd = 0.2$.

3. Update c via adaptive Metropolis-Hastings:
   (a) $\log(c^*) \sim N(\log(c_{m-1}), V^c_{m-1})$; $\rho = \frac{\pi(\mathcal{Y}_{m-1}, \boldsymbol{\theta}_m, c^*)}{\pi(\mathcal{Y}_{m-1}, \boldsymbol{\theta}_m, c_{m-1})}$
   (b) Generate $u \sim U(0,1)$; if $u < \rho$ then set $c_m = c^*$, otherwise $c_m = c_{m-1}$.
   (c) After m_0 iterations, update $V^c_m = \frac{m-2}{m-1} V^c_{m-1} + \frac{sd}{m-1}((m-1)\overline{\log(c)}^2_{m-2} - m\overline{\log(c)}^2_{m-1} + \log(c)^2_{m-1}) + sd\epsilon$, where $\overline{\log(c)}_m = \sum_{k=1}^m \log(c)_k/m$ and $\log(c)_k$ is the sampled $\log(c)$ at the kth iteration. We set $sd = 0.5$.

4. Update $\mathbf{w} \in \mathbb{R}^{2^J-1}$ in a block via adaptive Metropolis-Hastings:
   (a) $\mathbf{w}^* \sim N(\mathbf{w}_{m-1}, V^{\mathbf{w}}_{m-1})$; obtain $\mathcal{Y}^*$ from logistic transformations on elements of $\mathbf{w}^*$; $\rho = \frac{\pi(\mathcal{Y}^*, \boldsymbol{\theta}_m, c_m)}{\pi(\mathcal{Y}_{m-1}, \boldsymbol{\theta}_m, c_m)}$.
   (b) Generate $u \sim U(0,1)$; if $u < \rho$ then set $\mathcal{Y}_m = \mathcal{Y}^*$ and $\mathbf{w}_m = \mathbf{w}^*$, otherwise $\mathcal{Y}_m = \mathcal{Y}_{m-1}$ and $\mathbf{w}_m = \mathbf{w}_{m-1}$.
   (c) After m_0 iterations, update $V^{\mathbf{w}}_m = \frac{m-2}{m-1} V^{\mathbf{w}}_{m-1} + \frac{sd}{m-1}((m-1)\bar{\mathbf{w}}_{m-2}\bar{\mathbf{w}}'_{m-2} - m\bar{\mathbf{w}}_{m-1}\bar{\mathbf{w}}'_{m-1} + \mathbf{w}_{m-1}\mathbf{w}'_{m-1}) + sdI_{d\times d}$ and $sd = 1.0/d$, where $\bar{\mathbf{w}}_m = \sum_{k=1}^m \mathbf{w}_k/m$ and the elements $\mathbf{w}_k \in \mathbb{R}^{2^J-1}$ are the sampled $\mathbf{w}$ column vectors at the kth iteration.

The estimate of the density at a point with respect to squared-error loss is the posterior mean

$$f(x|\mathcal{D}) \approx \frac{1}{M}\sum_{m=1}^{M} f(x|\mathcal{Y}_m, \boldsymbol{\theta}_m). \tag{12}$$

The cumulative distribution function is also estimated by the posterior mean

$$F(x|\mathcal{D}) \approx \frac{1}{M}\sum_{m=1}^{M} F(x|\mathcal{Y}_m, \boldsymbol{\theta}_m) \tag{13}$$

where $f(\cdot|\mathcal{Y}, \boldsymbol{\theta})$ and $F(\cdot|\mathcal{Y}, \boldsymbol{\theta})$ are defined in (8) and (9) respectively.

4. SIMULATION STUDY

We examine the performance of the proposed method on data simulated under two scenarios. First, data are randomly truncated to $(-\infty, 1.5)$ or $(-1.25, \infty)$, each with probability one-half, and arise from a mixture of two normal distributions $F = 0.5N(-1, 0.5) + 0.5N(1, 1)$ restricted to either truncation interval. Five estimated densities (posterior means over a grid) are plotted from different samples of size $n = 500$, $n = 1{,}000$, $n = 2{,}000$, and $n = 4{,}000$, along with the true density in Figure 1. We set $J = 5$, $m = 0$, and $v = 10$. After a burn-in of 30,000 iterates, 30,000 were kept. The chains mixed well with acceptance rates around 0.60, 0.65, 0.50, and 0.5 for μ, $\log(\sigma)$, $\log(c)$, and $\mathbf{w}$ respectively. Figure 1 shows the estimator to estimate more accurately as the sample size increases, capturing the bimodality of the true density even though F is centered at the normal distribution.

We investigate another simulation where data arises from a bimodal distribution: $x_i \sim F$ where $F = 0.5N(-1, 0.5) + 0.5N(1, 0.5)$ randomly truncated to (a_i, b_i) in the following manner. With probabilities 0.25, 0.25, and 0.5, (a) $a = -\infty$ and $b_i \sim U(0, 3)$, (b) $b_i = \infty$ and $a_i \sim U(-3, 0)$, (c) $a_i \sim U(-3, -1)$ and $b_i \sim U(1, 3)$, respectively. Five estimated densities from samples of size $n = 500$ are plotted along with the true density; see the left panel of Figure 2. We keep $J = 5$, $m = 0$, $v = 10$, and $M = 30{,}000$ after a burn-in of $B = 30{,}000$ iterations. The Markov chains mix well with acceptance rates 0.58, 0.52, 0.53, and 0.32, for μ, $\log(\sigma)$, $\log(c)$, and $\mathbf{w}$. To study the influence of the proportion of truncation, we also generate data from the same $F = 0.5N(-1, 0.5) + 0.5N(1, 0.5)$ but under a different truncation scenario, (a) $a = -\infty$ and $b_i \sim U(1, 3)$, (b) $b_i = \infty$ and $a_i \sim U(-3, -1)$, (c) $a_i \sim U(-3, -1.5)$ and $b_i \sim U(1.5, 3)$ with probabilities 0.25, 0.25, and 0.50 respectively; these data have typically larger truncation intervals, so there is more information for the distribution function and density. Accordingly, the density estimates in the right panel of Figure 2 are more accurate (less variability and more bias), as one would expect.

We investigate one last simulation and compare our estimated density and cumulative distribution functions with the ones obtained from Efron and Petrosian (1999). The `DTDA` package for **R** (Moreira et al., 2010) includes the function `efron.petrosian` for obtaining estimation results from the approach of Efron and Petrosian (1999). Data are right-truncated to lie in the interval $(0, u_i)$ where $u_i \sim U(0.5, 3)$; the true density is exponential, $F(x) = 1 - e^{-x}$ for $x > 0$. The sample size considered is $n = 500$. We transform both the data and the truncation intervals using the natural log (so the resulting log-transformed data are right-truncated extreme value). The estimated density and cumulative distribution functions are plotted in Figure 3 with $J = 5$ and other prior settings as above. To obtain a density from Turnbull's (1976) estimate, we simply difference the estimated cumulative distribution function across a bin and set the bin area to this value. Since the density is not identifiable past the largest value $x_{(n)}$, we consider conditional density estimation on $(-\infty, x_{(n)})$. The Bayesian estimator "smooths" Turnbull's NPMLE estimator somewhat toward the normal centering density.

5. QUASARS DATA

Efron and Petrosian (1999) investigated truncated quasar luminosity data via the NPMLE. The data presented there are independently collected quadruplets, denoted as (z_i, m_i, a_i, b_i), for $i = 1, 2, \ldots, n$ with $n = 210$. Here z_i is the redshift of the ith quasar, m_i is the magnitude; the lower bound a_i is used to distinguish from non-quasar objects, while the upper bound b_i is used to exclude the quasars with magnitude above b_i, which cannot be observed

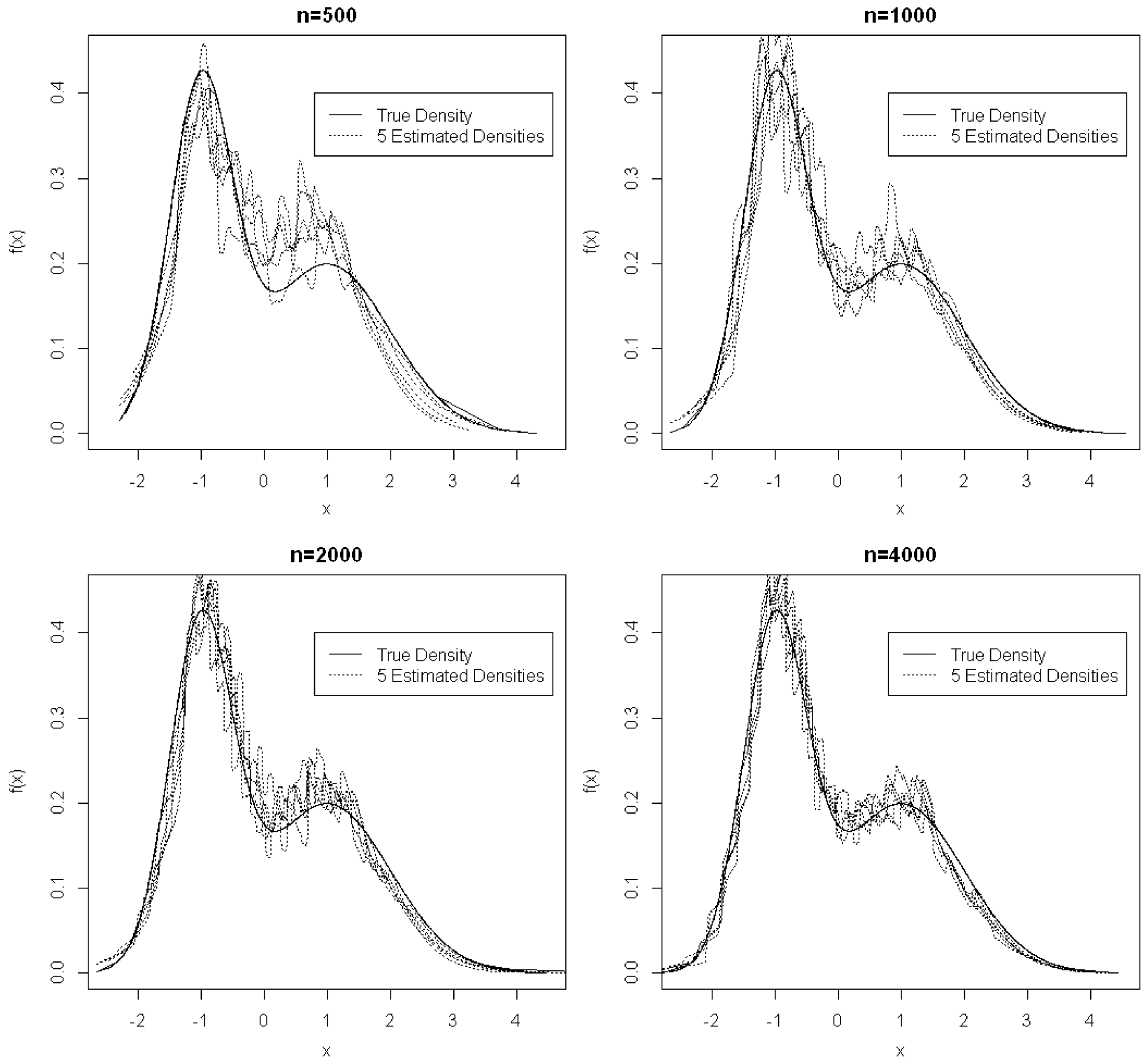

Figure 1. Estimated densities from five data sets randomly generated at four different sample sizes $n = 500$, $n = 1{,}000$, $n = 2{,}000$, *and* $n = 4{,}000$.

since they are too dim to yield dependent redshifts. We obtained the quasars data from the **R** package DTDA (Moreira, Uña-Álvarez, and Crujeiras, 2010); the distribution of each luminosity in the log-scale $x_i = H(z_i, m_i)$ is truncated to an interval (a_i, b_i), where $H(\cdot)$ represents a transformation depending on the cosmological model (Efron and Petrosian, 1999). We plot the estimated density and cumulative functions from the proposed method as well as the NPMLE method given in Efron and Petrosian (1999) in Figure 4. Efron and Petrosian (1999) also considered a parametric density of the form $\log\{f_d(x)\} = \sum_{j=0}^{d} \beta_j x^j$ over a finite interval (the degree of the polynomial d picked by large-sample hypothesis tests). Our cumulative distribution function tracks the NPMLE reasonably well, but allows for extrapolation to the left of a_i if one has faith in the underlying normal model. As before, the NPMLE density estimate is a simple histogram constructed by differencing the NPMLE and making the bin areas add up to one. In the next section we compare our the Bayesian nonparametric density estimate to a kernel-smoothed NPMLE. FORTRAN code for carrying out the analysis on the quasar luminosity data is available from the authors by request.

6. BAYESIAN AND KERNEL-SMOOTHED DENSITY ESTIMATES ON A CHILDHOOD CANCER DATA SET

Frequentist estimation of a density under double-truncation was only recently proposed by Moreira and Uña-Álvarez (2012). Moreira and Uña-Álvarez's nonparametric density estimator kernel-smooths Turnbull's NPMLE (presented in the introduction). They also consider a kernel-

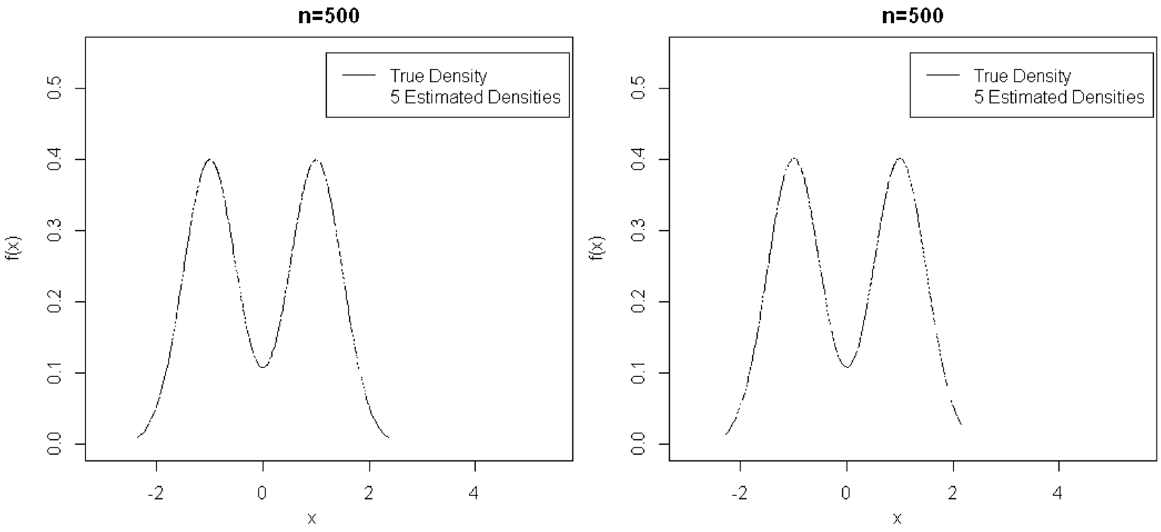

Figure 2. Estimated densities from five doubly-truncated data of size $n = 500$; left panel has smaller truncation intervals.

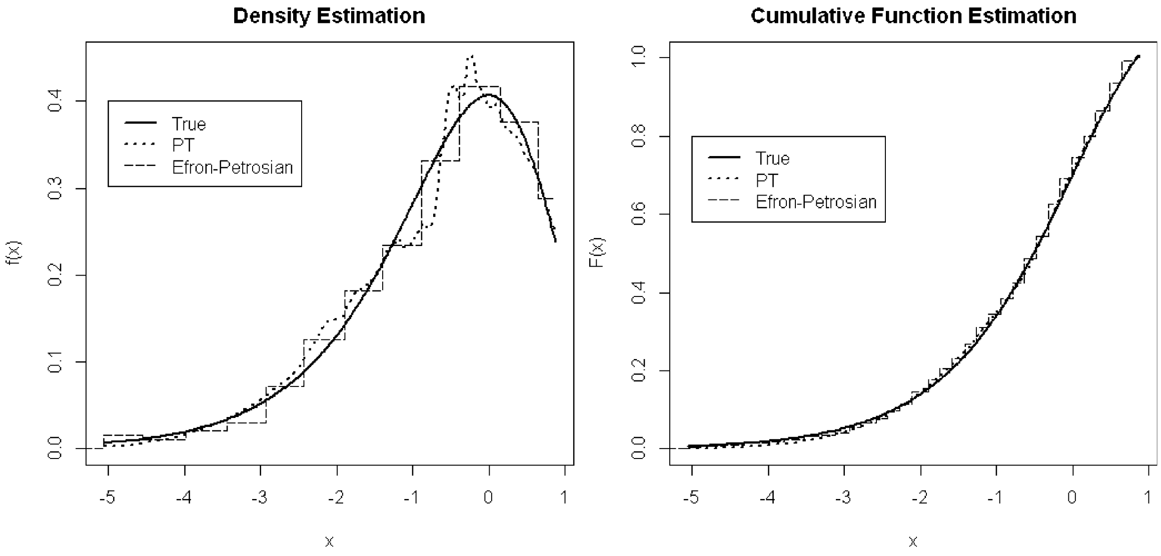

Figure 3. Density and cumulative function estimations from the proposed method and the method presented in Efron and Petrosian (1999).

smoothed semiparametric estimator as well which specifies a parametric distribution G for the conditional densities of the truncation times (a, b), given x. Moreira and Uña-Álvarez (2012) illustrate their estimator on a data set comprised of $n = 406$ diagnosis times in years for all childhood cancers in North Portugal between January 1, 1999 and December 31, 2003. Diagnosis time x_i is truncated to lie within (a_i, b_i) where b_i is the years between the child's birth and December 31, 2003; a_i is $b_i - 5$ (5 years is the window in which the study took place). All of x_i, a_i, and b_i are transformed to lie within $[0, 1]$ via $t(x) = (x + 5)/20$.

Figure 5 gives our estimator for these data with $J = 5$, $c \sim \Gamma(10, 1)$, and 30,000 iterates kept after a burn-in of 15,000. Also on the plot are the nonparametric, semiparametric, and naive kernel-smoothed estimators of Moreira and Uña-Álvarez (2012) for bandwidth $h = 0.035$. The naive kernel-smoothed estimator does not correct for truncation. Overall, the Bayesian estimate tracks the frequentist kernel-smoothed estimate up until about 0.8 (diagnosis at 11 years), where the Bayesian estimate noticeably drops below the others. This is likely a result of the Bayesian estimator attenuating to zero in the tails more quickly due to the centering family; this may or may not be more accurate depending on whether the true density also dies down more quickly. It is not our intention to provide a comprehensive comparison between these density estimators here, but

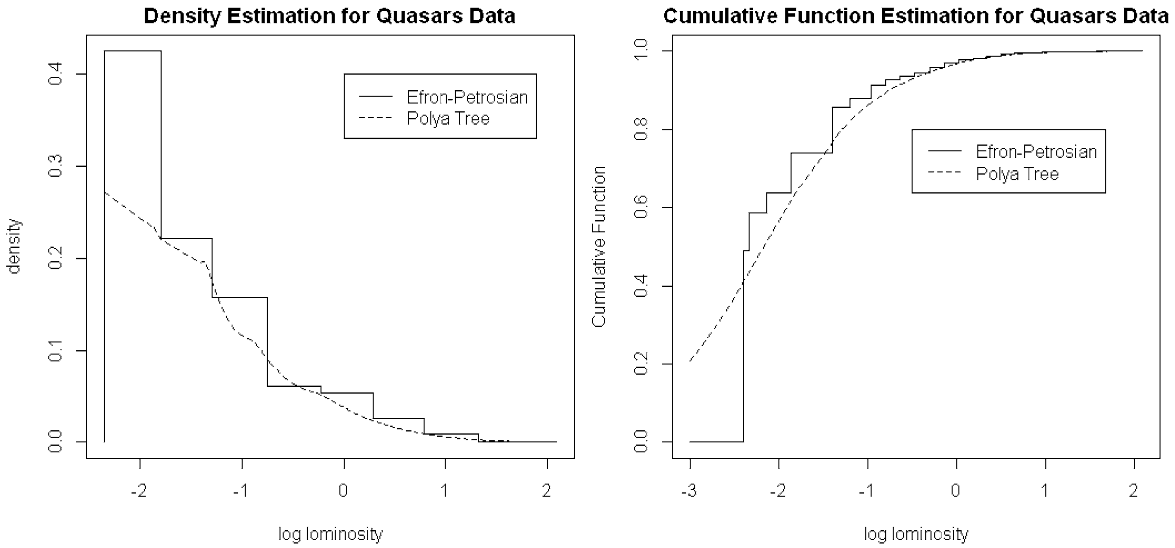

Figure 4. Density and cumulative function estimations for quasars data from the proposed method and the method presented in Efron and Petrosian (1999).

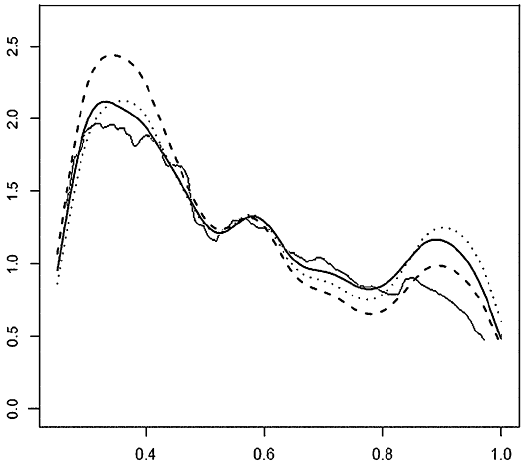

Figure 5. Doubly-truncated Portuguese childhood cancer density estimates; kernel-smoothed NPMLE (smooth-solid), Pólya tree estimate (spiky-solid), kernel-smoothed semiparametric (dashed), and kernel-smoothed estimate ignoring truncation (dotted).

rather show that they give similar results for this data set. Hanson (2006b) compared kernel-smoothed and Bayesian nonparametric density estimators (in this case a Dirichlet process mixture) in a small simulation study and found the Bayesian approach improved estimatation in moderate-to-large sample sizes; the kernel-smoothed estimator worked better for small samples. Also see Xu et al. (2013) for an extensive comparison between a Bayesian nonparametric density estimator and kernel-smoothed density estimator for the null hypothesis test of unimodality; the Bayesian test does significantly better overall.

7. TRUNCATED REGRESSION

The methods herein can be extended to semiparametric survival modeling using, e.g. linear models, for doubly-truncated and censored data. In particular, the doubly-truncated model can be extended to include covariates by simply specifying $\boldsymbol{\theta}_i = (\mathbf{z}_i'\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients and $\mathbf{z}_i$ is a p-dimensional vector of covariates associated with individual i. This is equivalent to the model

$$x_i = \mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} F, \quad F|c, \sigma \sim PT_J(c, \Phi_\sigma),$$

where Φ_σ is a mean-zero Gaussian distribution with variance σ^2. Furthermore the probabilities at the first level are fixed to $\mathcal{Y}^1(1) = \mathcal{Y}^1(2) = 0.5$ to ensure identifiability; this produces a median regression model. The algorithm for obtaining posterior inference is changed slightly to update $\boldsymbol{\beta}$ rather than μ using a block update akin to that used for the logit-transformed elements of $\mathcal{Y}$ in $\mathbf{w}$. This model without truncation was considered by Hanson and Johnson (2002), and further generalized to heteroscedastic error by Jara and Hanson (2011). A purely nonparametric frequentist method was introduced by Lewbell and Linton (2002).

We applied this model to the HIV incubation data of Lagakos, Barraj, and De Gruttola (1988), also analyzed by Shen (2012). There are 37 children and 258 adults in the data set who were infected with HIV through blood transfusion. The induction time x_i (time from HIV infection to diagnosis of AIDS) is truncated to $R_i = (0, 8 - y_i) = (a_i, b_i)$ where y_i is the infection time and 8 years is the total time by the end of the study (April 1977 through January 1986). The truncated Pólya tree model was fit where $\boldsymbol{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta}$, $\mathbf{z}_i = (1, 0)$ for children and $\mathbf{z}_i = (1, 1)$ for adults. We

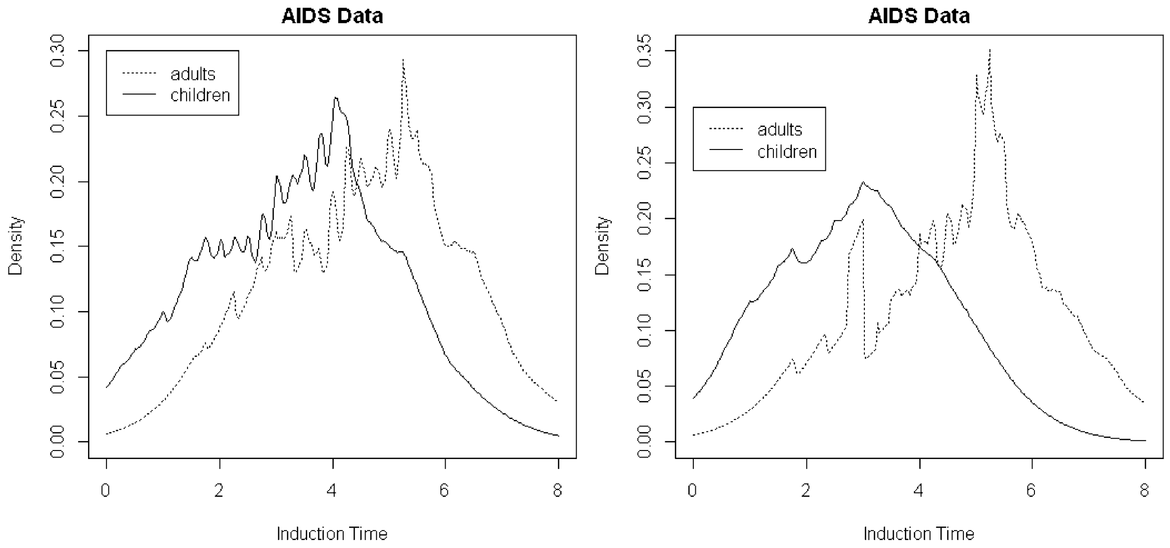

Figure 6. Density estimates for AIDS induction data; regression model (left) and independent fits (right).

considered the model with $J = 5$, $\boldsymbol{\beta} \sim N_2(\mathbf{0}_2, 100\mathbf{I}_2)$, $c \sim \Gamma(5, 1)$, and $\sigma^{-2} \sim \Gamma(0.1, 0.1)$ for illustrative purposes. After a burn-in of 30,000 iterates, 170,000 were kept; the Markov chain mixing was reasonably good based on history plots.

The posterior mean and 95% credible interval for β_0 is 3.51 (3.18, 3.96) and for β_1 is 1.20 (0.74, 1.50). The estimated median time to induction is estimated to be 3.51 years for children and $3.51 + 1.20 = 4.71$ years for adults. This agrees closely with Lagakos et al. (1988). The estimated densities for children and adults are plotted in the left panel of Figure 6. Fitting each group separately produces the right panel in Figure 6, with estimated median time to induction 3.45 years with 95% CI (2.31, 4.52) for children and 4.87 years for adults with 95% CI (4.59, 5.13). All density estimates are "spiky", but it is important to note that they *should* be spiky. The data were recorded only to the nearest quarter year (i.e. in three month increments), so there is a great deal of "ties" among the observed truncated induction times and these plots simply reflect this "piling up of mass" at these relatively few discrete values. Note that the NPMLE is also spiky—it places mass *only* on the observed tied induction times. When fit separately, the children's density estimate is smoother, reflecting that only 37 observations go into it, whereas 258 go into the adults' density.

Other authors have shown that Bayes estimators arising from nonparametric priors often mimic frequentist estimates mixed with parametric fits. Susarla and Van Ryzen (1976) show their survival curves reduce to Kaplan-Meier (1958) estimators when the precision gets small. Kalbfleisch (1978) finds Cox's (1975) partial likelihood as a limiting case of the gamma process proportional hazards model. Johnson and Christensen (1986) obtain Turnbull's (1976) estimator from grouped interval censored data.

8. DISCUSSION

This paper considers the nonparametric (or rather, richly parametric) estimation of a density in the presence of doubly-truncated data. The density is modeled as a finite Pólya tree with a type of "shrinkage prior" on the Pólya tree conditional probability parameters. The estimator works well in simulations and is compared to the NPMLE on simulated and real data. The Pólya tree resembles a smoothed version of the NPMLE, but conveys some additional advantages. Since the Pólya tree is centered at a parametric model, the tails are estimated and so extrapolation is possible, if one has some confidence in the centering family. The univariate model can be extended to bivariate truncated data as suggested in Yang, Hanson, and Christensen (2008); however, dimensions higher than two are problematic as the number of Pólya tree parameters grows exponentially with dimension unless the Pólya tree is marginalized.

We are currently working on fast approximations to maximizing the posterior with application to the k-sample problem with truncated data. Bilker and Wang (1996) suggest a test using a semiparametric approach akin to Wang (1989); note that Lagakos, Barraj, and De Gruttola (1988) consider a nonparametric approach to purely left- or right-truncated data whereas Chi, Tsai, and Hu, (2004) consider doubly-truncated discrete data. Chen and Hanson (2014) use Pólya trees for the k-sample problem in the presence of continuous, censored data with excellent results. Their method, which involves marginalizing the random F, cannot be directly applied here. However, if one rather takes the Pólya tree partition to coincide with the endpoints of censoring and truncation intervals, left-truncated and right-censored data can be accommodated rather easily, even with F marginalized.

Received 27 July 2013

REFERENCES

[1] A'Hearn, B. (2004). A restricted maximum likelihood estimator for truncated height samples. *Economics and Human Biology* **2** 5–20.

[2] Bilker, W. and Wang, M. (1996). A semiparametric extension of the Mann-Whitney test for randomly truncated data. *Biometrics* **52** 10–20.

[3] Chen, Y. and Hanson, T. (2014). Bayesian nonparametric k-sample tests for censored and uncensored data. *Computational Statistics and Data Analysis* **71** 335–346. MR3131974

[4] Chi, Y., Tsai, W., and Hu, C. (2004). Testing the equality of two survival functions with doubly truncated data. *Journal of the Chinese Statistical Association* **42** 223–244.

[5] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509

[6] Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association* **94** 824–834. MR1723343

[7] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230. MR0350949

[8] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2** 615–629. MR0438568

[9] Gasparini, M. (1996). Nonparametric Bayesian estimation of a distribution function with truncated data. *Journal of Statistical Planning and Inference* **55** 361–369. MR1422139

[10] Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

[11] Haario, H., Saksman, E., and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics* **20** 265–273. MR2323976

[12] Hanson, T. (2006a). Inference for mixtures of finite Pólya tree models. *Journal of the American Statistical Association* **101** 1548–1565. MR2279479

[13] Hanson, T. (2006b). Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Analysis* **1** 575–594. MR2221289

[14] Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of Pólya trees. *Journal of the American Statistical Association* **97** 1020–1033. MR1951256

[15] Jara, A. and Hanson, T. (2011). A class of mixtures of dependent tailfree processes. *Biometrika* **98** 553–566. MR2836406

[16] Johnson, W. O. and Christensen, R. (1986). Bayesian nonparametric survival analysis for grouped data. *Canadian Journal of Statistics* **14** 307–314. MR0876756

[17] Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40** 214–221. MR0517442

[18] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53** 457–481. MR0093867

[19] Lagakos, S., Barraj, L., and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **75** 515–523. MR0967591

[20] Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *The Annals of Statistics* **20** 1222–1235. MR1186248

[21] Lewbell, A. and Linton, O. (2002). Nonparametric censored and truncated regression. *Econometrica* **70** 765–779. MR1913830

[22] Moreira, C. and Uña-Álvarez, J. (2012). Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics* **6** 501–521. MR2988417

[23] Moreira, C., Uña-Álvarez, J., and Crujeiras, R. (2010). DTDA: An **R** package to analyze randomly truncated data. *Journal of Statistical Software* **37** Issue 7.

[24] Shen, P. (2010). Nonparametric analysis of doubly truncated data. *Annals of the Institute of Statistical Mathematics* **62** 835–853. MR2669740

[25] Shen, P. (2013). Regression analysis of interval censored and doubly truncated data with linear transformation models. *Computational Statistics* **28** 581–596. MR3064469

[26] Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association* **71** 897–902. MR0436445

[27] Tiwari, R. and Zalkikar, J. N. (1993). Nonparametric Bayesian estimation of survival function under random left truncation. *Journal of Statistical Panning and Inference* **35** 31–45. MR1220403

[28] Turnbull, B. (1976). The empirical distribution with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38** 290–295. MR0652727

[29] Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association* **84** 742–748. MR1132590

[30] Woodroofe, M. (1985). Estimating a distribution function with truncated data. *The Annals of Statistics* **13** 163–177. MR0773160

[31] Xu, L., Bedrick, E., Hanson, T., and Restrepo, C. (2014). A comparison of statistical tools for identifying modality in body mass distributions. *Journal of Data Science* **12** 175–196.

[32] Yang, M., Hanson, T., and Christensen, R. (2008). Nonparametric Bayesian estimation of a bivariate density with interval censored data. *Computational Statistics and Data Analysis* **52** 5202–5214. MR2526586

[33] Zhou, M. and Luan, J. (2005). Nonparametric Bayes estimator of survival function for right-censoring and left-truncation data. *Technical Report*, University of Kentucky, Department of Statistics.

Yuhui Chen
Department of Statistics
University of South Carolina, SC
USA
E-mail address: yuhchen.mail@gmail.com

Timothy Hanson
Department of Statistics
University of South Carolina, SC
USA
E-mail address: hansont@stat.sc.edu