Les Cahiers du GERAD ISSN: 0711–2440

Bayesian estimation of disease prevalence from continuous diagnostic test data using Polya tree distributions

M. Kaouache, B. MacGibbon,
L. Joseph

G–2015–125

November 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Bayesian estimation of disease prevalence from continuous diagnostic test data using Polya tree distributions

Mohammed Kaouache [a]
Brenda MacGibbon [b]
Lawrence Joseph [c]

[a] *McGill University Health Centre, The Maude Unit, Royal Victoria Hospital, Montréal (Québec) Canada, H3A 1A1*

[b] *GERAD & Département de mathématiques, Université du Québec à Montréal, Montréal (Québec) Canada, H3C 3P8*

[c] *McGill University Health Centre, Division of Clinical Epidemiology, Royal Victoria Hospital, Montréal (Québec) Canada, H3A 1A1, and McGill University, Department of Epidemiology and Biostatistics, Montréal (Québec) Canada, H3A 1A2*

mohammed.kaouache@mail.mcgill.ca
macgibbon.brenda@gmail.com
lawrence.joseph@mcgill.ca

November 2015

Les Cahiers du GERAD
G–2015–125

Abstract: Inferences about the prevalence of a given disease or condition can be drawn from results of a diagnostic test applied to a sample from the target population. For example, knowledge about disease clustering in tuberculosis (TB) can be estimated from the nearest genetic distance (NGD), a continuous test measuring the relatedness of TB strains. Most diagnostic tests, including the NGD test for TB clustering, are imperfect, which for continuous tests implies overlap in the measure between positive and negative cases. The resulting misclassification errors must be taken into account when estimating the prevalence. In creating models for continuous test results, one can use either a standard parametric form, such as normally distributed data through a bi-normal model, or attempt to fit a nonparametric model that makes fewer distributional assumptions. Nonparametric models include those based on Dirichlet Process priors and Polya trees. While Polya tree models have been applied to continuous diagnostic testing data, their properties in this context concerning prevalence estimation have not been rigorously examined. We extend this past work in three directions. First, we use simulations to learn about the performance of the model in practice. Second, we derive a method to calculate the Bayes Factor to select between a parametric and a nonparametric model. Third, we investigate the dependence of a fixed partition Polya tree model on the particular partition selected by comparison of the results with a random partition Polya tree model. Finally, we apply our methods to estimate the prevalence of TB clustering from NGD data.

Key Words: Bayesian nonparametric methods, bi-normal model, continuous diagnostic test, misclassification, Polya tree, prevalence.

Résumé : Les inférences sur la prédominance d'une maladie ou d'une condition donnée sont déduites parfois des résultats de tests de diagnostic continus. Par exemple, de l'information sur la présence de grappes en tuberculose (TB) peut être estimée à partir de la distance génétique la plus proche (NGD), un test mesurant la parenté de souches de TB. La plupart des tests de diagnostic sont imparfaits, ce qui se traduit par un chevauchement entre les mesures provenant de sujets dont le vrai statut est positif et celles provenant de sujets dont le vrai statut est négatif. Les erreurs de classification qui en résultent doivent être prises en compte quand on estime la prédominance. En formulant des modèles pour les tests de diagnostic, on peut soit utiliser un modèle paramétrique standard tel qu'un modèle bi-normal, qui suppose que les données suivent une loi normale, soit considérer un modèle non paramétrique, qui offre plus de flexibilité en imposant moins de contraintes sur la forme des lois de probabilité. Parmi ces modèles non paramétriques, on peut nommer ceux qui reposent sur le processus de Dirichlet et ceux qui reposent sur le processus d'arbres de Pólya. Bien que les modèles d'arbres de Pólya aient été utilisés pour modéliser les tests de diagnostic continus, leurs propriétés dans le contexte de l'estimation de la prédominance n'ont pas été rigoureusement étudiées. Nous étendons ces travaux passés dans trois directions. Premièrement, nous utilisons des simulations pour connaître l'efficacité du modèle en pratique. Deuxièmement, nous établissons une méthode pour calculer les facteurs de Bayes pour fin de sélection entre modèle paramétrique et modèle non paramétrique. Troisièmement, nous étudions la dépendance d'un modèle d'arbres de Pólya, où la partition est fixe, sur la partition particulière sélectionnée, en le comparant à un modèle dont la partition est aléatoire. Finalement nous appliquons nos méthodes à l'estimation de la prédominance de formation de grappes en TB en utilisant des données NDG.

1 Introduction

In tuberculosis (TB) epidemiology it is of interest to estimate the prevalence of clustering defined by the proportion of recently transmitted cases as opposed to cases that are reactivations of previously acquired infection. To date there is no gold standard test to definitively determine whether a case is clustered or not, but several imperfect diagnostic measures have been devised. One marker of recent transmission is a continuous measure of genetic relatedness, the nearest genetic distance or NGD [1]. Lower values of NGD are more likely to be clustered compared to higher values. While the probability density function of NGD values from recently transmitted cases is located to the left of the density from the reactivated cases, some overlap is expected. Furthermore, NGD data tend to be skewed, so normality assumptions may not hold. Given the possibility of non-normality of NGD data, what statistical model should be applied to estimate the prevalence of TB clustering, and how should this model be selected? Once a model is chosen, given the imperfect nature of the NGD test, how well can it be expected to perform in estimating the prevalence of clustering?

Analysis of diagnostic test data in the absence of a gold standard is usually done by latent class models. Frequentist latent class models are reviewed by Goetghebeur et al. [2] and Rutjes et al. [3] , although if data from only a single diagnostic test is available the problem is non-identifiable, and Bayesian methods may be preferable [4,5]. The non-normality can be handled by non-parametric methods, and indeed several Bayesian nonparametric latent class models have been proposed. Ladouceur et al. [6] used approximate Dirichlet process priors [7], extending the bi-normal model of Scott et al. [8]. Dirichlet process models appear to work well in practice, but are problematic theoretically since they select a discrete probability distribution for test results with probability one. Conversely, Polya tree priors result in a continuous probability distribution. Branscum et al. Branscum et al. [9] used a mixture of finite Polya tree priors in the diagnostic testing setting but their model requires additional diagnostic tests for inferences to be made on the prevalence.

Here we propose to study a Bayesian finite Polya tree model for continuous diagnostic test data with no gold standard, focusing on the case where no additional diagnostic test data or other information is available. We extend past work on Bayesian nonparametric latent class models in three directions. First, we derive a method to calculate the Bayes Factor to select between a parametric and a nonparametric model. Second, we use simulations to learn about the performance of the model in practice, comparing inferences from Polya trees to those assuming normality, both when normality holds and when data come from a different distribution. Third, we investigate the dependence of a fixed partition Polya tree model on the particular partition selected by comparison of the results to those from a random partition Polya tree model [10]. This is of interest because some have criticized Polya tree models for the need to select a single partition, upon which the smoothness of the resulting density estimate can depend.

Section 2 presents our Bayesian Polya tree model for continuous diagnostic tests, including the likelihood function and prior and posterior distributions. This section also discusses a model based on randomized Polya trees. Results from our simulation study investigating the properties of our model and comparisons to the bi-normal parametric model are given in Section 3. These simulations include examples of prototypic scenarios under which diagnostic testing data may arise. In Section 4, we use Bayes factors computed via marginal likelihood to select between the parametric and nonparametric models. In Section 5 we return to the analysis of NGD data for TB clustering in Montreal, Canada. Section 6 concludes the paper with a discussion.

2 Bayesian finite Polya tree models for continuous diagnostic test data

We first describe a standard Polya tree model with a fixed partition, followed by the extension to random partition Polya trees.

2.1 A fixed partition Polya tree model

Polya trees were introduced by Freedman [11], Fabius [12] and Ferguson [13], and further developed by Lavine [14,15] and Mauldin et al. [16]. Polya trees are generalizations of Dirichlet processes that can place positive

mass on the set of absolutely continuous distributions, constructed by recursive binary partitioning of the sample space into finer and finer disjoint sets.

A Polya tree distribution, F, on a sample space Ω can be defined as follows. We start by constructing a recursive tree of partitions of Ω. At the top of the tree (level 0) is the entire space Ω, followed by level 1 which consists of B_0 and B_1, with 0 denoting the left interval and 1 the right side. While in general these can be any two disjoint intervals, in this paper we use the canonical construction [13]. Let G be a fixed parametric base distribution, for example a normal distribution, and then choose B_0 and B_1 such that $G(B_0) = G(B_1)$. Continuing to level 2, split B_0 into B_{00} and B_{01} such that $G(B_{00}) = G(B_{01})$, and similarly split B_1 into B_{10} and B_{11}, such that $G(B_{10}) = G(B_{11})$. The disjoint intervals B_{00}, B_{01}, B_{10}, B_{11}, form the partition of Ω at level 2, and so on to deeper levels of the tree.

Any interval at an arbitrary level k can be uniquely identified from a sequence of 0's and 1's of length k, defining the unique path from the top of the tree to level k. Let $\varepsilon = \varepsilon_1 ... \varepsilon_k$ be the binary sequence that identifies the interval $B_{\varepsilon_1 ... \varepsilon_k}$ at a certain level k of the tree. We define the conditional probability, also called the branch probability, $Y_{\varepsilon_1 ... \varepsilon_k} = F(B_{\varepsilon_1 ... \varepsilon_k 0} | B_{\varepsilon_1 ... \varepsilon_k})$ that, at level $k+1$, we end-up in $B_{\varepsilon_1 ... \varepsilon_k 0}$ (1-$Y_{\varepsilon_1 ... \varepsilon_k}$ is then the probability of ending up in $B_{\varepsilon_1 ... \varepsilon_k 1}$), given that at level k we were in $B_{\varepsilon_1 ... \varepsilon_k}$.

We complete the definition of the Poly tree by assuming that the branch probabilities are independent and follow beta distributions with fixed parameters. At level $k = 0$, let $Y_\emptyset = F(B_0)$ so that $F(B_1) = 1 - Y_\emptyset$, and assume $Y_\emptyset \sim Beta(\alpha_0, \alpha_1)$. For $k \geq 1$, and conditional on $B_{\varepsilon_1 ... \varepsilon_k}$, the mass $Y_{\varepsilon_1 ... \varepsilon_k}$ given by F to $B_{\varepsilon_1 ... \varepsilon_k 0}$ follows a $Beta(\alpha_{\varepsilon_1 ... \varepsilon_k 0}, \alpha_{\varepsilon_1 ... \varepsilon_k 1})$ density. Let $\alpha_0 = \alpha_1 = c$ and for $k \geq 1$, we let $\alpha_{\varepsilon_1 ... \varepsilon_k 0} = \alpha_{\varepsilon_1 ... \varepsilon_k 1} = c(k+1)^2$, where c is a weight parameter, which implies an absolutely continuous distribution with probability 1 [13]. F is then centered about the base distribution G, and the weight parameter c represents the strength in our belief that the data are in fact generated from the base distribution G, with higher values of c indicating stronger belief.

In theory, Polya trees have an infinite number of parameters, approximated in practice by finite Polya trees which are specified only up to a certain level, denoted here by K. We use the notation $F \sim PT_K(G, c)$ to denote that F is distributed according to a finite Polya tree prior with base distribution G and a weight parameter equal to c.

We now embed Polya trees within our diagnostic testing model. Let π be the prevalence of a disease or medical condition in a certain population. Suppose that we apply a continuous diagnostic test to a random sample of n individuals from this population, and let $X = (x_1, ..., x_n)$ be the observed test outcomes. A subset of these n individuals belongs to the diseased population (the $D+$ group) and the rest are disease free (the $D-$ group), although the true disease status for each individual is unknown. Let $Z = (z_1, ..., z_n)$ be latent data representing the unknown disease status for each individual, so that $z_i = 1$ if the individual has the disease and $z_i = 0$ otherwise. Let F_0 (F_1) be the cumulative distribution function of test results for the $D-$ ($D+$) group. The model is bi-normal when F_0 and F_1 are assumed to be normally distributed, and a nonparametric model arises when F_0 and F_1 are distributed according to Polya trees, the latter denoted as $x_i|_{z_i} \sim F_{z_i}$ and $F_j \sim PT_K(G_j, c)$, j=0,1. Let $\mathcal{Y}^0$ and $\mathcal{Y}^1$ denote the branch probabilities associated with F_0 and F_1, respectively. Further, let $G_0 \sim N(\mu_0, \sigma_0^2)$ and $G_1 \sim N(\mu_1, \sigma_1^2)$ be the base densities, where $\mu_1, \sigma_1^2, \mu_0, \sigma_0^2$ are selected according to prior information about F_0 and F_1.

Prior densities are also required for the branch probabilities, which are given beta distributions with parameters determined by a weight c. For simplicity we assume the same value of c for both F_0 and F_1. The set of prior distributions is completed by setting a beta (a,b) prior over the range [0, 1] for the prevalence π. In this paper we select $a = b = 1$, the low information because it is the main parameter of interest to estimate. Of course, other choices of prior densities can be used as appropriate in any given problem.

Conditional on being diseased (disease free), the contribution of a data point x_i to the likelihood is equal to the re-normalized density function of the base distribution G_1 (G_0) restricted to the element of the partition, $B^1_{\varepsilon_{i_1} ... \varepsilon_{i_K}}(B^0_{\varepsilon_{i_0} ... \varepsilon_{i_K}})$ at the lowest level K, associated with F_1 (F_0) and containing x_i, multiplied by all the branch probabilities leading to $B^1_{\varepsilon_{i_1} ... \varepsilon_{i_K}}(B^0_{\varepsilon_{i_0} ... \varepsilon_{i_K}})$. The full likelihood function of the observed and latent

data can be written as:

$$f(x_1,...,x_n,z_1,...,z_n|\pi,\mathcal{Y}^1,\mathcal{Y}^0)=\prod_{i=1}^{n}(\pi f_1(x_i))^{z_i}\prod_{i=1}^{n}((1-\pi)f_0(x_i))^{1-z_i}, \quad (1)$$

where

$$f_j(x_i)=\frac{g_j(x_i)}{G_j(B^j_{\varepsilon_{i_1}...\varepsilon_{i_K}})}\prod_{k=1}^{K}(Y^j_{\varepsilon_{i_1}...\varepsilon_{i_{k-1}}})^{1-\varepsilon_{i_k}}(1-Y^j_{\varepsilon_{i_1}...\varepsilon_{i_{k-1}}})^{\varepsilon_{i_k}},\ j=0,1. \quad (2)$$

The parameters $\varepsilon_{i_{1:k}}=\varepsilon_{i_1}...\varepsilon_{i_k}$, $\varepsilon_{i_j}=0,1$ for $j=1,...,k$ trace the unique path followed by the data point x_i down the tree to level k, and g_0 and g_1 are, respectively, the probability densities of the normal base distributions G_0 and G_1. There are two different paths associated with one single data point, one with respect to the $D+$ partition and the other with respect to the $D-$ partition. To avoid cumbersome notation, we use the same $\varepsilon_{i_{1:k}}$ to denote the two different paths, easily distinguishable from their context.

The joint posterior distribution over all parameters is obtained from Bayes Theorem by multiplying the likelihood (1) by the prior distributions. While there is no closed form formula for this posterior density, the Gibbs sampler can be used. This requires the full conditional density for each parameter, which we now describe.

Conditional on the branch probabilities $(\mathcal{Y}^0,\mathcal{Y}^1)$ and π, the true latent disease status z_i is, up to a normalizing constant, a posteriori equal to 1 with probability $\pi f_1(x_i)$ and equal to 0 with probability $(1-\pi)f_0(x_i)$, where f_0 and f_1 are the Polya tree densities given by (2). Thus z_i follows a Bernoulli with parameter $p=\frac{\pi f_1(x_i)}{\pi f_1(x_i)+(1-\pi)f_0(x_i)}$. Conditional on π and the z_i's, the full conditional densities of the branch probabilities are also easily derived by noting that every time a data point with $z_i=0$ falls in a certain interval of the partition at level k indexed by the binary sequence (of length k) $\varepsilon_1...\varepsilon_{k-1}0$, it makes a contribution of size $Y^0_{\varepsilon_1...\varepsilon_{k-1}}$ to the likelihood. Conversely, every time that data point falls in the interval indexed by the binary sequence $\varepsilon_1...\varepsilon_{k-1}1$ it makes a contribution to the likelihood of size $1-Y^0_{\varepsilon_1...\varepsilon_{k-1}}$. Since the prior distribution of $Y^0_{\varepsilon_{i_1}...\varepsilon_{i_{k-1}}}$ is $Beta(ck^2,ck^2)$, its full conditional posterior density is $Beta(ck^2+n_{\varepsilon_1...\varepsilon_{k-1}0},ck^2+n_{\varepsilon_1...\varepsilon_{k-1}1})$, where $n_{\varepsilon_1...\varepsilon_{k-1}0}$ and $n_{\varepsilon_1...\varepsilon_{k-1}0}$ are the number of data points with $z_i=0$ falling in the intervals indexed respectively by $\varepsilon_1...\varepsilon_{k-1}0$ and $\varepsilon_1...\varepsilon_{k-1}1$. The full conditional density of $Y^1_{\varepsilon_1...\varepsilon_{k-1}}$ is similarly obtained. The full conditional density of π can also easily be obtained as a $Beta(a+n_1,b+n_0)$, where n_0 and n_1 are the number of data points with $z_i=0$ and $z_i=1$ respectively.

As usual, the Gibbs sampler selects random values from the above full conditional densities in cyclic fashion, and the output used to approximate the marginal posterior densities across all parameters.

2.2 Randomized Polya tree models

One of the main drawbacks to Polya trees is the dependence of final inferences on the chosen partition. Branscum et al. [9] proposed mixtures of Polya trees to avoid this problem, which employ prior densities rather than fixed values for the means and standard deviations of the base distributions. While this can work well when data from multiple tests are available, poorer inferences can result when data from only a single test is applied. This is because allowing the base distributions to be moved around may cause the estimated $D+$ and $D-$ distributions to wander too far from their respective true distributions, especially when the true densities have appreciable overlap. By fixing the base distributions of $D+$ and $D+$, stronger prior information is put on the location of the true distributions within the diseased and the non-diseased population, while allowing for their shapes to be determined by the data, but this creates discontinuities at the fixed partition endpoints.

To smooth out these discontinuities, Paddock et al. [10] proposed randomized Polya tree models that add randomness to the location of the partition endpoints. Under this random partition model, each data point x_i follows a random distribution F_{0i} if disease free, and a random distribution F_{1i} if diseased. F_{0i} and F_{1i} are conditional on, and can be considered as, random perturbations on the quantile scale of F_0 and

F_1 respectively. Each of F_0 and F_1 are random probabilities distributed according to distinct Polya tree distributions.

Paddock et al. [10] illustrated this strategy using the special case where the sample space is the interval (0,1], but generalization to any space Ω on the real line is straightforward. Let G_0 and G_1 be the base distributions of F_0 and F_1, respectively. At each level j of the tree a new parameter β_{ij} is introduced. These beta parameters will determine the partition endpoints at level j associated with the data point i. The β_{ij} are, a priori, randomly generated in a small interval $(0.5 \pm \tau)$, where τ is a constant generally fixed between 0.01 and 0.1. At level 1, for example, instead of splitting the sample space Ω into B_0^0 and B_1^0 such that $G_0(B_0^0)$=G$_1(B_1^0) = 0.5$, as would be done for a simple Polya tree, we split Ω into B_0^{0i} and B_1^{0i} such that $G_0(B_0^{0i})=\beta_{i1}$ and $G_0(B_1^{0i})$=1-β_{i1}.

A priori, the beta parameters are uniformly distributed over a small interval centered at 0.5. Other than the additional step to sample $\{\beta_{i1}^0, ..., \beta_{iK}^0\}_{i=1}^n$ and $\{\beta_{i1}^1, ..., \beta_{iK}^1\}_{i=1}^n$ from their posterior distributions using a Metropolis-Hastings algorithm [10], the Gibbs sampler can proceed as under a fixed partition model. Note that within each iteration of the MCMC algorithm, the partition endpoints change and in order to determine the beta marginal posterior densities, the number of data points with $z_i = 0$ (with $z_i = 1$) falling within each interval of the $D-$ ($D+$) partition at each level of the tree needs to be recalculated, rendering these random partition models more computationally intensive compared to fixed partition models.

3 Evaluating the performance of latent class Polya tree models

To evaluate the accuracy and precision with which Polya tree models can estimate the prevalence of a given condition from continuous test data, we conducted a simulation study under two main sets of conditions. In the first set of simulations the data are assumed to arise from two normal distributions. The flexibility of a nonparametric model is thus not needed, and the simulations will serve to quantify any disadvantages of using a nonparametric model when the data arise from a standard bi-normal model. In the second set of simulations the test results from both the diseased and the non-diseased populations will be non-normally distributed. Under these conditions the nonparametric model is expected to perform better than the parametric model, and we can thus estimate the degree to which estimation is improved.

3.1 Simulated data sets

For both sets of simulations we chose a relatively low degree of overlap, corresponding to an area under the ROC curve of 0.9. We first experimented with larger overlaps and realized that without a high degree of *a priori* information to distinguish between the diseased and the non-diseased populations the prevalence could not be accurately estimated. Similar findings using Dirichlet priors were reported by Ladouceur et al. [6]. We simulated data with both moderate ($n = 1000$) and large ($n = 5000$) sample sizes. We avoid smaller sample sizes which are generally insufficient for accurate prevalence estimation in the absence of a gold standard [17]. In the first set of simulations, data are simulated from a bi-normal model where the true density within the $D+$ population is $N(0, 1)$ and that within the $D-$ population is $N(2, 1)$. In the second set of simulations, we simulate the $D+$ results from a chi-squared distribution with 3 degrees of freedom and the $D-$ results from an even mixture of $N(5, 1.5)$ and $N(11, 1.5)$. These densities are depicted in Figure 1, and were chosen to be far from normality. A total of 200 data sets were simulated, and each dataset was fitted by both parametric and nonparametric models. We next present the prior densities used.

3.2 Prior distributions

Under a parametric model, test results within the diseased and the disease free population are assumed to be normally distributed and prior distributions are specified on their respective means and standard deviations. As discussed earlier, the Polya tree priors here will be centered around a base probability distribution G which is also assumed to be normal, with the parameter c representing the strength of our prior belief that the data are in fact normally distributed.

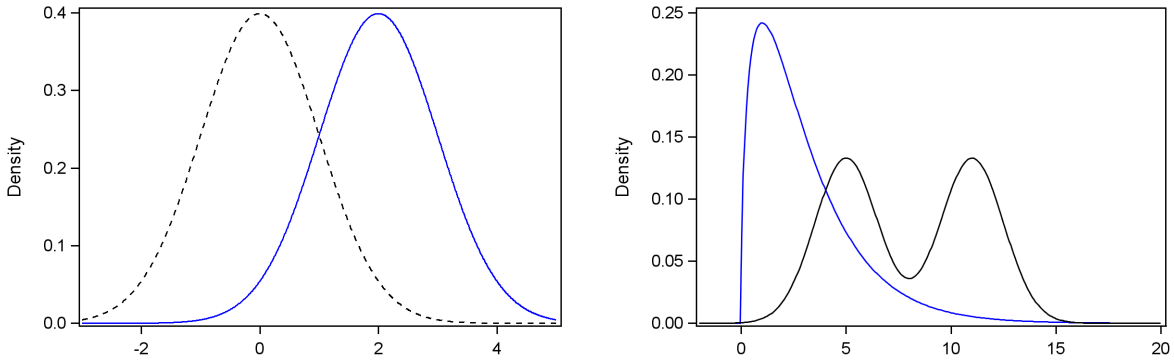

Figure 1: True probability densities of the diseased (dashed) and non-diseased (solid) populations used in the first set of simulations (left) and true probability densities of the diseased (dashed) and non-diseased (solid) populations used in the second set of simulations (tight).

We consider two sets of priors. The first set is centered about the true means and variances, representing the best case scenario where accurate prior information is available. The second set of priors are off-centered, that is, they are no longer centered on their true values. The exact values used are summarized in Table 1. For the parametric models the centered priors are elicited by putting normal prior distributions on the means μ_0 and μ_1 and uniform prior distributions on the standard deviations σ_0 and σ_1, centered and concentrated around their true values.

Table 1: Prior distributions across all simulations for both parametric and nonparametric models.

First set of simulations. Test data are normally distributed. True $D+ \sim N(0,1)$ and true $D- \sim N(2,1)$			
First set of priors (centered)		Second set of priors (not centered)	
Parametric model	Nonparametric model	Parametric model	Nonparametric model
$\mu_1 \sim N(0,1)$	$g_1 \sim N(0,1)$	$\mu_1 \sim N(-1,1)$	$g_1 \sim N(-1,1)$
$\mu_0 \sim N(2,1)$	$g_0 \sim N(2,1)$	$\mu_0 \sim N(3,1)$	$g_0 \sim N(3,1)$
$\sigma_1 \sim unif(0.8,1.2)$	$c = 5$	$\sigma_1 \sim unif(0.8,1.2)$	$c = 5$
$\sigma_0 \sim unif(0.8,1.2)$	$c = 10$	$\sigma_0 \sim unif(0.8,1.2)$	$c = 10$
	$c = 100$		$c = 100$
Second set of simulations. Test data are not normally distributed. True $D+ \sim \chi^2_3$ and true $D- \sim 0.5N(5,1.5) + 0.5N(11,1.5)$			
First set of priors (centered)		Second set of priors (not centered)	
Parametric model	Nonparametric model	Parametric model	Nonparametric model
$\mu_1 \sim N(2.366,0.25)$	$g_1 \sim N(2.366,\sqrt{6})$	$\mu_1 \sim N(1.366,0.25)$	$g_1 \sim N(1.366,\sqrt{6})$
$\mu_0 \sim N(8,0.25)$	$g_0 \sim N(8,3.35)$	$\mu_0 \sim N(9,0.25)$	$g_0 \sim N(9,3.35)$
$\sigma_1 \sim unif(2.25,2.65)$	$c = 5$	$\sigma_1 \sim unif(2.25,2.65)$	$c = 5$
$\sigma_0 \sim unif(3.15,3.55)$	$c = 10$	$\sigma_0 \sim unif(3.15,3.55)$	$c = 10$
	$c = 100$		$c = 100$

The centered priors for the finite Polya tree models, $F_j \sim PT_5(N(G_j, c), j = 0, 1$ consist of a normal base distribution centered at the true means and standard deviations of the data. Three values of c are used, 5, 10 and 100. The non-centered priors are obtained by shifting the location of the normal prior distributions for μ_0 and μ_1 one unit to the left and to the right, respectively. The off-centered ones were handled in a similar fashion. For the Polya tree models, all programming was carried out in SAS (Version 9.2, Cary NC).

3.3 Results from simulations

Our main focus will be on estimating the prevalence parameter π, but we will also report on the posterior predictive densities of the test results. Posterior distributions were generated using the Gibbs sampler with 100,000 posterior iterations after a burn-in of 5000. Results for the prevalence are presented in terms of bias (defined as the average median estimates over the 200 simulated data sets minus the true value of the parameter), the average length of the 95% credible intervals (CrI), and the coverage probabilities, estimated as the proportion of the 200 credible intervals containing the true prevalence value.

Table 2 summarizes the results when estimating the prevalence under the first set of conditions, that is, when the data are simulated from normal densities. Using centered priors, both the parametric and the nonparametric models estimate the prevalence with high accuracy. The nonparametric models generally offer greater precision than the parametric model, with precision increasing with the value of c and with the sample size. Coverage probabilities are close to one. The prevalence is still estimated with low bias across all models for the second set of priors, although bias is increased compared to estimates from the first set of priors. Credible intervals are narrower but coverage probabilities are below the nominal level value of 95%. Therefore, the nonparametric models offer greater precision at the expense of lower coverage, especially with $n = 5000$ where the highest coverage is only equal to 0.425.

Table 2: Bias, average length and coverage probabilities of the 95% Credible intervals across all simulations, based on 200 simulated data sets.

		n=1000			n=5000		
		Bias	Length of 95% CrI	Coverage of 95% CrI	Bias	Length of 95% CrI	Coverage of 95% CrI
Data are simulated from two normal distributions as described in Table 1	First set of priors (centered)						
	Parametric	0.003	0.329	0.995	0.017	0.221	1
	Nonparametric						
	c=5	0.003	0.154	1.000	0.002	0.124	1
	c=10	0.004	0.124	1.000	0.003	0.093	1
	c=100	0.001	0.084	0.995	0.001	0.048	1
	Second set of priors (not centered)						
	Parametric	0.024	0.339	0.995	0.019	0.227	0.925
	Nonparametric						
	c=5	0.020	0.082	0.865	0.023	0.048	0.400
	c=10	0.019	0.076	0.915	0.023	0.042	0.375
	c=100	0.021	0.067	0.855	0.017	0.032	0.425
Data are simulated from two non-normal distributions as described in Table 1	First set of priors (centered)						
	Parametric	0.143	0.122	0	0.223	0.046	0
	Nonparametric						
	c=5	0.066	0.294	0.94	0.174	0.140	0
	c=10	0.040	0.237	0.99	0.115	0.137	0.06
	c=100	0.019	0.113	0.98	0.038	0.064	0.29
	Second set of priors (not centered)						
	Parametric	0.168	0.109	0	0.227	0.046	0
	Nonparametric						
	c=5	0.060	0.219	0.86	0.162	0.147	0
	c=10	0.052	0.178	0.84	0.121	0.135	0.05
	c=100	0.034	0.095	0.74	0.039	0.064	0.34

Table 2 also summarizes results when estimating the prevalence when data are simulated from non-normal densities. Large bias and very low coverages render the use of the parametric model a very poor choice compared to the nonparametric models. When $n = 1000$, the nonparametric models have coverage probabilities ranging from 0.94 to 0.99. Bias and precision are acceptable under the nonparametric model for $n = 1000$, with the smallest bias, 0.019, and the smallest interval width, 0.113, obtained when $c = 100$. For a sample size of 5000, credible interval widths decrease under the nonparametric models, resulting in very poor coverage probabilities, ranging from 0 to 0.29. Despite the low coverage probabilities, the bias

when $c = 100$ is only 0.038. Similar bias is obtained under the second set of priors with slightly higher precision resulting in coverage probabilities ranging from 0.74 to 0.86 for $n = 1000$. For $N = 5000$ and with $c = 5$ and $c = 10$, the nonparametric models performed poorly because they are too flexible, with insufficient information from the priors. Sample size has to be taken into account when selecting a c parameter that achieves a trade-off between flexibility and identifiability is required, meaning in practice that small values of c should not generally be used with large sample sizes.

3.4 Randomized Polya trees: A simulated example

As an illustrative example of the use of randomized Polya trees, we simulated one data set, and compared results from a randomized Polya tree to that from a fixed Polya tree. This dataset was simulated under the conditions from the second set of simulations described above (Figure 1). We used the centered set of priors from Table 2 and a sample size of 1000. We used two values for the smoothing parameter in the randomized Polya tree model, $\tau = 0.01$ and 0.05.

Prevalence estimates (95% CrI) from a simple Polya tree was 0.223 (0.113, 0.331). The randomized Polya tree model estimates were 0.229 (0.122, 0.353) for $\tau = 0.01$ and 0.238 (0.105, 0.373) for $\tau = 0.05$. We note very similar point estimates and slightly larger CrI with the randomized Polya tree model comperd to the non-randomized model, with CrI width increasing with τ. The posterior predictive densities are presented in Figure 2. The posterior predictive densities are relatively close to the true densities, and as expected, there was additional smoothness with the randomized Polya tree model.

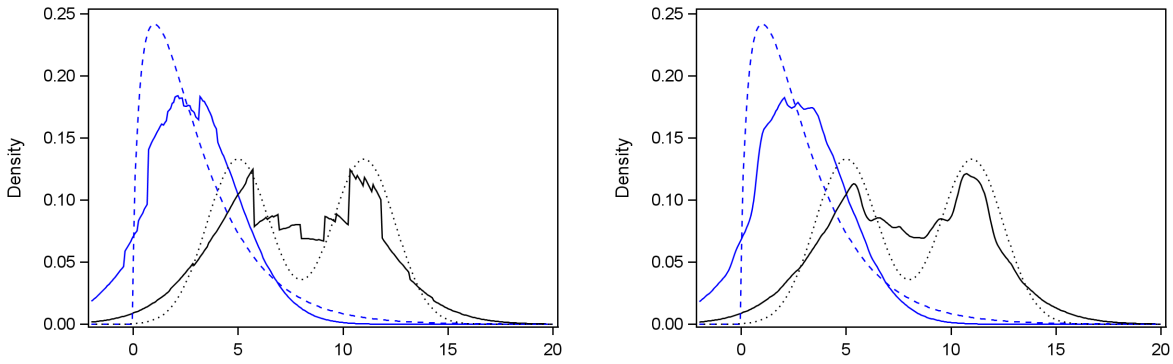

Figure 2: $D+$ and $D-$ posterior predictive density estimates under a fixed (left) and a randomized (right) Polya tree model. In both cases, the first set of priors with $c = 10$, as described in Table 2, were used. Dashed and dotted curves represent, respectively, the true $D+$ and $D-$ densities. Solid blue and black lines represent, respectively, the estimated $D+$ and $D-$ test result densities.

4 Comparing the performance of parametric to nonparametric models using Bayes Factors

In any given analysis of continuous diagnostic testing data, one might need to choose between a parametric and non-parametric model. The choice can be important, since as our simulations have shown, results can substantially differ between models. We now demonstrate how Bayes Factors can be calculated to aid in this decision.

Suppose that we would like to compare two plausible models, M_1 and M_2. The Bayes factor [18] is defined as the ratio of the posterior to prior odds of M_1 compared to M_2. When the two models are equally probable a priori, the Bayes factor reduces to the ratio of two marginal likelihoods. Many Monte Carlo methods for computing a marginal likelihood have appeared in the literature. The approach of Chib [19] uses the

output from the Gibbs sampler. An extension of Chib's method that uses the Markov chain Monte Carlo (MCMC) output produced by the Metropolis-Hastings algorithm has been presented by Chib and Jeliazkov [20]. Chen and Shao [21] used ratio importance sampling and Gelman and Meng [22] used path sampling for this problem.

Chen [21] proposed a new Monte Carlo method for computing a marginal likelihood in the presence of latent data from a single (MCMC) run using samples from the joint posterior distribution. Accommodating latent data is important here, since the unknown disease status for each subject is a latent variable in our diagnostic testing context. Chen's method does not require the specific structure or the details of the MCMC sampling algorithm to be known, and so is widely applicable.

Let $X = (x_1, ..., x_n)$ be the observed data, $m(X) = \int f(X, Z|\theta) f(\theta) d\theta dZ$, the marginal likelihood to be calculated, and let $L(\theta^*|X) = \int f(X, Z|\theta^*) dZ$. The point θ^* is in theory arbitrary, but for efficiency should be chosen to be a point of high posterior density. According to the method of Chen [21], given a random sample of size M, $\{\theta_j, Z_j = (z_1, ..., z_n)_j, j = 1, ..., M\}$ from the joint posterior distribution $f(\theta, Z|X)$, the log-marginal likelihood can be estimated as

$$\log m(X) = \log L(\theta^*|X) - \log \left[\frac{1}{M} \sum_{j=1}^{M} \frac{f(\theta^*|Z_j, X)}{f(\theta_j|Z_j, X)} \right]. \tag{3}$$

4.1 Marginal likelihood under a parametric model

According to (3), application of Chen's method requires a closed form expression for the posterior density of $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1, \pi)$ given the latent data. This condition is not satisfied here, but we know that in a normal model with fixed variance the mean has a closed form density [23]. Therefore, the marginal likelihood can be obtained by conditioning on (σ_0, σ_1) and then averaging over the prior density of (σ_0, σ_1). In order to apply Chen's method to calculate $m(X|\sigma_0, \sigma_1)$, it remains to calculate the integral $L(\mu_0^*, \mu_1^*, \pi^*|\sigma_0, \sigma_1, X)$, since the closed form density $f(\mu_0, \mu_1, \pi|\sigma_0, \sigma_1, Z, X)$ is available. Since σ_0 and σ_1 are held fixed here, $L(\mu_0^*, \mu_1^*, \pi^*|\sigma_0, \sigma_1, X)$ can be calculated analytically as

$$\begin{aligned} L(\mu_0^*, \mu_1^*, \pi^*|\sigma_0, \sigma_1, X) &= \prod_{i=1}^{n} \left\{ \int (\pi f_1(x_i))^{z_i} ((1-\pi) f_0(x_i))^{1-z_i} dz_i \right\} \\ &= \prod_{i=1}^{n} E_{bernouli(z_i|\pi)} \left\{ (f_1(x_i))^{z_i} (f_0(x_i))^{1-z_i} \right\} \\ &= \prod_{i=1}^{n} \pi f_1(x_i) + (1-\pi) f_0(x_i). \end{aligned} \tag{4}$$

4.2 Marginal likelihood under a nonparametric model

Recall that under a Polya tree model the vector of parameters consists of $\theta = (\pi, \mathcal{Y}^1, \mathcal{Y}^0)$, where π is the prevalence and $\mathcal{Y}^j$ denotes the set of branch probabilities associated with F_j, $j = 0, 1$. Let $M(X)$ be the marginal likelihood under the nonparametric model. From Bayes theorem, we can write

$$f(\pi, \mathcal{Y}^1, \mathcal{Y}^0, Z|X) = \frac{f(X, Z|\pi, \mathcal{Y}^1, \mathcal{Y}^0) f(\pi) f(\mathcal{Y}^1) f(\mathcal{Y}^0)}{M(X)}.$$

Integrating both sides with respect to $\mathcal{Y}^1$ and $\mathcal{Y}^0$, we obtain

$$f(\pi, Z|X) = \frac{\int f(X, Z|\pi, \mathcal{Y}^1, \mathcal{Y}^0) f(\pi) f(\mathcal{Y}^1) f(\mathcal{Y}^0) d\mathcal{Y}^1 d\mathcal{Y}^0}{M(X)}.$$

Since the branch probabilities have closed form beta posterior densities, we can factor out $\prod_{i=1}^{n} (\pi g_1(x_i))^{z_i}$ $\prod_{i=1}^{n} ((1-\pi)g_0(x_i))^{1-z_i} f(\pi)$ and obtain that

$$\int f(X, Z|\pi, \mathcal{Y}^1, \mathcal{Y}^0) f(\pi) f(\mathcal{Y}^1) f(\mathcal{Y}^0) d\mathcal{Y}^1 d\mathcal{Y}^0 = \phi(X, Z) \times \prod_{i=1}^{n} (\pi g_1(x_i))^{z_i} \prod_{i=1}^{n} ((1-\pi)g_0(x_i))^{1-z_i} f(\pi).$$

Let $h(X, Z|\pi) = \phi(X, Z) \times \prod_{i=1}^{n} (\pi g_1(x_i))^{z_i} \prod_{i=1}^{n} ((1-\pi)g_0(x_i))^{1-z_i}$, so that

$$f(\pi, Z|X) = \frac{h(X, Z|\pi) f(\pi)}{M(X)}. \tag{5}$$

Following Chen [21], we can replace $f(\pi, Z|X)$ by $h(X, Z|\pi)$, thus obtaining an expression for the marginal likelihood similar to what would be obtained in a latent class model where the only parameter is π.

It remains to calculate the integral $L(\pi^*|X) = \int h(X, Z|\pi^*) dZ$. We can write

$$L(\pi^*|X) = \left\{ \prod_{i=1}^{n} (\pi^* g_1(x_i) + (1-\pi^*) g_0(x_i)) \right\}$$

$$\times \int h(X, Z|\pi) \times \prod_{i=1}^{n} \left(\frac{\pi^* g_1(x_i)}{\pi^* g_1(x_i) + (1-\pi^*) g_0(x_i)} \right)^{z_i} \prod_{i=1}^{n} \left(\frac{(1-\pi^*) g_0(x_i)}{\pi^* g_1(x_i) + (1-\pi^*) g_0(x_i)} \right)^{1-z_i} dZ$$

We recognize

$$f(Z|\pi^*, X) = \prod_{i=1}^{n} \left(\frac{\pi^* g_1(x_i)}{\pi^* g_1(x_i) + (1-\pi^*) g_0(x_i)} \right)^{z_i} \prod_{i=1}^{n} \left(\frac{(1-\pi^*) g_0(x_i)}{\pi^* g_1(x_i) + (1-\pi^*) g_0(x_i)} \right)^{1-z_i}$$

as the posterior density of Z in a bi-normal model with fixed parameters, where the prevalence is fixed and equal to π^* and where the $D-$ and the $D+$ densities are fixed and equal to g_0 and g_1, respectively. In practice, given a sample of size M, $\{Z_j = (z_1, ..., z_n)_j, j = 1, ..., M\}$ from the posterior density $f(Z|\pi^*, X)$ under this bi-normal model, $L(\pi^*|X)$ can be estimated by

$$\left\{ \prod_{i=1}^{n} (\pi^* g_1(x_i) + (1-\pi^*) g_0(x_i)) \right\} \times \frac{1}{M} \sum_{j=1}^{M} h_0(x, Z_j, \pi^*) \times h_1(X, Z_j, \pi^*).$$

Having presented a method for computing the marginal likelihood under both parametric and nonparametric models, we next illustrate the calculation of Bayes Factors using two simulated data sets, one from normally distributed data, the other where the non-parametric model should better fit the data.

4.3 Simulated examples

We consider data simulated from normal and non-normal densities, similar to the simulations from Section 3, and illustrated in Figure 1. We again consider sample sizes of 1000, and a true prevalence of 0.25. The two datasets are fit first to a bi-normal model and then to a Polya tree model, each time using the centered priors presented in Table 1. We then compare the fit from each model using Bayes factors.

Table 3 contains MCMC estimates of the log marginal likelihood under the parametric models and the three nonparametric models. As expected, the parametric model has a better fit when compared to nonparametric models with low values of c ($c = 5$). The parametric model has slightly better fit than the nonparametric model with $c = 10$ and slightly worse fit than the nonparametric model with $c = 100$. When

Table 3: Log marginal likelihood and Bayes factor (BF) estimates for both normal and non-normal sampling situations.

	Model	Log Marginal Likelihood	BF	P (parametric model is better)
	Parametric	-1696.480	1	
	Nonparametric (fixed partition)			
True population distributions are normal	c=5	-1699.175	14.806	0.937
	c=10	-1697.800	3.743	0.789
	c=100	-1695.932	0.578	0.366
$D+ \sim N(0,2)$ and $D- \sim N(0,1)$	Nonparametric (random partition, $\tau = 0.05$)			
	c=5	-1699.435	0	0
	c=10	-1697.998	4.563	0.820
	c=100	-1695.954	0.591	0.371
	Parametric	-2737.187	1	
	Nonparametric (fixed partition)			
True population distributions are not normal	c=5	-2687.952	0	0
	c=10	-2699.132	0	0
	c=100	-2741.697	90.922	0.989
$D+ \sim \chi^2_3$ and $D- \sim 0.5N(5,1.5) + 0.5N(11,1.5)$	Nonparametric (random partition, $\tau = 0.05$)			
	c=5	-2687.298	0	0
	c=10	-2698.506	0	0
	c=100	-2742.421	187.541	0.995

$c = 100$, there is probably less uncertainty about the distributions of test results in the nonparametric model than in the parametric model, which explains why the nonparametric model has a slightly better fit.

Nonparametric models with smaller values of c have better fit when test results are not normally distributed. Compared to the parametric model, Polya tree models with $c = 5$ and $c = 10$ have considerably better fit. With $c = 100$, the nonparametric model has poorer fit than the parametric model, with a log-marginal likelihood difference of 6.5 for the second data set and a difference of 4.5 for the third data set. This implies that when the data are not normally distributed, using a nonparametric model where the random probability distribution are forced to be very concentrated around normal base distributions (i.e., using very high values of c) can lead to a model with very poor fit. For these three simulated datasets randomized Polya tree models are similar to simple Polya tree models in term of fit.

5 Application to tuberculosis clustering from DNA fingerprint data using Nearest Genetic Distance data

To return to the tuberculosis clustering problem described in the introduction, we now fit Polya tree models to an NGD data set. We will compare inferences from a model with fixed and random partitions as well as compare the fit of the Polya tree model to that from a parametric bi-normal model, using Bayes factors. The data set consists of test results from 393 subjects with recently diagnosed active TB in Montreal [24].

For our Polya tree model, we used the same prior means and standard deviations for the base distributions elicited by Ladouceur et al. [6] under a Dirichlet process model. In particular, we take $g_1 \sim N(28, 17)$ and $g_0 \sim N(134, 63)$. The means and standard deviations of these base densities were created by averaging the information provided by three experts who provided prior densities independently [6]. We set the depth of our Polya tree to $K = 5$. The small sample size (393) suggests avoiding large values for c, as that would be similar to using a parametric model with fixed hyperparameters. Accordingly, in order for the nonparametric models to retain some flexibility, we chose to run models with $c = 2$ and $c = 5$. For the random partition model we set $\tau = 0.05$. For comparison purposes, we will also fit the same data set to a bi-normal parametric model, using the following prior densities for the means and the standard deviations: $\mu_0 \sim N(134, 20)$, $\mu_1 \sim N(28, 10)$, $\sigma_0 \sim Uniform(43, 83)$ and $\sigma_1 \sim Uniform(7, 27)$. These priors were de-

rived by placing normal densities on the means and uniform priors on the standard deviations and centering these priors around the means and standard deviations of the base densities elicited by Lacouceur et al. [6].

Posterior predictive densities for the $D+$ and $D-$ populations are displayed in Figure 3, where we can see that the random partition model has smoother density plots. Prevalence of clustering estimates are summarized in Table 4. Results from all parametric and nonparametric models are similar, with an estimate of the prevalence of approximately 20%. Lengths of the 95% credible intervals from all models are similar and are close to 10%, a high precision considering the small sample size ($n = 393$). This can be explained by the low degree of overlap between the disease and disease-free densities estimated, as shown in Figure 3. The small overlap can also explain the similarities between the results obtained using parametric and nonparametric models, even though the estimated densities from the nonparametric model have non-normal shapes.

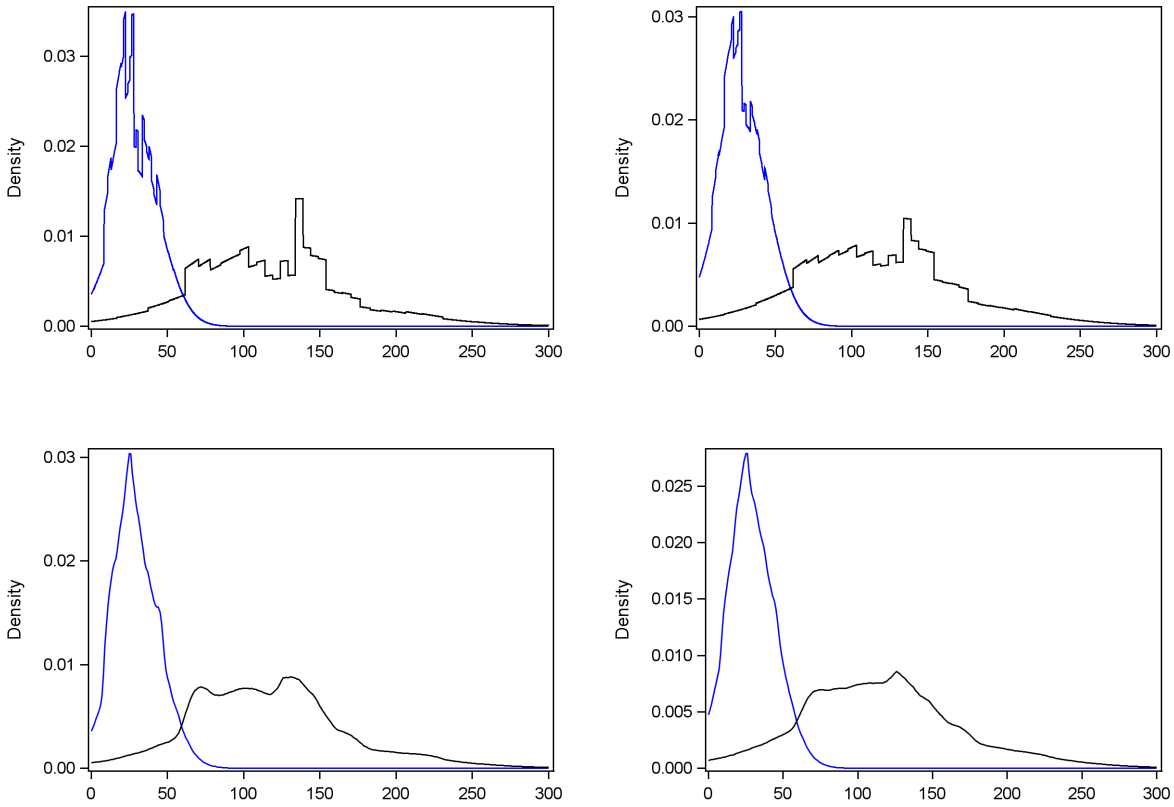

Figure 3: Posterior density estimates from Polya tree models with a fixed partition with $c = 2$ (top left) and $c = 5$ (top right). The bottom figures represent density estimates from a randomized Polya tree with $c = 2$ (bottom left) and $c = 5$ (bottom right).

Table 4: Posterior estimates of the prevalence parameter across parametric and nonparametric models.

Model	Median (95%CrI) for π
Parametric	0.184 (0.121, 0.250)
Nonparametric (fixed parition)	
$c = 2$	0.193 (0.132, 0.254)
$c = 5$	0.180 (0.127, 0.238)
Nonparametric (random partition, $\tau = 0.05$)	
$c = 2$	0.192 (0.129, 0.254)
$c = 5$	0.177 (0.119, 0.237)

Table 5 presents log-marginal likelihood estimates from the parametric and the nonparametric models under the prior distributions described in Section 4. The nonparametric models offer a better fit than the parametric bi-normal model, an expected result given the non-normal shape of the density estimates. We also notice that the simple Polya tree model offers a better fit than the randomized Polya tree model with low values of c. We note that even though the data is not normal as shown by the Bayes factor, the parametric and the nonparametric models gave very similar estimates for the prevalence. This can be explained by similarity of curve shapes in the area of overlap between the clustered and the nonclustered test results.

Table 5: Log marginal likelihood and Bayes factor (BF) estimates across parametric and nonparametric models.

Model	Log Marginal Likelihood	BF	P (parametric model is better)
Parametric	-2138.482	1	
Nonparametric (fixed partition)			
c=2	-2108.488	0	0
c=5	-2119.717	0	0
c=10	-2127.663	0	0
Nonparametric (random partition, $\tau = 0.05$)			
c=2	-2118.064	0	0
c=5	-2123.310	0	0
c=10	-2127.831	0	0

6 Discussion

In this paper we carefully investigated the properties of finite Polya tree models for continuous diagnostic tests via simulations. These simulations have been carried out under two different scenarios. Acknowledging that when test results from different disease classifications have a large overlap, strong and accurate prior information is needed to have any hope of reasonable inferences in this non-identifiable problem, we limited our simulations to the case with a reasonable overlap given that no other information is available about the disease status of each individual. Following Ladouceur et al. [6], we considered relatively large sample sizes only ($n = 1000$ and $n = 5000$). Smaller sizes will rarely provide accurate inferences for such difficult problems. In order to keep the number of simulations manageable, we used the same degree of belief in the distribution of test data being close to normal from both the diseased and the disease free populations. Other scenarios are possible where the shape of one density is believed to be closer to normal than the other.

Overall, as expected, the nonparametric Polya tree models that are not too flexible performed better than the normal model when the true distributions of the data are not normal. To answer the concern about partition dependence of the Polya tree model, we considered a model based on randomized Polya trees. Although we did not carry out a full set of simulations to investigate the properties of such model, credible intervals obtained from a randomized Polya tree model were slightly larger than those obtained under a simple Polya tree model. While median prevalence estimates were similar for the two models, the randomized Polya tree presented a more aesthetic appearance, owing to greater smoothness. Finally, we showed how Bayes Factors can be used to help decide whether a nonparametric or a bi-normal model may better fit the data.

References

[1] Salamon. H., Behr. M., Rhee. J., Small. P. (2000). Genetic distances for the study of infectious disease epidemiology. American Journal of Epidemiology 151, 324–334.

[2] Goetghebeur, E., Liinev, J., Boelaert, M., and Van der Stuyft, P. (2000). Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. Statistical Methods in Medical Research 9, 231–248.

[3] Rutjes, A., Reitsma, J., Coomarasamy, A., Khan, K., and Bossuyt, P. (2007). Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment 50, 1–51.

[4] Joseph, L., Gyorkos, T., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. American Journal of Epidemiology 141, 263–272.

[5] Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. Statistical Science 20, 111–140.

[6] Ladouceur, M., Rahme, E., Scott, A., Schwartzman, K. and Joseph, L. (2011). Modeling continuous diagnostic test data using approximate Dirichlet process distributions. Statistics in Medicine 30(21), 2648–2662.

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

[8] Scott, A., Joseph, L., Bélisle, P., Behr, M. and Schwartzman, K. (2008). Bayesian estimation of tuberculosis clustering rates from DNA sequence data. Statistics in Medicine 27, 140–156.

[9] Branscum, A.J., Johnson, W.O., Hanson, T.E., and Gardner, I.A., (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. Statistics in Medicine 27, 2474–2496.

[10] Paddock, S., Ruggeri, F., Lavine, M. and West, M. (2003). Randomized Polya tree models for nonparametric Bayesian inference. Statistica Sinica 13, 443–460.

[11] Freedman, D.A. (1963), On the Asymptotic Behavior of Bayes Estimates in the Discrete Case, The Annals of Mathematical Statistics, 34, 1386–1403.

[12] Fabius, J. (1964), Asymptotic behavior of Bayes estimates, The Annals of Mathematical Statistics, 35, 846–856.

[13] Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. The Annals of Statistics 2, 615-629.

[14] Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. Annals of Statistics 20, 1222–1235.

[15] Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. Annals of Statistics 22, 1161–1176.

[16] Mauldin, R.D., Sudderth, W.D. and William, S.C. (1992). Polya trees and random distributions. The Annals of Statistics 20, 1203–1221.

[17] Dendukuri, N., Rahme, E., Bélisle, P., Joseph, L. (2004). Bayesian sample size determination for prevalence and diagnostic studies in the absence of a gold standard test. Biometrics 60, 388–397.

[18] Kass, R. and Raftery, A. (2005). Bayes factors. Journal of the American Statistical Association 90, 773–795.

[19] Chib, S. (1995). Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

[20] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. Journal of the American Statistical Association 96, 270–281.

[21] Chen, M.H. (2005). Computing marginal likelihoods from a single MCMC output. Statistica Neerlandica 59, 16–29.

[22] Gelman, A., and Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science 13, 163–185.

[23] Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (2004). Bayesian Data Analysis, Chapman and Hall.

[24] Scott, AN., Menzies, D., Tannenbaum, TN., Thibert, L., Kozak, R., Joseph, L., Schwartzman, K. and Behr, MA. (2005). Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. Journal of Clinical Microbiology 43, 89–94.