Hindawi Publishing Corporation
Epidemiology Research International
Volume 2013, Article ID 131232, 8 pages
http://dx.doi.org/10.1155/2013/131232

*Research Article*

# Robust Medical Test Evaluation Using Flexible Bayesian Semiparametric Regression Models

**Adam J. Branscum,[1] Wesley O. Johnson,[2] and Andre T. Baron[3]**

[1] *Biostatistics Program, College of Public Health and Human Sciences, Oregon State University, Corvallis, OR 97331, USA*
[2] *Department of Statistics, University of California, Irvine, CA 92697, USA*
[3] *Tumor Biology Investment Group, Inc., Richmond, KY 40475, USA*

Correspondence should be addressed to Adam J. Branscum; adam.branscum@oregonstate.edu

Received 6 August 2013; Accepted 31 October 2013

Academic Editor: Leo J. Schouten



The application of Bayesian methods is increasing in modern epidemiology. Although parametric Bayesian analysis has penetrated the population health sciences, flexible nonparametric Bayesian methods have received less attention. A goal in nonparametric Bayesian analysis is to estimate unknown functions (e.g., density or distribution functions) rather than scalar parameters (e.g., means or proportions). For instance, ROC curves are obtained from the distribution functions corresponding to continuous biomarker data taken from healthy and diseased populations. Standard parametric approaches to Bayesian analysis involve distributions with a small number of parameters, where the prior specification is relatively straight forward. In the nonparametric Bayesian case, the prior is placed on an infinite dimensional space of all distributions, which requires special methods. A popular approach to nonparametric Bayesian analysis that involves Polya tree prior distributions is described. We provide example code to illustrate how models that contain Polya tree priors can be fit using SAS software. The methods are used to evaluate the covariate-specific accuracy of the biomarker, soluble epidermal growth factor receptor, for discerning lung cancer cases from controls using a flexible ROC regression modeling framework. The application highlights the usefulness of flexible models over a standard parametric method for estimating ROC curves.

## 1. Introduction

Bayesian analysis is often used in the support of epidemiologic research [1–7]. A contemporary area of research in the population health sciences involves the development and application of statistical methods for evaluating the accuracy of medical tests. With binary outcome data, statistical methods focus on estimating sensitivity and specificity, while with quantitative data, standard objects of interest are the receiver operating characteristic (ROC) curve and area under the curve (AUC).

The ROC curve can be regarded as a graphical portrayal of the degree of separation between the distributions of test outcomes for "diseased" and nondiseased populations. The formula for an ROC curve depends on $\text{Se}(k)$ and $\text{Sp}(k)$, the sensitivity and specificity of the test at classification threshold $k$. We proceed under the innocuous assumption that test outcomes tend to be higher for individuals in the diseased population. Let $y$ denote a continuous test outcome for a disease $D$, where disease status is labeled as $D = 0$ for disease absent and $D = 1$ for disease present. In general, $y$ can be any continuously measured classifier that varies according to a cumulative distribution function $G_0$ for nondiseased individuals and varies according to $G_1$ for diseased individuals. The ROC curve is a plot of true positive probability versus false positive probability across all classification thresholds, that is, a plot of pairs $(1 - \text{Sp}(k), \text{Se}(k))$ for all $k$. Since $\text{Se}(k) = \Pr(y > k \mid D = 1) = 1 - G_1(k)$ and $1 - \text{Sp}(k) = \Pr(y > k \mid D = 0) = 1 - G_0(k)$, accurate direct estimation of the ROC curve depends crucially on accurate estimation of the distribution functions $G_0$ and $G_1$.

This motivates the use of flexible data-driven methods for estimating functions, which is a central goal of nonparametric Bayesian analysis. The standard parametric approach

imposes the strong condition that $G_0$ and $G_1$ be members of the normal family (i.e., the normal-normal model), while a flexible nonparametric approach treats them as arbitrary continuous distribution functions. A flexible method is desirable because $G_0$ and/or $G_1$ could exhibit unanticipated right or left skewness, multimodality, or they could be symmetric but not bell-shaped. As a result, the methodology we present has the flexibility to handle data from many different settings, and notably it will be designed to include the normal-normal model as a special case.

The development of modern nonparametric and semiparametric Bayesian procedures for ROC analysis is an active area of research [8–17]. The main goals of this paper are (i) to use flexible Bayesian models and ROC curves to evaluate the accuracy of a biomarker that may depend on patient characteristics and (ii) to provide a nontechnical description of Polya tree priors and show how they can be implemented in epidemiological research using standard statistical software.

## 2. Materials and Methods

Robust inference for ROC curves stems from allowing $G_0$ and $G_1$ to be members of broad classes of distributions. Our approach is to model them using Mixtures of Finite Polya Trees (MFPT) priors, which have been carefully discussed in the statistics literature [18, 19]. We avoid a full technical description because it requires considerable mathematical detail. Instead, a conceptual introduction appears in Appendix A.

Briefly, a major advantage beyond flexibility is that MFPT priors for $G_0$ and $G_1$ can be "centered" on separate parametric families. In part, this means that the flexible model generalizes the standard parametric model since the normal-normal model is allowed as a special case. Additionally, it means that the expected value of $G_\ell$ under the prior will be the distribution function that corresponds to the centering family. In our illustrations, the centering families will be $F_\ell \equiv \{N(\mu_\ell, \sigma_\ell^2) : \mu_\ell \in \mathbb{R}, \sigma_\ell^2 \in \mathbb{R}^+\}, \ell = 0, 1$. The MFPT priors will also have positive weight parameters, $c_0$ and $c_1$, and positive integers $J_0$ and $J_1$ that define the (finite) length of the trees. Large values for the weights indicate higher prior confidence in the centering distributions, while values near zero allow for considerable flexibility around the centering family. To aid the selection of a weight parameter, consider that the parametric model $F_\ell$ is the special case obtained when $c_\ell \to \infty$. Hence, using a large weight value is similar to fitting the parametric centering model to the data (possible oversmoothing). The opposite extreme of setting $c_\ell \doteq 0$ produces empirical estimates (possible undersmoothing). In our experience, a weight parameter between 0.5 and 1 is often a positive compromise between these extremes. Tree lengths generally range between 4 and 10, depending on the sample size. Small values for $J_\ell$ tend to oversmooth the data. A fully nonparametric prior is obtained as $J_\ell \to \infty$, but large values of $J_\ell$ can substantially increase computing time. Finally, there will be priors on $(\mu_\ell, \sigma_\ell)$; write them as $p_\ell(\mu_\ell, \sigma_\ell)$, for $\ell = 0, 1$. We assume $G_0$ and $G_1$ are independent, with priors denoted by $G_\ell \sim \text{MFPT}(F_\ell, p_\ell, c_\ell)$, for $\ell = 0, 1$.

*2.1. Models.* In the absence of covariates, consider two independent samples of continuous biomarker measurements, where $y_1, y_2, \ldots, y_{n_0}$ constitute a random sample from an unknown distribution $G_0$, and $z_1, z_2, \ldots, z_{n_1}$ are realizations from $G_1$. We regard the $y_i$'s as biomarker outcomes from $n_0$ nondiseased individuals and the $z_j$'s as outcomes from $n_1$ diseased individuals. Our model is represented as $y_i \mid G_0 \overset{\text{indep}}{\sim} G_0$ with $G_0 \sim \text{MFPT}(F_0, p_0, c_0)$ and similarly for the $z_j$'s. The prior $p_\ell(\mu_\ell, \sigma_\ell)$ in our illustrations involves a normal distribution on $\mu_\ell$ and an independent uniform distribution on $\sigma_\ell$.

There are many possible extensions of this two-group model depending on the complexity of the data. We describe a semiparametric regression model that can be easily adapted to handle a variety of scenarios. The model specifies separate linear regressions with arbitrary residual distributions for the data from the nondiseased and diseased populations:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_g x_{ig} + \epsilon_i \equiv x_i\beta + \epsilon_i, \\ \epsilon_i \mid G_0 &\sim G_0, \quad G_0 \sim \text{MFPT}\left(N\left(\mu_0, \sigma_0^2\right), p_0, c_0\right); \\ z_j &= \alpha_1 \widetilde{x}_{j1} + \alpha_2 \widetilde{x}_{j2} + \cdots + \alpha_h \widetilde{x}_{jh} + \widetilde{\epsilon}_j \equiv \widetilde{x}_j\alpha + \widetilde{\epsilon}_j, \\ \widetilde{\epsilon}_j \mid G_1 &\sim G_1, \quad G_1 \sim \text{MFPT}\left(N\left(\mu_1, \sigma_1^2\right), p_1, c_1\right). \end{aligned} \tag{1}$$

There are no intercepts in the regression part of the models since each residual distribution has an arbitrary unknown mean. The covariate vectors $x_i$ and $\widetilde{x}_j$ can be distinct or have some overlap (e.g., both might include age) or complete overlap ($x_i = \widetilde{x}_j$). In the absence of available prior information, independent normal prior distributions with mean 0 and large variance can be used for the regression coefficients. Alternatively, when prior information is available, we recommend conditional means priors [20]. Special cases include a model with no covariates for the nondiseased population (no $x_{ir}$ variables), a model with no covariates for the diseased population (no $\widetilde{x}_{js}$ variables) and a model without covariates (the two-group model).

Our model can be used to estimate covariate-specific ROC curves and AUC. Let $x^* = (x, \widetilde{x})$ define two particular subpopulations of nondiseased and diseased people. Then, $\text{ROC}(x^*)$ and its corresponding $\text{AUC}(x^*)$ are measures of the tests' accuracy when applied to nondiseased people with covariate vector $x$ and diseased people with covariate vector $\widetilde{x}$ (often, the researcher will set $x$ equal to $\widetilde{x}$). Approximate posterior inference for ROC curves can be made by calculating posterior summaries of $(1 - G_0(k - x\beta), 1 - G_1(k - \widetilde{x}\alpha))$ across a fine grid of values for $k$ that spans the sample space.

## 3. Results

*3.1. Fitting MFPT Models in SAS.* The SAS 9.3 (SAS Institute, Inc., Cary, North Carolina) code that was used to fit a one-sample MFPT model to the simulated data described in Appendix A is in Appendix B.

*3.2. ROC Analysis of sEGFR Data.* We investigated a soluble isoform of the epidermal growth factor receptor (sEGFR)

as a biomarker for lung cancer in premenopausal and postmenopausal women. sEGFR also has been associated with ovarian cancer [21, 22]. The data were collected from a case-control study conducted at the Mayo Clinic in Rochester, Minnesota between 1998 and 2003. There were 140 controls and 101 lung cancer cases, and 92 premenopausal and 149 postmenopausal women enrolled in the study.

In a preliminary analysis, we found no clear evidence of a difference in the distributions of $y = \sqrt{\text{sEGFR}}$ for premenopausal versus postmenopausal lung cancer cases (Wilcoxon $P = 0.15$). However, there was evidence for a statistical difference in the distribution of $y$ for controls based on menopausal status (Wilcoxon $P < 0.01$). We thus included menopausal status as a covariate in models for the control data. A flexible approach is supported over the standard parametric approach because normal quantile plots (not shown) indicated a right skew in the distributions of $y$ for cases and postmenopausal controls.

We compared two analyses. The first analysis had menopausal status as a covariate for the control group in an ROC regression model. The second analysis modeled outcomes among pre- and postmenopausal controls with completely separate (flexible) distributions, without regression structure. We compared these models using the log pseudomarginal likelihood (LPML) [23], [24, Section 4.9].

For all models that used Polya trees, we set weight parameters (the $c_\ell$'s) equal to 0.5 and tree lengths (the $J_\ell$'s) equal to 5. In this setting, values of $y$ tended to be higher among controls instead of cases, so we reversed the roles for cases and controls. Let $\widetilde{x} = 1$ for premenopausal women and $\widetilde{x} = 0$ for postmenopausal women. For both models under consideration, the data from cases were modeled as independently distributed according to an unknown distribution $G_0$. The prior was $G_0 \sim \text{MFPT}(N(\mu_0, \sigma_0^2), p_0, c)$, with $p_0(\mu_0, \sigma_0)$ selected to be $\mu_0 \sim N(50, 400)$ independent of $\sigma_0 \sim \text{Uniform}(0, 100)$. The prior mean of 50 for $\mu_0$ was chosen because that was approximately the mean of $y$ for male lung cancer cases in a similar study, but we allowed for a high degree of uncertainty about the value of $\mu_0$ through a large prior variance.

For controls, model 1 for the data was the regression

$$\begin{aligned} z_j &= \alpha_1 \widetilde{x}_j + \widetilde{\epsilon}_j, \\ \widetilde{\epsilon}_j \mid G_1 &\sim G_1, \\ G_1 &\sim \text{MFPT}\left(N\left(\mu_1, \sigma_1^2\right), p_1, c\right), \end{aligned} \tag{2}$$

where, independently, $\alpha_1 \sim N(0, 400)$, $\mu_1 \sim N(50, 400)$, and $\sigma_1 \sim \text{Uniform}(0, 100)$.

Model 2 specified separate distributions for data from premenopausal and postmenopausal controls. Denote the distributions by $G_2$ and $G_3$, respectively, where $G_w$ was assigned an MFPT prior that was centered at $N(\mu_w, \sigma_w^2)$, with $\mu_w \sim N(50, 400)$ and $\sigma_w \sim \text{Uniform}(0, 100)$, for $w = 2, 3$. Using the same priors, we also considered the standard parametric analysis (model 3) in which $G_0$, $G_2$, and $G_3$ were normal distributions with distinct means and variances.

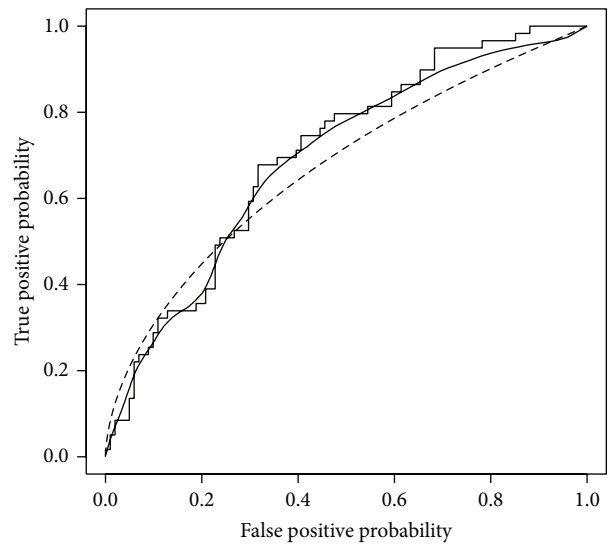


Figure 1: Empirical ROC curve for postmenopausal women and estimates of the ROC curve from a parametric normal model (dashed line) and a mixture of finite Polya trees analysis (smooth solid line).

The LPML statistics were similar for models 1 and 2 (LPML $\doteq -1187$). We therefore selected the more flexible model 2 that had separate distributions across $\widetilde{x}$. For the standard parametric model 3, LPML $= -1201$, indicating a strong preference for the more flexible models (pseudo-Bayes factor $= e^{14}$).

Estimated ROC curves were obtained from model 2, together with estimates from the parametric normal model 3 and from nonparametric empirical distribution functions. The estimated ROC curve from the parametric analysis differs drastically from estimates obtained from the MFPT and empirical distribution functions for postmenopausal women (Figure 1). Not surprisingly, the parametric model fails to track either of the data-driven estimates. Note that although the parametric model is its prior expectation, the MFPT model has the flexibility to allow the data to reallocate probability mass away from a bell-curve to produce an appropriately smooth and flexible estimate of the ROC curve. For premenopausal women, the parametric and Polya tree models also give different inferences, including at clinically important cutoffs corresponding to false positive probabilities up to 10% (Figure 2). The empirical partial AUC over that region was 0.026, while the estimate from the parametric analysis was increased by 37% to 0.041. A 90% posterior interval for the partial AUC from the parametric model was (0.027, 0.055), which does not contain the empirical estimate. The Polya tree analysis is a middle ground between these extremes, with an estimated ROC curve that is essentially a smoothed version of the empirical curve and an estimated partial AUC of 0.019 (0.009, 0.036). Similar results were obtained from a sensitivity analysis that placed diffuse Gamma priors with mean 1 and variance 10000 on the $\mu_\ell$'s and $\sigma_\ell$'s.

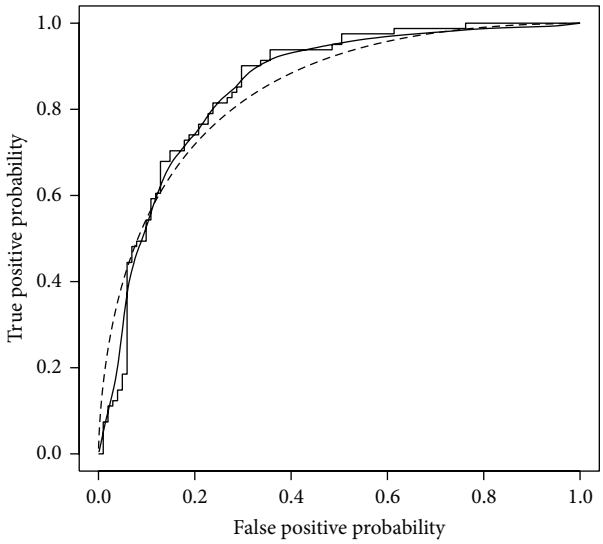


Figure 2: Empirical ROC curve for premenopausal women and estimates of the ROC curve from a parametric normal model (dashed line) and a mixture of finite Polya trees analysis (smooth solid line).

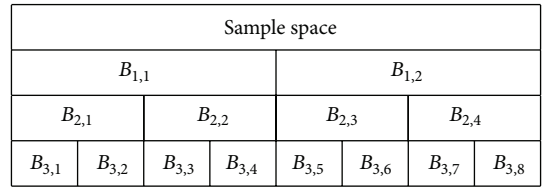


Figure 3: Partition structure of a Polya tree with three levels.

## 4. Conclusions

We have described the use of MFPT priors in Bayesian analysis. Even when a parametric model is thought to hold, a Polya tree analysis offers a way to assess parametric assumptions and to perform a sensitivity analysis to address deviations from them. For example, a simple approach to sensitivity analysis involves comparing estimated ROC curves and AUC from the normal-normal and semiparametric models.

A finite Polya tree prior is not technically nonparametric. Bayesian nonparametric models are characterized by having an infinite number of parameters. Such models often use finite approximations, as in our case. It is thus more accurate to view models that contain finite Polya tree priors as parametric, with a large number of parameters. In fact, Bayesian statistical analysis with finite Polya tree priors has been called "parametric nonparametric statistics," because it uses parametric models (with many parameters) that maintain flexibility [25]. In our experience, finite approximations of Polya trees are sufficiently flexible and scalable to perform well for a wide variety of complex settings.

## Appendices

## A. Polya Tree Priors

*A.1. Finite Polya Tree Priors.* A random sample $y_1, \ldots, y_n$ is obtained from an unknown distribution $G$. We describe informally what it means to place a finite Polya tree prior on $G$. Polya tree priors place a distribution on a collection of cumulative distribution functions. The first step in constructing a finite Polya tree prior is to create nested partitions of the sample space. The first partition splits the sample space into two nonoverlapping intervals. In the second partition, those two intervals are each split, yielding a finer partition that contains four intervals. Then, these four intervals are each split to give an 8 interval third-level partition of the sample space. This sequential process of splitting the sample space into finer and finer partitions continues up to a certain (finite) level $J$.

An analogy is to a convention center with $J$ floors. The rooftop encompasses the entire sample space. Down one level is a floor with two large meeting rooms. Two levels down is a floor with four meeting rooms, then a floor with 8 rooms, and so on until we get to the ground floor $J$ levels down, which contains $2^J$ meeting rooms (Figure 3 contains an example with $J = 3$). Suppose the unknown distribution $G$ assigns multiple people (the $y_i$'s in this analogy) to rooms on the ground floor, and our task is to use the observed distribution of people to rooms to estimate $G$. Although all levels (floors) of the tree (convention center) are important for the purpose of estimating $G$, of primary importance is level $J$ (the ground floor).

The goal is to produce a data-driven estimate of $G$ that assigns high probability to intervals (rooms) that contain lots of data (people), assigns low probability to empty intervals, and assigns midrange probability to intervals that contain some (but not a lot) of the data. Consider first the simplest case where the convention center has only one floor ($J = 1$). Then, people are assigned to one of two meeting rooms on floor 1. Call them rooms $B_{1,1}$ and $B_{1,2}$. The probability of being assigned to the first room on floor 1 is $G(B_{1,1}) = \Pr(y_i \in B_{1,1})$, which, since $G$ is a cumulative distribution function and $B_{1,1}$ is an interval of the form $(L, U)$, is to be interpreted as $G(B_{1,1}) = G(U) - G(L)$. Denote this unknown probability by the parameter $\pi_{1,1}$. Then, by the complement rule, $\pi_{1,2} = 1 - \pi_{1,1}$ is the probability of being assigned to room 2 on floor 1. In notation, $\Pr(y_i \in B_{1,1}) = G(B_{1,1}) = \pi_{1,1}$ and $\Pr(y_i \in B_{1,2}) = G(B_{1,2}) = \pi_{1,2}$. Since $G$ is unknown, $\pi_{1,1}$ and $\pi_{1,2}$ are unknown. To help interpret these parameters, suppose the data arise from a right-skewed distribution (lots more people in room 1 than room 2), then $\pi_{1,1}$ would be large and hence $\pi_{1,2}$ would be small, and vice-versa for left-skewed data. In practice, $J = 1$ is never used because it would lead to a crude estimate of the density function $g$, much like estimating a density using a relative frequency histogram that contains only two bins. In our experience, setting $J$ equal to 4, 5, or 6 often works well in practice. Another option is to select $J$ so that roughly $2^J \doteq n$ [18].

Now suppose we return a year later to find that the convention center has expanded and contains a second floor ($J = 2$). Suppose the new ground floor has four rooms,

$B_{2,1}$, $B_{2,2}$, $B_{2,3}$, and $B_{2,4}$ (third row of Figure 3). Moreover, suppose that room assignments are made based on whether the individual was in room $B_{1,1}$ or $B_{1,2}$ during the previous visit. If the previous assignment was to room $B_{1,1}$, then the individual is assigned again to the left side of the convention center, namely, to either room $B_{2,1}$ with unknown probability $\pi_{2,1}$ or to room $B_{2,2}$ with probability $\pi_{2,2} = 1 - \pi_{2,1}$. Similarly, define $\pi_{2,3}$ and $\pi_{2,4}$ $(= 1 - \pi_{2,3})$ to be the probability of being assigned to room $B_{2,3}$ or $B_{2,4}$, respectively, given that the previous assignment was to room $B_{1,2}$.

The $\pi_{j,k}$'s are *conditional* probability parameters, since $\pi_{2,1} = \Pr(y_i \in B_{2,1} \mid y_i \in B_{1,1})$ and $\pi_{2,3} = \Pr(y_i \in B_{2,3} \mid y_i \in B_{1,2})$. To relate the $\pi_{j,k}$'s to $G$, we must determine the *marginal* probability of assignment to the various ground-floor rooms. Observe that interval $B_{2,1}$ is a subset of interval $B_{1,1}$ (see Figure 3), so the marginal probability of interval $B_{2,1}$ is

$$\begin{aligned} G(B_{2,1}) &= \Pr(y_i \in B_{2,1}) \\ &= \Pr(y_i \in B_{2,1}, y_i \in B_{1,1}) \\ &= \Pr(y_i \in B_{2,1} \mid y_i \in B_{1,1}) \Pr(y_i \in B_{1,1}) \\ &= \pi_{2,1}\pi_{1,1}. \end{aligned} \tag{A.1}$$

Similar steps lead to $G(B_{2,2}) = (1 - \pi_{2,1})\pi_{1,1}$, $G(B_{2,3}) = \pi_{2,3}\pi_{1,2}$, and $G(B_{2,4}) = (1 - \pi_{2,3})\pi_{1,2}$. Suppose again that $G$ is right skewed. Then the data will estimate $\pi_{1,1}$ to be large, and it will estimate $\pi_{2,1}$ to be large since most of the $n$ people will be assigned to room $B_{2,1}$. Therefore, the estimate of $G(B_{2,1})$ will be (relatively) large.

Notice that level 1 has only one unique parameter $(\pi_{1,1})$ associated with it because $\pi_{1,2}$ is completely determined by $\pi_{1,1}$; similarly, level 2 has two unique parameters ($\pi_{2,1}$ and $\pi_{2,3}$) associated with it. We can continue the convention center storyline to any level $J$. For $J = 3$ (Figure 3), we add conditional probability parameters $\pi_{3,1}$, $\pi_{3,2}$, $\pi_{3,3}$, $\pi_{3,4}$, $\pi_{3,5}$, $\pi_{3,6}$, $\pi_{3,7}$, and $\pi_{3,8}$, but only 4 of these are unique. For instance, $\pi_{3,1}$ is the probability that $y_i$ is in interval $B_{3,1}$ given that it is in interval $B_{2,1}$, and $G(B_{3,1}) = \Pr(y_i \in B_{3,1}) = \Pr(y_i \in B_{3,1}, y_i \in B_{2,1}) = \Pr(y_i \in B_{3,1} \mid y_i \in B_{2,1}) \Pr(y_i \in B_{2,1}) = \pi_{3,1}\pi_{2,1}\pi_{1,1}$.

The key point is that if we can estimate all of the $\pi_{j,k}$'s, then we can estimate the probability that is allocated by $G$ to each interval at level $J$. Estimating continuous $G$ requires continuing ad infinitum as $J$ grows. However, in practice we truncate to a fixed $J$, hence the term, finite Polya tree. But this is incomplete for the purpose of estimating a continuous $G$, because we have not modeled how probability mass is distributed *within* each interval at level $J$. In terms of the convention center analogy, we have modeled the probability of assignment to each of the ground-floor rooms, but we have not modeled how people are distributed within those rooms. For instance, they could all be clustered in the middle of the room (all the $y_i$'s clumped together in the center of the interval). Alternatively, the data could be uniformly distributed, clumped to the right or left side of the interval, or have any other dispersion pattern within each interval. To address this issue, we model the data according to how a user-specified distribution $G^*$ allocates probability mass within the intervals at level $J$.

The distribution $G^*$ is important in two other ways. First, it is used to determine the lower and upper endpoints of all intervals in the tree. The median of $G^*$ is used to split the sample space into two intervals at level 1 of the tree. The quartiles of $G^*$ define cutpoints for intervals at level 2. Writing the 25th percentile as $G^{*-1}(1/4)$, the median as $G^{*-1}(2/4)$, and the 75th percentile as $G^{*-1}(3/4)$, we have for a sample space that covers the real line

$$\begin{aligned} B_{2,1} &= \left(-\infty, G^{*-1}\left(\frac{1}{4}\right)\right), \\ B_{2,2} &= \left(G^{*-1}\left(\frac{1}{4}\right), G^{*-1}\left(\frac{2}{4}\right)\right), \\ B_{2,3} &= \left(G^{*-1}\left(\frac{2}{4}\right), G^{*-1}\left(\frac{3}{4}\right)\right), \\ B_{2,4} &= \left(G^{*-1}\left(\frac{3}{4}\right), +\infty\right). \end{aligned} \tag{A.2}$$

In general, the $(j,k)$th interval is $B_{j,k} = (G^{*-1}((k-1)/2^j), G^{*-1}(k/2^j))$, for $j = 1, \ldots, J$ and $k = 1, \ldots, 2^j$.

Second, $G^*$ is the prior expectation of the unknown distribution function $G$. This provides guidance for selecting $G^*$; we select it based on our best prior assessment of the data-generating distribution $G$. If our prior assessment is that the data will be governed by a certain normal distribution, we use that normal for $G^*$. If our prior assessment is accurate, the posterior estimate of $G$ from the finite Polya tree model will take the shape of that normal distribution. But the real advantage of Polya tree priors is their flexibility to allow for data-driven prior-to-posterior reallocation of probability mass away from the shape of $G^*$ to any shape supported by the data. This happens by estimating the $\pi_{j,k}$'s that were equated to $G$.

The collection $\{\pi_{j,k} : j = 1, \ldots, J;\ k = 1, \ldots, 2^j\}$ constitutes the unknown parameters corresponding to $G$. We need to specify a prior distribution over this collection. Recall that when $k$ is an even number between 2 and $2^j$, $\pi_{j,k} = 1 - \pi_{j,k-1}$. Therefore, priors are needed only on $\pi_{j,k}$ when $k$ is odd. It is standard to use independent beta priors, specifically $\pi_{j,k} \sim \text{Beta}(cj^2, cj^2)$ for $k$ odd and all $j$. The prior mean of 0.5 gives equal weight to the left and right intervals formed by splitting a parent interval. The positive constant $c$ is often set equal to 0.5 or 1. The value of $c$ reflects our confidence in the prior estimate $G^*$ for the unknown $G$. Large values of $c$ (e.g., $c \geq 10$) translate into higher prior confidence in the particular $G^*$, so a relatively larger amount of data are needed in order for the posterior estimate to stray from $G^*$ if the data conflict with it. A low value (e.g., $c = 0.1$) will often lead to an estimate of $G$ that is similar to the empirical distribution function. We have found that $c = 0.5$ or 1 performs well in practice because it often is a middle ground between a purely data-based and model-based estimate.

```
proc mcmc data=MFPT nbi=5000 nmc=50000 thin=10 monitor=(G littleg mu sigma);
begincnst; J=4; c=1; endcnst;
parms pi11 pi21 pi23;
parms pi31 pi33 pi35 pi37;
parms pi41 pi43 pi45 pi47 pi49 pi411 pi413 pi415;
parms mu sigma / t(3);
array p[16];
p1 = pi11*pi21*pi31*pi41; p2 = pi11*pi21*pi31*(1-pi41);
p3 = pi11*pi21*(1-pi31)*pi43; p4 = pi11*pi21*(1-pi31)*(1-pi43);
p5 = pi11*(1-pi21)*pi33*pi45; p6 = pi11*(1-pi21)*pi33*(1-pi45);
p7 = pi11*(1-pi21)*(1-pi33)*pi47; p8 = pi11*(1-pi21)*(1-pi33)*(1-pi47);
p9 = (1-pi11)*pi23*pi35*pi49; p10 = (1-pi11)*pi23*pi35*(1-pi49);
p11 = (1-pi11)*pi23*(1-pi35)*pi411; p12 = (1-pi11)*pi23*(1-pi35)*(1-pi411);
p13 = (1-pi11)*(1-pi23)*pi37*pi413; p14 = (1-pi11)*(1-pi23)*pi37*(1-pi413);
p15 = (1-pi11)*(1-pi23)*(1-pi37)*pi415; p16 = (1-pi11)*(1-pi23)*(1-pi37)*(1-pi415);
prior pi11 ~ beta(c,c); prior pi21 pi23 ~ beta(c*2**2,c*2**2);
prior pi31 pi33 pi35 pi37 ~ beta(c*3**2,c*3**2);
prior pi41 pi43 pi45 pi47 pi49 pi411 pi413 pi415 ~ beta(c*4**2,c*4**2);
prior mu ~ normal(50,sd=10); prior sigma ~ uniform(0,20);
k = int(2**J * cdf("normal", y, mu, sigma) + 1);
llike=J*log(2) + log(p[k]) + lpdfnorm(y, mu, sigma);
model general(llike);
array setID[52]; array G[52]; array littleg[52];
do i = 1 to dim(G) by 1;
setID[i] = int(2**J * cdf("normal", 39 + i/2, mu, sigma) + 1);
G[i] = p1*(setID[i] ge 2) + p2*(setID[i] ge 3) + p3*(setID[i] ge 4)
       + p4*(setID[i] ge 5) + p5*(setID[i] ge 6) + p6*(setID[i] ge 7)
       + p7*(setID[i] ge 8) + p8*(setID[i] ge 9) + p9*(setID[i] ge 10)
       + p10*(setID[i] ge 11) + p11*(setID[i] ge 12) + p12*(setID[i] ge 13)
       + p13*(setID[i] ge 14) + p14*(setID[i] ge 15) + p15*(setID[i] ge 16)
       + p[setIDi] *(2**J * cdf("normal", 39 + i/2, mu, sigma)- setID[i] + 1);
littleg[i] = 2**J * p[setIDi] * pdf("normal", 39 + i/2, mu, sigma);
end; run;
```

Algorithm 1

Once we have selected $J$, $G^*$, and $c$, we have all the elements needed to specify a finite Polya tree prior for $G$. Formulas for $G$ and its corresponding density function $g$ are

$$\begin{aligned} G(y) &= \sum_{i=1}^{k(y)-1} p\left(B_{J,i}\right) + p\left(B_{J,k(y)}\right) \\ &\quad \times \left[2^J G^*(y) - k(y) + 1\right], \\ g(y) &= 2^J p\left(B_{J,k(y)}\right) g^*(y), \end{aligned} \tag{A.3}$$

where $k(y)$ is an integer between 1 and $2^J$ that identifies the interval at level $J$ containing $y$, and $p(B_{J,k(y)})$ is the probability of that interval (it is the product of $J$ of the $\pi_{j,k}$'s) [19]. The interval that contains $y$ at level $J$ can be determined using the formula $k(y) = \text{Int}(2^J G^*(y)+1)$, where the function Int$(\cdot)$ returns the integer portion of a decimal number (e.g., Int$(3.14) = 3$).

To motivate an extension to *mixtures* of finite Polya tree priors, consider Figure 4(a), where $g$ is estimated using a finite Polya tree analysis of 200 simulated data values from a mixture of two normal distributions (70% of observations from $N(70, 4)$ and 30% from $N(57, 2.25)$). The prior mean, $G^*$, was the $N(50, 25)$, and we set $J = 4$. The estimate from the finite Polya tree analysis is choppy, which turns out to be the result of using a fixed $G^*$. These bumps get smoothed over if instead we replace $G^*$ with a parametric distribution $G^*_\theta$, where $\theta$ denotes a collection of unknown parameters that is assigned a prior distribution. The posterior estimate of $g$ in Figure 4(b) was obtained using a $N(\mu, \sigma^2)$ for $G^*_\theta$, with $\mu \sim N(50, 100)$ and $\sigma \sim U(0, 20)$. This is an example of our preferred prior in nonparametric Bayesian statistics, namely, the mixture of finite Polya trees prior. The model is $y_i \mid G \sim G, G \sim \text{MFPT}(G^*_\theta, p, c)$, where $p$ is a prior distribution on $\theta$. With large $c$, the model for $y_i$ is simply $N(\mu, \sigma^2)$ and the analysis becomes completely parametric. For a more elaborate but still elementary description, see [25].

## B. SAS Code

SAS 9.3 code to fit the one-sample MFPT model to simulated data from the mixture of two normal distributions described in Appendix A is presented in Algorithm 1. The SAS data step is not shown in the code; in it, the variable y is read into a data set that was named MFPT, where y contains 200 evenly spaced percentiles from the true mixture distribution. The

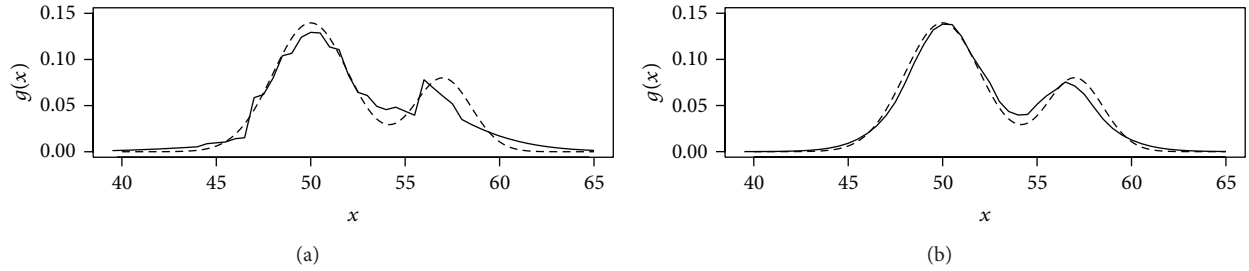


Figure 4: Posterior estimates of $g$ from simulated bimodal data using a finite Polya tree prior ((a), solid line) and a mixture of finite Polya trees prior ((b), solid line). The dashed lines represent the true data density $g$.

first line instructs the mcmc procedure to use the data in `MFPT` and specifies options for the number of burn-in iterates (`nbi`), the number of iterates to simulate post burn-in (`nmc`), to retain every 10th iterate (`thin=10`), and the parameters to monitor for posterior inference. Constants that are used in the program appear in line 2; in this example we set $J = 4$ and $c = 1$. Each parameter $\pi_{j,k}$ is coded as `pijk`, with the appropriate numbers substituted for `j` and `k`. The marginal probabilities at level $J = 4$ are named `pk`, where `k` ranges from 1 to 16; these variables represent $p(B_{J,k(y)})$ as defined in Appendix A. Priors on the $\pi_{j,k}$'s, $\mu$, and $\sigma$ follow. SAS code to fit the finite Polya tree model described in Appendix A can be obtained by omitting the `prior` and `parms` lines associated with $\mu$ and $\sigma$ and setting `mu` and `sigma` equal to constants in line 2 of the code. The variable `k` identifies the set each $y_i$ belongs to at level 4. The log likelihood (`llike`) is obtained by taking the natural log of the formula for $g$ that appears in (A.3). The remaining code is used to estimate the distribution function (`G`) and the density function $g$ (coded as `littleg` since SAS is not case sensitive).

The model is fit using Gibbs sampling with block updating occurring for the groups of parameters that are defined in the `parms` statements. The posterior distributions of $g(y)$ and $G(y)$ are evaluated at $y = 39.5, 40, 40.5, \ldots, 65$. Figure 4(b) plots the posterior means of $g(39.5)$, $g(40)$, $g(40.5), \ldots, g(65)$, and a continuous density is obtained by interpolation.

## Disclosure

Andre Baron is a coinventor of patents related to sEGFR and cofounder of Tumor Biology Investment Group, Inc., a biotechnology company that holds the rights to several sEGFR patents. Of note, Dr. Baron was not involved in the statistical analyses or interpretation of the statistical results of these analyses.

## Acknowledgments

This work was supported by National Institutes of Health (Grants K07 CA76170, R21 CA82520, and RO3 CA82091 to A.T.B). The authors thank both referees for their helpful and encouraging comments.

## References

[1] D. B. Dunson, "Commentary: practical advantages of Bayesian analysis of epidemiologic data," *American Journal of Epidemiology*, vol. 153, no. 12, pp. 1222–1226, 2001.

[2] S. Greenland, "Bayesian perspectives for epidemiological research: I. Foundations and basic methods," *International Journal of Epidemiology*, vol. 35, no. 3, pp. 765–775, 2006.

[3] S. Greenland, "Bayesian perspectives for epidemiological research. II. Regression analysis," *International Journal of Epidemiology*, vol. 36, no. 1, pp. 195–202, 2007.

[4] S. Greenland, "Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods," *International Journal of Epidemiology*, vol. 38, no. 6, pp. 1662–1673, 2009.

[5] A. Lawson, *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, CRC Press, Boca Raton, Fla, USA, 2009.

[6] R. F. MacLehose, J. M. Oakes, and B. P. Carlin, "Turning the Bayesian crank," *Epidemiology*, vol. 22, no. 3, pp. 365–367, 2011.

[7] S. R. Cole, H. Chu, S. Greenland, G. Hamra, and D. B. Richardson, "Bayesian posterior distributions without Markov chains," *American Journal of Epidemiology*, vol. 175, no. 5, pp. 368–375, 2012.

[8] A. Erkanli, M. Sung, E. J. Costello, and A. Angold, "Bayesian semi-parametric ROC analysis," *Statistics in Medicine*, vol. 25, no. 22, pp. 3905–3928, 2006.

[9] T. E. Hanson, A. Kottas, and A. J. Branscum, "Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches," *Journal of the Royal Statistical Society C*, vol. 57, no. 2, pp. 207–225, 2008.

[10] A. J. Branscum, W. O. Johnson, T. E. Hanson, and I. A. Gardner, "Bayesian semiparametric ROC curve estimation and disease diagnosis," *Statistics in Medicine*, vol. 27, no. 13, pp. 2474–2496, 2008.

[11] T. E. Hanson, A. J. Branscum, and I. A. Gardner, "Multivariate mixtures of Polya trees for modeling ROC data," *Statistical Modelling*, vol. 8, no. 1, pp. 81–96, 2008.

[12] J. Gu, S. Ghosal, and A. Roy, "Bayesian bootstrap estimation of ROC curve," *Statistics in Medicine*, vol. 27, no. 26, pp. 5407–5420, 2008.

[13] C. Wang, B. W. Turnbull, Y. T. Gröhn, and S. S. Nielsen, "Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 12, no. 1, pp. 128–146, 2007.

[14] G. T. Fosgate, H. M. Scott, and E. R. Jordan, "Development of a method for Bayesian nonparametric ROC analysis with

application to an ELISA for Johne's disease in dairy cattle," *Preventive Veterinary Medicine*, vol. 81, no. 1–3, pp. 178–193, 2007.

[15] V. Inácio, A. A. Turkman, C. T. Nakas, and T. A. Alonzo, "Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface," *Biometrical Journal*, vol. 53, no. 6, pp. 1011–1024, 2011.

[16] I. V. de Carvalho, A. Jara, T. E. Hanson, and M. de Carvalho, "Bayesian nonparametric ROC regression modeling," *Bayesian Analysis*, vol. 8, no. 3, pp. 623–646, 2013.

[17] M. Ladouceur, E. Rahme, P. Bélisle, A. N. Scott, K. Schwartzman, and L. Joseph, "Modeling continuous diagnostic test data using approximate Dirichlet process distributions," *Statistics in Medicine*, vol. 30, no. 21, pp. 2648–2662, 2011.

[18] T. Hanson and W. O. Johnson, "Modeling regression error with a mixture of Polya trees," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1020–1033, 2002.

[19] T. E. Hanson, "Inference for mixtures of finite Polya tree models," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1548–1565, 2006.

[20] E. J. Bedrick, R. Christensen, and W. Johnson, "A new perspective on priors for generalized linear models," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1450–1460, 1996.

[21] A. T. Baron, J. M. Lafky, C. H. Boardman et al., "Serum sErbB1 and epidermal growth factor levels as tumor biomarkers in women with stage III or IV epithelial ovarian cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 8, no. 2, pp. 129–137, 1999.

[22] A. T. Baron, C. H. Boardman, J. M. Lafky et al., "Soluble epidermal growth factor receptor (SEG-FR) and cancer antigen 125 (CA125) as screening and diagnostic tests for epithelial ovarian cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 14, no. 2, pp. 306–318, 2005.

[23] S. Geisser and W. F. Eddy, "A predictive approach to model selection," *Journal of the American Statistical Association*, vol. 74, pp. 153–160, 1979.

[24] R. Christensen, W. Johnson, A. Branscum, and T. Hanson, *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press, Boca Raton, Fla, USA, 2010.

[25] R. Christensen, T. Hanson, and A. Jara, "Parametric nonparametric statistics: an introduction to mixtures of finite Polya trees," *The American Statistician*, vol. 62, no. 4, pp. 296–306, 2008.