Journal of Data Science **12**(2014), 325-338

A New Polya Tree Construction Facilitating A Goodness-of-Fit Test

Yuhui Chen
Department of Mathematics, The University of Alabama

Abstract: Polya tree, by embedding parametric families as a special case, provides natural suit to test goodness of fit of a parametric null with non-parametric alternatives. For this purpose, we present a new construction on Polya tree for random probability measure, which aims to perform an easy multiple χ^2 test for goodness of fit. Examples of data analyses are provided in simulation studies to highlight the performance of the proposed methods.

Key words: Goodness-of-Fit Test, Nonparametric Model, Polya tree.

1. Introduction

Polya tree (PT), as a random probability measure, has recently received more and more attention due to its much greater tractability over tailfree processes as well as its continuous or absolutely continuous probability measure over Dirichlet processes (Ferguson, 1973, 1974). The basic idea for Polya trees arises from Ferguson (1974), Lavine (1992, 1994), and Mauldin, Sudderth, and Williams (1992). Later on, more generalizations, such as the mixture of Polya trees (Hanson and Johnson, 2002; Hanson, 2006), the multivariate Polya trees (Paddock et. al, 2003), and the optional Polya trees (Wong and Ma, 2010), are developed and flourished in many areas, for example, the Polya trees for spatial frailty modeling (Zhao et. al 2009), the Polya tree sampler (Hanson et. al, 2011), the Polya trees for multiple sample tests (Chen and Hanson, 2014), and the Polya trees for truncated data (Chen and Hanson, 2014).

A Polya tree F is defined as $F \sim \mathrm{PT}(\Pi, \mathcal{A})$ (Lavine, 1992, 1994). The parameter $\Pi = \cup_{j=0}^{\infty} \Pi_j$ is a nested sequence of partitions with the jth partition $\Pi_j = \cup_{\forall \epsilon_j} B_{\epsilon_j}$, where B_{ϵ_j} is the one of the 2^j base sets of Π_j with $B_{\epsilon_j} \cap B_{\epsilon_j'} = \emptyset$ if $\epsilon_j \neq \epsilon_j'$, and $\epsilon_j = \{0,1\}^j$ is defined as a j digits binary fold with $\epsilon_0 = \emptyset$. Specifically, $\Pi_0 = \{\Omega\}$, where Ω is a separable measurable space, and $B_{\epsilon_0} \stackrel{def}{=} B_{\emptyset} = \Omega$.

Generally, by assuming a parametric distribution F_θ (also called centering distribution), the nested partitions are obtained in such a way that, at each partition Π_j for $j \in \{1, 2, \ldots, J\}$, we obtain all 2^j base sets with each base set B_{ϵ_j} satisfying $F_\theta(B_{\epsilon_j}) = 1/2^j$. For example, at Π_1,$B_0 = [F_\theta^{-1}0, F_\theta^{-1}(\frac{1}{2})]$ and $B_1 = [F_\theta^{-1}(\frac{1}{2}), F_\theta^{-1}(1)]$; at Π_2, $B_{00} = [F_\theta^{-1}(0), F_\theta^{-1}(\frac{1}{4})]$, $B_{01} = [F_\theta^{-1}(\frac{1}{4}), F_\theta^{-1}(\frac{1}{2})]$, $B_{10} = [F_\theta^{-1}(\frac{1}{2}), F_\theta^{-1}(\frac{3}{4})]$, and $B_{11} = [F_\theta^{-1}(\frac{3}{4}), F_\theta^{-1}(1)]$. The refinement of each partition, say Π_{j+1}, is obtained by firstly independently assigning Beta priors to 2^j conditional probabilities. Specifically, for one of the 2^{j+1} base sets $B_{\epsilon_j 0}$, we define $\mathcal{Y}_{\epsilon_j 0} = F(B_{\epsilon_j 0}|B_{\epsilon_j})$ and assign a prior $\mathcal{Y}_{\epsilon_j 0} \sim \text{Beta}(\alpha_{\epsilon_j 0}, \alpha_{\epsilon_j 1})$. Note, B_{ϵ_j} is one of the base sets of Π_j and the parent of the two children of $B_{\epsilon_j 0}$ and $B_{\epsilon_j 1}$ with $B_{\epsilon_j 0} \cup B_{\epsilon_j 1} = B_{\epsilon_j}$ and $B_{\epsilon_j 0} \cap B_{\epsilon_j 1} = \emptyset$. Also, we have $\mathcal{Y}_{\epsilon_j 0} + \mathcal{Y}_{\epsilon_j 1} = 1$ with $\mathcal{Y}_{\epsilon_j 1} = F(B_{\epsilon_j 1}|B_{\epsilon_j})$. The posterior of $\mathcal{Y}_{\epsilon_j 0}$ has a conjugate property, i.e., $\mathcal{Y}_{\epsilon_j 0}|\text{Data} \sim \text{Beta}(\alpha_{\epsilon_j 0} + n_{\epsilon_j 0}, \alpha_{\epsilon_j 1} + n_{\epsilon_j 1})$, where $n_{\epsilon_j 0} = \sum_{i=1}^n \text{I}(x_i \in B_{\epsilon_j 0})$ and $n_{\epsilon_j 1} = \sum_{i=1}^n \text{I}(x_i \in B_{\epsilon_j 1})$ with $n_{\epsilon_j 0} + n_{\epsilon_j 1} = n_{\epsilon_j} = \sum_{i=1}^n \text{I}(x_i \in B_{\epsilon_j})$; $\text{I}(x \in A)$ is an indicator function and equals to 1 if $x \in A$ and 0 otherwise. Note, $n_{\epsilon_0} \equiv n$ is the size of the data set. The nonnegative parameter $\mathcal{A}$ is the set of the hyperparameter α(s) for Beta distributions. Keeping refinement for each base set at each partition, asymptotically, the true density F will be obtained (Lavine, 1992, 1994).

In Section 2, we introduce a new Polya tree construction similar to this scheme except: At each partition Π_j, the base sets are directly obtained from the previous refined distribution F_{j-1} at Π_{j-1} instead of from the common assumed F_θ. In Section 3, a multiple χ^2 test is constructed with adjusted Bonferroni corrections (BC) to examine whether or not a nonparametric model is needed when the centering parametric model F_θ is pre-specified. We conduct simulation studies in Section 4.1 and Section 4.2 to examine the performance of the proposed methods in Section 2 and Section 3 respectively. The Galaxy data set is studied in Section 5. Then we come up the conclusion in Section 6.

2. Construction

Let $x_i \sim F$ for $i = 1, 2, \ldots, n$ and n is the sample size. F is an unknown continuous probability distribution function with the support domain Ω and constructed as follows. Given the refined partition Π_j and the associated refined distribution F_j for $j = 0, 1, 2, \ldots$ with $\Pi_0 = \{\Omega\}$ and $F_0 \overset{def}{=} F_\theta$, for each base set $B_{\epsilon_j} \in \Pi_j$ with $B_{\epsilon_0} = B_\emptyset \overset{\text{def}}{=} \Omega \in \Pi_0$, define $r_{\epsilon_j 0} = F_{j+1}(B_{\epsilon_j 0}|B_{\epsilon_j})$. Similar to the typical Polya trees construction, let $r_{\epsilon_j 1} = F_{j+1}(B_{\epsilon_j 1}|B_{\epsilon_j})$, we have $r_{\epsilon_j 0} + r_{\epsilon_j 1} = 1$ with $B_{\epsilon_j 0}$ and $B_{\epsilon_j 1}$ is the two children of B_{ϵ_j} satisfying $B_{\epsilon_j 0} \cup B_{\epsilon_j 1} = B_{\epsilon_j}$ and $B_{\epsilon_j 0} \cap B_{\epsilon_j 1} = \emptyset$. By independently assigning a Beta prior $r_{\epsilon_j 0} \sim \text{Beta}(\alpha_{\epsilon_j 0}, \alpha_{\epsilon_j 1})$, the conjugate posterior of $r_{\epsilon_j 0}$ is then obtained by

$$r_{\epsilon_j 0}|\text{Data} \sim \text{Beta}(\alpha_{\epsilon_j 0} + n_{\epsilon_j 0}, \alpha_{\epsilon_j 1} + n_{\epsilon_j 1}), \tag{1}$$

where $n_{\epsilon_j 0}$ and $n_{\epsilon_j 1}$ are defined as the same with the typical Polya trees illustrated in Section 1. Obviously, the posterior of $r_{\epsilon_j 0}$ refines F_j on the base set $B_{\epsilon_j 0}$ given B_{ϵ_j}. If we keep refining all base sets at Π_{j+1} in this way, F_j is eventually refined at the level $j+1$. We denote this refinement at Π_{j+1} as F_{j+1}. However, unlike the typical Polya trees, the base set $B_{\epsilon_j 0}$ is obtained by

$$B_{\epsilon_j 0} = \{x : F_j(x) \leq m_{\epsilon_j 0}, \text{ for all } x \in B_{\epsilon_j}\}, \tag{2}$$

where

$$m_{\epsilon_j 0} = \frac{\min_{\epsilon_j} + \max_{\epsilon_j}}{2}. \tag{3}$$

Before illustrating the way to obtain $\min_{\epsilon_j}$ and $\max_{\epsilon_j}$, let firstly define the nested partitioning m-matrix as

$$m = \begin{bmatrix} m_0 & \dots & \dots & \dots & \dots & \dots & \dots; & J=1 \\ m_{00} & b_0 & m_{10} & \dots & \dots & \dots & \dots; & J=2 \\ m_{000} & b_{00} & m_{010} & b_0 & m_{100} & b_{10} & m_{110}; & J=3 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots; & \dots \end{bmatrix} \tag{4}$$

and the nested refining b-matrix as

$$b = \begin{bmatrix} b_0 & \dots & \dots & \dots & \dots & \dots & \dots; & J=1 \\ b_{00} & b_0 & b_{10} & \dots & \dots & \dots & \dots; & J=2 \\ b_{000} & b_{00} & b_{010} & b_0 & b_{100} & b_{10} & b_{110}; & J=3 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots; & \dots \end{bmatrix} \tag{5}$$

with $m_{\epsilon_j 0}$ obtained from Equation (3) and
$b_{\epsilon_j 0}$ obtained from

$$b_{\epsilon_j 0} = \min_{\epsilon_j} + r_{\epsilon_j 0}(\max_{\epsilon_j} - \min_{\epsilon_j}), \tag{6}$$

here $r_{\epsilon_j 0}$ is the posterior obtained from (1). Now we define $\min_{\epsilon_j} = b[j, (k_{\epsilon_j 0} - 1)/2]$ if B_{ϵ_j} is not the most left children in Π_j, otherwise $\min_{\epsilon_j} = 0$; and $\max_{\epsilon_j} = b[j, (k_{\epsilon_j 0} + 1)/2]$ if B_{ϵ_j} is not the most right children in Π_j, otherwise $\max_{\epsilon_j} = 1$. $k_{\epsilon_j 0}$ is the decimal expression of the binary fold $\epsilon_j 0$, for example, $k_{00} = 1$ and $k_{10} = 3$; $b[i, j]$ is the element of the b-matrix at the ith row and the jth column. Then, for $x \in B_{\epsilon_j} \in \Pi_j$, the refined distribution function $F_{j+1}(x)$ at Π_{j+1} can be obtained from:

$$F_{j+1}(x) = \begin{cases} \min_{\epsilon_j} + \frac{b_{\epsilon_j 0} - \min_{\epsilon_j}}{m_{\epsilon_j 0} - \min_{\epsilon_j}} \left[F_j(x) - \min_{\epsilon_j}\right] & x \in B_{\epsilon_j 0} \\ b_{\epsilon_j 0} + \frac{\max_{\epsilon_j} - b_{\epsilon_j 0}}{\max_{\epsilon_j} - m_{\epsilon_j 0}} \left[F_j(x) - m_{\epsilon_j 0}\right] & x \in B_{\epsilon_j 1}. \end{cases} \tag{7}$$

Here, F_0 (i.e. F_θ) can be arbitrarily chosen from any continuous parametric family as long as $\Omega_F \subseteq \Omega_{F_\theta}$ is satisfied, where Ω_F and Ω_{F_θ} are the support domains of the true distribution F and the centering distribution F_θ respectively, to guarantee a convergence to the true distribution F (Theorem 1 in Lavine, 1994). Note, through the construction we have $F_j(B_{\epsilon_j 0}|B_{\epsilon_j}) = \frac{1}{2}$, by letting $\alpha_{\epsilon_j 0} = \alpha_{\epsilon_j 1}$ for $\forall \epsilon_j$ and $j = 0, 1, \ldots$, if $F = F_\theta$, then we have $E(F_1) = F_\theta, E(F_2) = F_1, \ldots, E(F_{j\to\infty}) = F_{j-1}$.

Illustratively, starting from the first partition level $j = 1$, since $\min_\emptyset = 0$ and $\max_\emptyset = 1$, the partition point $m_0 = \frac{0+1}{2} = 0.5$ and thus $B_0 = \{x : F_\theta(x) \leq m_0, \text{ for all } x \in \Omega\}$ and $B_1 = \Omega - B_0$. From Equation (6), the refined b_0 equals to the posterior of r_0, i.e. $F_1(B_0|\Omega)$. The refined distribution at $j = 1$ and $x \in \Omega$, corresponding to Equation (7), then can be written as:

$$F_1(x) = \{\frac{b_0}{m_0} F_\theta(x) \mathrm{I}(x \in B_0)\} + \{b_0 + \frac{1-b_0}{1-m_0}(F_\theta(x) - m_0)\mathrm{I}(x \in B_1)\}. \tag{8}$$

Continuously refining the base set B_0 at $j = 2$, we get $B_{00} = \{x : F_1(x) \leq m_{00}, \text{ for all } x \in B_0\}$, where $m_{00} = \frac{0+b_0}{2}$. Thus we have:

$$F_2(x) = \begin{cases} \frac{b_{00}}{m_{00}}\left[F_1(x)\right] & x \in B_{00} \\ b_{00} + \frac{b_0 - b_{00}}{b_0 - m_{00}}\left[F_1(x) - m_{00}\right] & x \in B_{01}, \end{cases} \tag{9}$$

here b_{00} is obtained by Equation (6). Recursively, for each base set at each partition level, the corresponding elements in the m-matrix and in the b-matrix are refined nestedly in order. And as long as $\Omega_F \subseteq \Omega_{F_\theta}$, the above refinement process has full support in the set of all continuous probability measures as $j \to \infty$ (Theorem 1 in Lavine, 1994). However, it's not necessary to fit an infinite Polya tree due to the robust Bayesian consideration since at the bottom of the tree, it's much possible that the refinement of the partitions only depends on the prior (Lavine, 1994). Thus the Polya tree can be updated to a predetermined level as long as α's increase rapidly toward the bottom of the tree to guarantee to give a probability *one* to the set of absolutely continuous distributions. Throughout the paper, we cap j at $J = \lceil \log_4(n) \rceil$ and find it works well in the simulation studies.

There are many ways to choose the hyper-parameters αs. One typical choice is to let $\alpha_{\epsilon_j} = cj^2$, which makes F is absolutely continuous with respect to 1 (Ferguson, 1974); or let $\alpha_{\epsilon_j} = c/2^j$ to make F to be a Dirichlet Process (Blackwell, 1973; Blackwell and MacQueen, 1973). The parameter c here can be considered as a precision parameter (Ferguson, 1973, 1974; Lavine, 1992, 1994) and $c > 0$. When $c \to \infty$, $F \to F_\theta$; and when $c \to 0$, an empirical distribution will be obtained. Usually, one can let $c = 1$ or assign a prior as $c \sim \Gamma(5, 1)$ (Hanson, 2006).

For inference purpose, the quantile function $F_J^{-1}(q)$ of a finite Polya tree, where $0 \leq q \leq 1$, can be easily calculated from a grid searching algorithm by starting from an arbitrarily chosen x^*, and then increasing x^* by a small value δ if $F_J(x^*) < q$ or decreasing x^* by δ if $F_J(x^*) > q$.

3. Goodness-of-Fit Test

Polya trees are naturally suited to a goodness-of-fit test as they can embed continuous parametric families as a special case (Berger and Guglielmi, 2001). Along this line, we will propose a new approach in this section to answer the question whether or not it is necessary to fit a nonparametric model when the centering parametric model F_θ is pre-specified. By the construction in Section 2, the answer of this question is naturally the answer for a multiple comparisons test. Let define this multiple comparisons test as the following:

$$\begin{aligned} H_0 : &\underbrace{F_\theta = F_1}_{\text{Test 1}};\ \underbrace{F_1 = F_2}_{\text{Test 2}};\ \ldots;\ \underbrace{F_{J-1} = F_J}_{\text{Test J}} \\ H_1 : &\text{at least one is not equal} \end{aligned} \tag{10}$$

A multiple χ^2 test, a natural approach to answer such a goodness of fit test question, is then constructed by comparing the currently refined frequencies of the base sets to the previous ones for each partition. But how to obtain the χ^2 test statistics at each partition level? One might have already noticed, by the construction, for example, at the partition Π_2, before the refinement, we have $F_1(B_{00}|B_0) = m_{00}$, $F_1(B_{01}|B_0) = b_0 - m_{00}$, $F_1(B_{10}|B_1) = m_{10} - b_0$, and $F_1(B_{11}|B_1) = 1 - m_{10}$. However, after the refinement, $F_2(B_{00}|B_0) = b_{00}$, $F_2(B_{01}|B_0) = b_0 - b_{00}$, $F_2(B_{10}|B_1) = b_{10} - b_0$, and $F_2(B_{11}|B_1) = 1 - b_{10}$. Consequently, the frequencies of the base sets in Π_2, for the before and the after refinements, is nothing but the inverse of those probabilities. Thus, we could define the test statistics for the jth individual hypothesis test as

$$\mathrm{ST}_j = \sum_{k=1}^{2^j} \left[O_k - E_k\right]^2 / E_k, \text{ for } j \in \{1, 2, \ldots, J\}, \tag{11}$$

where $\mathrm{O}_k = 1/(b[j,k] - b[j,k-1])$ and $\mathrm{E}_k = 1/(m[j,k] - m[j,k-1])$ for $k \in \{2, \ldots, 2^j - 1\}$, $\mathrm{O}_k = 1/b[j,k]$ and $\mathrm{E}_k = 1/m[j,k]$ for $k = 1$, and $\mathrm{O}_k = 1/(1 - b[j,k-1])$ and $\mathrm{E}_k = 1/(1 - m[j,k-1])$ for $k = 2^j$. If the jth equality holds, i.e. $F_{j-1} = F_j$, then ST_j follows a χ^2 distribution with the degree of freedom $2^j - 1$, denoted as $\mathrm{ST}_j \sim \chi^2_{2^j-1}$. Simultaneously, if we can show all the J equalities holds in (10), then we can conclude that there is no need to

fit a nonparametric model since the parametric one, i.e. F_θ, has already well interpreted the data.

Maintaining the familywise error rate (FWER) in multiple dependent or independent hypotheses test is another critical concern. An easy and often used approach is Bonferroni correction which corrects each individual hypothesis at a statistical significance level of α/J, here J is the number of hypotheses test under the null. The proof for Bonferroni correction follows from Boole's inequality: FWER $= Pr\{\cup_j(p_j \leq \frac{\alpha}{J})\} \leq \sum_j Pr(p_j \leq \frac{\alpha}{J}) \leq J\frac{\alpha}{J} = \alpha$. However, based on our construction, giving each individual hypothesis the same correction might not be appropriate. Since as j increases, a prior driven pattern is more likely to be obtained. Thus we might need to adjust Bonferroni correction by assigning a small correction α_j for a large j for compensation of more likely witnessing a prior driven pattern at this large j. Let define a weight w_j as

$$w_j = \frac{1/2^j}{\sum_{k=1}^{J} 1/2^k}, \text{ for } j \in \{1, 2, \ldots, J\}. \tag{12}$$

Clearly, if $j > j^{'}$, then $w_j < w_{j^{'}}$. The adjusted Bonferroni correction for the jth individual hypothesis test thus can be defined as $\alpha_j = w_j\alpha$. And if $p_j < \alpha_j$, where p_j is the p-value of the jth individual hypothesis test and obtained from the jth χ^2 test, then we can conclude there exists a significant difference between F_{j-1} and F_j; otherwise, the equality $F_{j-1} = F_j$ holds.

4. Simulation Study

4.1 Model Fit

In this section, we will examine the performance of our proposed PT construction in Section 2. Data is generated from the following simulation cases with different sample sizes $n = 100$, $n = 200$, $n = 500$ and $n = 1000$. However, to make the paper concise, here we only report the results from $n = 500$ and $n = 1000$. The results from $n = 100$ and $n = 200$ behaviors quite similar except that the variance is larger than the one fitted from a larger sample size. Under each simulation case, we consider two different centering distribution F_θ. The J is capped at $\lceil \log_4(n) \rceil$ and $c = 1$. 100 fitted distribution function F_Js are plotted in each Figure 1 to Figure 12.

1. $x \sim \Gamma(2, 5)$
   - a: $F_\theta = \mathrm{T}(d.f. = 2)$
   - b: $F_\theta = \mathrm{Exp}(\mathrm{scale} = 1)$

2. $x \sim \text{Beta}(2, 1)$
   - a: $F_\theta = \Gamma(\text{shape} = 1, \text{scale} = 0.5)$
   - b: $F_\theta = \text{N}(\text{mean} = 0, \text{sd} = 1)$
3. $x \sim \text{N}(\text{mean} = 0, \text{sd} = 1)$
   - a: $F_\theta = \text{T}(d.f. = 1)$
   - b: $F_\theta = \Gamma(\text{shape} = 1, \text{scale} = 2)$

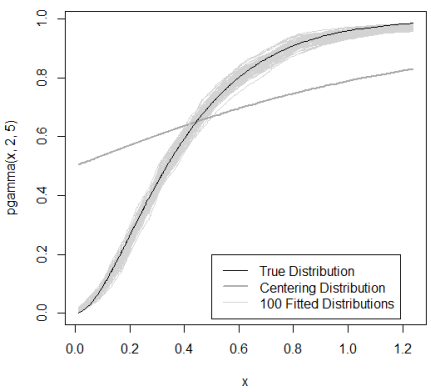

Figure 1: simulation case $1(a)$ with $n = 500$

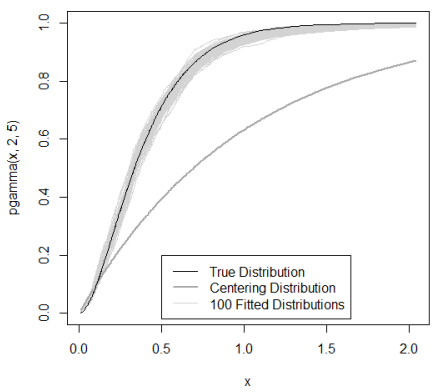

Figure 2: simulation case $1(b)$ with $n = 500$

From the simulation study, see Figure 1 – Figure 9 and Figure 11, we find, as long as $\Omega_F \subseteq \Omega_{F_\theta}$ is satisfied, the fitted distribution F_J approaches the true distribution F very well and the deviation between the F and F_J decreases as the sample size increases. However, when $\Omega_F \supset \Omega_{F_\theta}$, see Figure 10 and Figure 12, F_J fails to approach the true distribution F due to the reason that $F_\Gamma(x)$ is undefined when $x \leq 0$, see the simulation case $3(b)$. Solving this problem to get a robust F_J is actually very simple: one, for example, can always choose a centering distribution with $\Omega_{F_\theta} \equiv \mathbb{R}$ to guarantee $\Omega_F \subseteq \Omega_{F_\theta}$; or one can conduct a pilot study, which not only can help to find an appropriate support domain, but also can help the convergence rate. Here, we refer to a *good* convergence rate as the rate at which F_J can approach F with a small sample size.

To further illustrate the help of the pilot study, we conduct another simulation study. Let $x \sim \text{T}(d.f. = 30)$ and $n = 50$. By applying the R built-in function *qqnorm*, we found the data is approximately normal distributed;

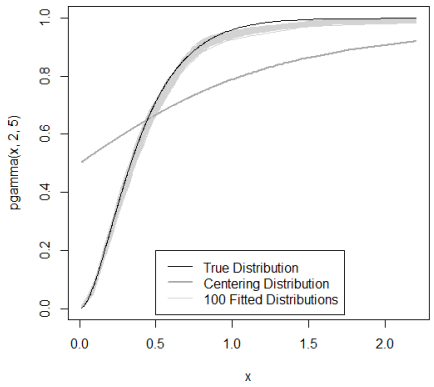

Figure 3: simulation case $1(a)$ with $n = 1000$

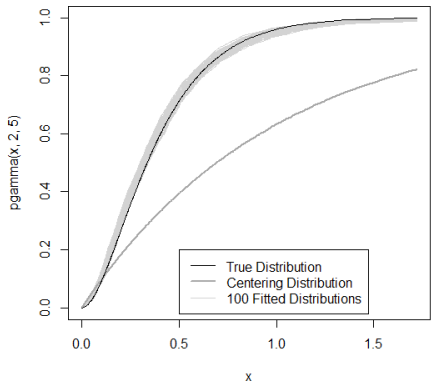

Figure 4: simulation case $1(b)$ with $n = 1000$

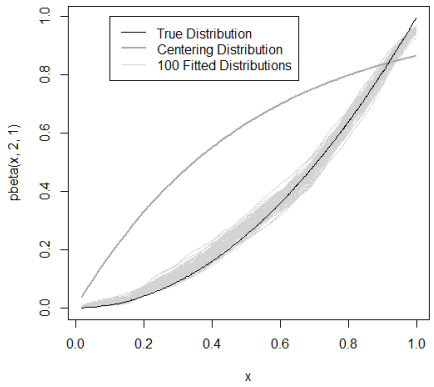

Figure 5: simulation case $2(a)$ with $n = 500$

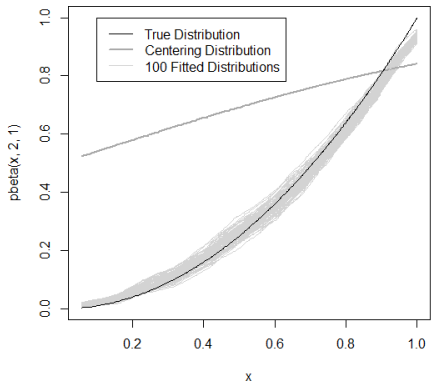

Figure 6: simulation case $2(b)$ with $n = 500$

and by calculating the sample mean and the sample variance, we found choosing a standard normal distribution as F_θ might be appropriate. We then set $F_\theta^1 = \mathrm{N}(\mathrm{mean} = 0, \mathrm{sd} = 1)$. For comparison purpose, we choose another centering distribution $F_\theta^2 = \mathrm{T}(d.f. = 1)$. Obviously, by centering at F_θ^1 (see Figure 13), the 100 fitted distribution F_Js well approach the true distribution F compared to the ones (see Figure 14) which are centered at F_θ^2. Conclusively, a good centering distribution can be a help for improving the convergence rate when other

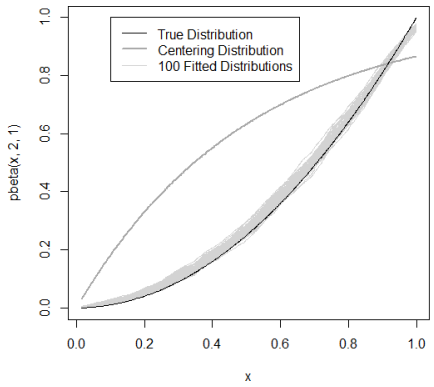

Figure 7: simulation case $2(a)$ with $n = 1000$

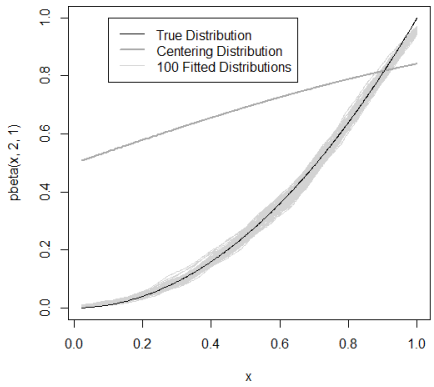

Figure 8: simulation case $2(b)$ with $n = 1000$

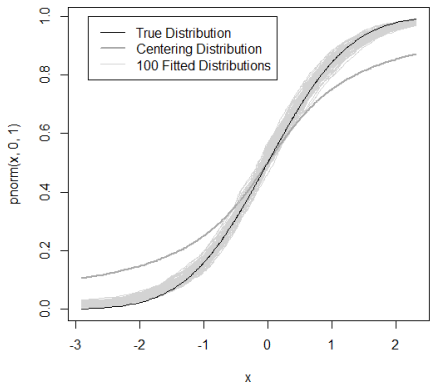

Figure 9: simulation case $3(a)$ with $n = 500$

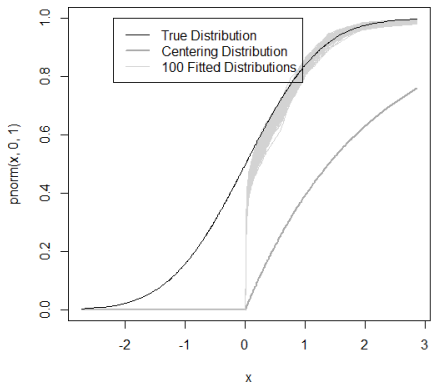

Figure 10: simulation case $3(b)$ with $n = 500$

conditions, such as the sample size, are the same.

4.2 Goodness-of-Fit Test

In this section, we will study the performance of the proposed method for a goodness-of-fit test introduced in Section 3. The data is generated from the followings with different sample sizes $n = 100$, $n = 200$, and $n = 500$. For each

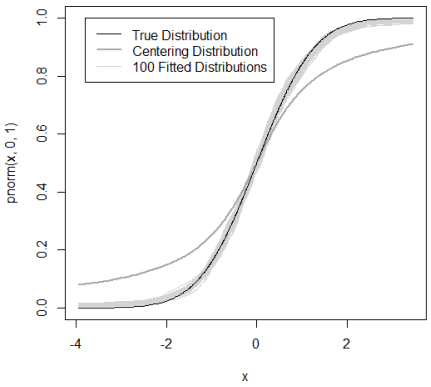

Figure 11: simulation case 3(a) with $n = 1000$

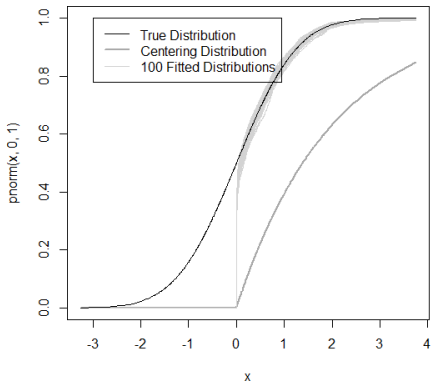

Figure 12: simulation case 3(b) with $n = 1000$

generated data, we consider different centering distributions, see below:

- (i) $x \sim \mathrm{N}(0, 1)$
  - a: $F_\theta = \mathrm{N}(0, 1)$
  - b: $F_\theta = \mathrm{N}(1, 1)$
  - c: $F_\theta = \mathrm{N}(0, 2)$
- (ii) $x \sim \mathrm{T}(1)$
  - a: $F_\theta = \mathrm{T}(1)$
  - b: $F_\theta = \mathrm{N}(0, 1)$
- (iii) $x \sim \mathrm{SN}(\mathrm{location} = 0, \mathrm{scale} = 1, \mathrm{shape} = 2)$, (Azzalini, 1985)
  - a: $F_\theta = \mathrm{SN}(\mathrm{location} = 0, \mathrm{scale} = 1, \mathrm{shape} = 2)$
  - b: $F_\theta = \mathrm{N}(0, 1)$
  - c: $F_\theta = \mathrm{T}(5)$

For each (a) scenario, the centering distribution F_θ will be chosen to be exactly equal to the true distribution F. But for others, F_θ is chosen from a set of competitive distributions. For example, in case (i)(a), the centering distribution F_θ equals to the true distribution $\mathrm{N}(0, 1)$; other two competitive centering distributions are a mean shift normal distribution, see case (i)(b), and a variance shift normal distribution, see case (i)(c).

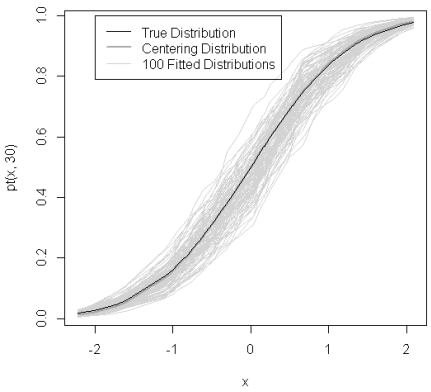

Figure 13: F_θ^1 and $n = 50$

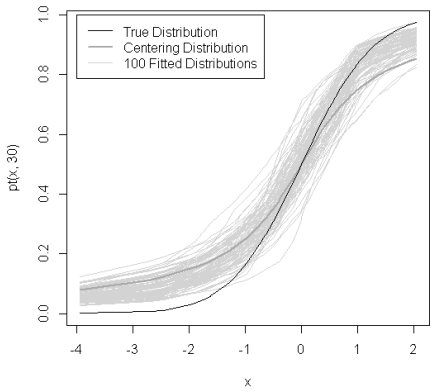

Figure 14: F_θ^2 and $n = 50$

We simulate MC = 1000 datasets for each scenario and then calculate the proportion of detecting the significant difference between F and F_θ, i.e. Power = $\sum_{i=1}^{\text{MC}} \text{I(Significant)}/\text{MC}$. Obviously, for each (a) scenario, this should be the Type I Error since this calculates the $Pr(\text{Reject } H_0|H_0)$; and for other scenarios, this is the power since it calculates $Pr(\text{Reject } H_0|H_1)$. The simulation results is listed in Table 1.

Table 1: Powers for different simulation scenarios.

	$n = 100$	$n = 200$	$n = 500$
(i)(a).	0.021	0.008	0.003
(i)(b).	0.838	0.905	0.983
(i)(c).	0.832	0.967	1.000
(ii)(a).	0.015	0.007	0.004
(ii)(b).	0.162	0.199	0.531
(iii)(a).	0.012	0.004	0.000
(iii)(b).	0.956	0.994	1.000
(iii)(c).	0.967	0.996	1.000

Overall, this proposed method works well for the above simulated scenarios. For example, in each (a) scenario, the Type I Error has a value less than the significant level 0.05 and decreases as the sample size increases as well. It might suggest, for a large sample size n, a different criteria for the truncated J should be used to increase the test power. We therefore applied $J = \lceil \log_3(n) \rceil$ for $n = 500$,

it turns out this criteria works well. As for other scenarios, power increases as the sample size n goes up and quickly approaches to 1 except for the scenario (ii)(b). We further investigate this case and find the reason behind it is that only a little information could be used on the tails. Thus if we want more power, we have to increase the sample size to help the update of the tree on the tails.

5. Galaxy Data

In this section, we study the Galaxy data set reported by Roeder (1990) using the proposed methods. The data set can be found in DPpackage (Jara et. al, 2011) in R and only contains the physical information on velocities (km/second) for 82 galaxies, denoted as $x_1^*, x_2^*, \ldots, x_n^*$ with $n = 82$. The transformed data, $x_i = \log(x_i^*/1000)$ for $i = 1, 2, \ldots, n$, is adopted for fitting the model. A pilot study shows it might be appropriate to consider $F_\theta = \mathrm{SN}(\mathrm{location} = 3, \mathrm{scale} = 0.25, \mathrm{shape} = -1)$. Let $J = \lceil \log_4(n) \rceil = 4$ and $c = 1$. The median speed is obtained by a grid searching algorithm introduced in Section 2. Repeating the fitting process for re-sampled galaxy data with replacement 1000 times and computing the median speed each time, we then obtain a sample of the median speed. The 95% confidence intervals then can be easily obtained from this sample, see Table 2 for the results.

Table 2: Median Speed and 95%CIs for the Galaxy Data

Speed (km/second)	Median	LCL	UCL
Transformed scale	3.03	3.01	3.05
Original scale	20697	20287	21115

Meanwhile, for each re-sampled galaxy data, we also conducted the goodness-of-fit test introduced in Section 3 and then calculated the proportion of rejecting H_0. We obtained this proportion is 980 out of 1000. Thus, we conclude: A non-parametric model is needed if the above skew normal distribution is specified as the centering distribution. This conclusion coincides to the one from a goodness-of-fit test for the family of skew-normal models (function *ks.sn* in package GOFSN in R (Romero, 2012)) since it has a p-value less than 0.001.

6. Conclusion

In Section 2, we proposed a new construction on Polya trees for random probability measures. Then, we constructed a multiple χ^2 test in Section 3 to examine whether or not a nonparametric model is needed when the parametric centering distribution is pre-specified. However, we found the capped level J

chosen for fitting a finite Polya tree is a critical concern since the fitting and the testing results depend on this choice. Usually, with a too small J, the model might be underfitted, however, with a too large J, the model might be prior driven and thus overfitted. Practically, we capped J at $J = \lceil \log_4(n) \rceil$, where n is the sample size. In simulation studies, we found this value works well for a small or a moderate large sample. But when the n is extremely large, the null might be incorrectly rejected. The reason behind this is that any tiny difference between the b-matrix and the m-matrix at the bottom of the tree would result in obtaining a very large test statistics and thus the null could be rejected incorrectly even if the null is held. Therefore, for an extremely large sample, one might consider to assign J a relative small value, or apply other criteria instead of $J = \lceil \log_4(n) \rceil$, or use different weighted Bonferroni corrections.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.

Berger, J. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96, 174–184.

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics*, 1, 356–358.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1, 353–355.

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

Chen, Y. and Hanson, T. (2014+). Bayesian nonparametric density estimation for doubly-truncated data. *Statistics and Its Interface*, accepted.

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

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

Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101, 1548–1565.

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

Hanson, T., Monteiro, J., and Jara, A. (2011). The Polya tree sampler: Towards efficient and automatic independent Metropolis-Hastings proposals. *Journal of Computational and Graphical Statistics*, 20, 41–62.

Jara, A., Hanson, T., Quintana, F., Muller, P., and Rosner, G. (2011). DPpackage: Bayesian non- and semi-parametric modeling in R. *Journal of Statistical Software*, 40, 1–30.

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

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

Mauldin, R., Sudderth, W. and Williams, S. (1992). Polya trees and random distributions. *The Annals of Statistics*, 20, 1203–1221.

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

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the America Statistical Association*, 85, 617–624.

Romero, V. (2012). Goodness-of-fit tests for the family of skew-normal models. *GOFSN Package in R.*

Wong, W. and Ma, L. (2010). Optional Polya tree and Bayesian inference. *The Annals of Statistics*, 38, 1433–1459.

Zhao, L., Hanson, T., and Carlin, B. (2009). Flexible spatial frailty modeling via mixtures of Polya trees. *Biometrika*, 96, 263–276.

Received April 9 2013; accepted September 24, 2013.

Yuhui Chen
Assistant Professor
Department of Mathematics
The University of Alabama, Tuscaloosa AL, USA
ychen164@ua.edu