STAT 322 Project

Anh Phan

A more ”gentle” introduction to Mixture of Polya Tree

1 Introduction

Throughout the development of Statistics, parametric statistic, with assumption of data from distribution of fixed (and often not large) set of parameters, has been a convenient way of conducting analysis, builing model, testing, etc. But one issue that it often faces is that data might based on more ”flexible” distribution, which based on a much more larger set of parameters. That is where nonparametric statistic come into play, where we assuming data come from a distribution with many (or even infinite) number of parameters, but it require a huge amount of computing power due to the high number of parameters and (possibly) more computation. At the same time, we also have the story of Frequentist and Bayesian statistic, where Frequentist method is often easier to conduct but cannot incorporate any belief of our data distribution like Bayesian statistics. With the computing power being improved day by day (we can even train models of billions of parameters now), as well as more and more methods that allow researchers to work with high-dimension distribution, it is natural that nonparametric statistic and Bayesian statistic also gain more and more popularity over time.

Among different methods in Bayesian nonparametric, the idea of Finite Polya Tree (Polya Tree) and Mixture of Finite Polya Trees (Mixture of Polya Trees) has gain interest from researchers due to its flexibility (how it can generalize from a simple normal distribution to a normal-like distribution that incorporate more parameters). And as any Bayesian statistician, prior beliefs are also incorporated in this method for a better accuracy in analyzing data. In this paper, we will look at this method for generalization of familiar distribution, with a focus on normal distribution, called Polya Tree (PT), and its extension, Mixture of Polya Tree (MPT). Due to its flexibility, it has been used various statical analysis task, but two example that we will look more into is a one sample analysis (which is the simplest case of using a distribution of analyze data), and its usage as a prior for the error term in mixed model (the class of model where we model fixed and random effect), specifically Generalized Linear Mixed Model (GLMM). The definition and MPT's application to GLMMs were based on the paper ”Parametric Nonparametric Statistics: An Introduction to Mixtures of Finite Polya Trees” by Christensen et al., and the work in the paper is extended with a quick example of one sample analysis.

2 Polya Tree

We will begin with Polya Tree (PT) (a formal definition can be found in the article by Lavine), the first extension to a parametric family of distribution. Assuming that we want to model some samples, some random variables, using a normal distribution, but this assumption can be sometimes too strict and cannot represent odd scenarios, for example, skewed distribution or distribution of more than 1 mode. To solve this problem, statisticians come up with the idea that instead of having a bell-shaped curve in the whole number line, we will split the number line into several parts, where each part will maintain the same ”normal distribution shape” but the density in that part will be scaled so that the probability that a data point to be in our interested part can be controlled by us. This would allow us to extend the benefit of working with a normal distribution to a variety of cases of different densities.

One reason why this method has "Tree" in its name is how the number line is split in when constructing a Polya Tree. In computer science and mathematics, a tree is defined as an object where we start with a parent node, and then each parent node will have some children nodes and this goes on. For Polya Tree, we begin with a normal distribution (we will consider normal distribution here, for other distribution, the process would be the same), then we split the number line into two part, called q_1 and q_2, where the probability that a data point is in part 1 will equal the probability that a data point is in part 2, or $P(X \in q_1) = 0.5 = P(X \in q_2)$ (we can see that $q_1 = (-\infty, \mu]$ and $q_2 = (\mu, +\infty)$). From here, we can choose to scale the probability that a data point is in either part, as long as $P(X \in q_1) + P(X \in q_2) = 1$. Assuming $P(X \in q_1) = \theta$ (and $P(X \in q_2) = 1 - \theta$), then in order to maintain the normal shape and the probability, we will have our p.d.f to be:

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2(I\{X \in q_1\}\theta + I\{X \in q_2\}(1-\theta))$$

with I is the indicator function. From this p.d.f, we will have:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2(I\{X \in q_1\}\theta + I\{X \in q_2\}(1-\theta))dx =$$

$$\int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2(I\{X \in q_1\}\theta + I\{X \in q_2\}(1-\theta))dx +$$

$$\int_{\mu}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2(I\{X \in q_1\}\theta + I\{X \in q_2\}(1-\theta))dx =$$

$$\int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2\theta dx + \int_{\mu}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2(1-\theta)dx =$$

$$2\theta\frac{1}{2} + 2(1-\theta)\frac{1}{2} = 1$$

so this can be a p.d.f of a probability distribution, and we also have:

$$P(x \in q_1) = P(x \le \mu) = \int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2(I\{X \in q_1\}\theta + I\{X \in q_2\}(1-\theta)) =$$

$$\int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2\theta = 2\theta(\int_{-\infty}^{\mu} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}) = 2\theta\frac{1}{2} = \theta$$

and this also leads to $P(x \in q_2) = 1-\theta$, similar to what we expect. For this p.d.f, we actually scale each part of the distribution by a certain factor such that we still have a probability distribution and the "normal" shape and property of the distribution is still maintained in each part. So we have started with a "parent node", which is the number line, and we have 2 "children node", which is each half of the number line. What we have just constructed is a 1 stage/level Polya Tree, which allow us to adjust the normal distribution based on the probability of data in each half of the support range. We can continue this process, for example, if we found that the probability of a data point in each of the 4 quantiles might be different, when we can split q_1 into q_{11}, q_{12}, which are the first two quantiles, and q_2 into q_{21}, q_{22}, which are the last two quantiles, then we choose $P(x \in q_{11} | x \in q_1) = \theta_1$ and $P(x \in q_{21} | x \in q_2) = \theta_2$ (hence $P(x \in q_{12} | x \in q_1) = 1 - \theta_1$ and $P(x \in q_{22} | x \in q_2) = 1 - \theta_2$). From here, we can see that:

$$P(x \in q_{11}) = P(x \in q_1)P(x \in q_{11} | x \in q_1) = \theta\theta_1$$

$$P(x \in q_{12}) = P(x \in q_1)P(x \in q_{12} | x \in q_1) = \theta(1 - \theta_1)$$

$$P(x \in q_{21}) = P(x \in q_2)P(x \in q_{21}|x \in q_2) = (1-\theta)\theta_2$$
$$P(x \in q_{22}) = P(x \in q_2)P(x \in q_{22}|x \in q_2) = (1-\theta)(1-\theta_2)$$

From here, in order to modify the normal distribution to fit these new change of probability of being inside a quantile, we can multiply the p.d.f of the normal distribution with:

$$4(I\{x \in q_{11}\}\theta\theta_1 + I\{x \in q_{12}\}\theta(1-\theta_1) + I\{x \in q_{21}\}(1-\theta)\theta_2 + I\{x \in q_{22}\}(1-\theta)(1-\theta_2))$$

in order to make sure the 4 claim above is true while the normal form of the distibution is maintaine. We can construct a Polya Tree of higher level based on this idea. The process will be as follow: given a range $(a_1, a_2]$ (or even $(-\infty, a)$ or $[a, \infty)$, depends on the range that you are considering), we will find a point c such that we can split the range into $(a_1, c] \cup (c, a_2]$ such that

$$P(x \in (a_1, c]|x \in (a_1, a_2]) = P(x \in [c, a_2)|x \in (a_1, a_2]) = \frac{1}{2}$$

And then we will choose $\theta \in (0, 1)$ such that $P(x \in (a_1, c]|x \in (a_1, a_2]) = \theta$, which means that: $P(x \in [c, a_2)|x \in (a_1, a_2]) = 1 - \theta$. By splitting each part that we are having into two smaller parts in each turn, if we create a Polya tree of J level, we will end up splitting the original support range of the distribution into 2^J part. At level i, or the i step, in order to create 2^i parts from 2^{i-1} parts, we need to define a conditional probability on each of the 2^{i-1} part, so at the end in order to make a Polya Tree of 2^J level, we need to define:

$$2^0 + 2^1 + ... + 2^{J-1} = 2^J - 1$$

conditional probability. If we define θ_{ik} to be the conditional probability that at level i, our data point is in the k-th part out of the 2^i part that we divided from the original support range of the distribution (for normal distribution, it is the number line). Based on this, we will have:

$$\theta_{ik} = P(x \in k^{th}\ part\ at\ step\ i|x \in \lceil\frac{k}{2}\rceil^{th}\ part\ at\ step\ i-1)$$

Where $\lceil r\rceil$ is the smallest integer larger than or equal to r. From here, we notice that the probability of a data point to be in one of the 2^J part at the end will be a multiple of J conditional probability that we defined in the construction of Polya Tree, one from each level. If we let Ω_k to be the list of those value for part k in the set of 2^J parts divided by construction of Polya Tree, then we can see that if x is in the k-th part, then the p.d.f at that point would be:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}2^n\prod_{\theta\in\Omega_k}\theta$$

Based on this, we have defined a way to extend from the normal distribution to a distribution with more parameters that can represent a wide variety of samples or random variables. When we look at the plot of the density function, as in this example from the original paper by Christensen et al., we can see that the p.d.f of $N(\mu, \sigma)$ is very smooth, while the plot of the p.d.f of 1-level PT has a drop in the middle, and the plot of the p.d.f of 2-level PT is much less smooth due to the 3 "sudden jump/drop" at the Q_1, Q_2, Q_3 of the normal distribution. This is also a property of the p.d.f of the PT to be less smooth due to the fact that we can assign different probability of being in a range of the distribution. Later on, we will look at a way to smoothen this distribution using the concept of mixture distribution.

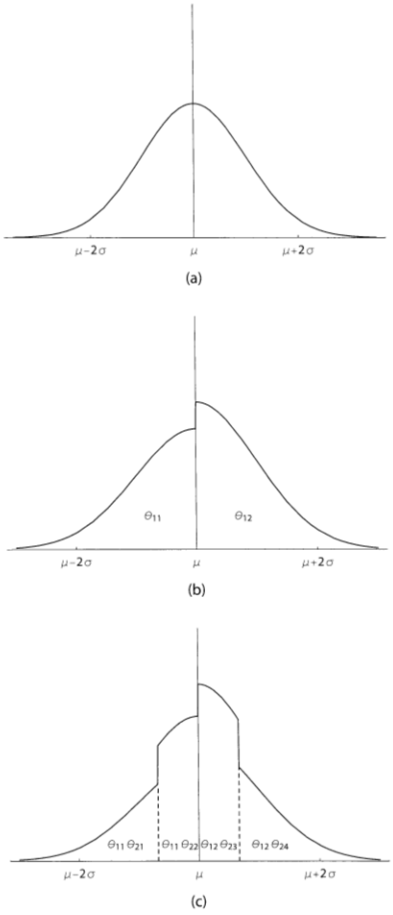


Figure 1: Plot of $N(\mu, \sigma)$, 1-level PT based on $N(\mu, \sigma)$, 2-level PT based on $N(\mu, \sigma)$

With the distribution that we just created, if we want to conduct Bayesian analysis on the distribution, especially on the parameter of the conditional probability of being in a range, we would need to define a prior distribution of these parameters (we will call these parameters as the θs now). Notice that we only need to define the distribution for θ_{ik} for k odd because we have: $\theta_{i,k+1} = \theta_{ik}$, and since these parameters represent probability value, we can use a Beta prior on these parameters (Christensen el al., 2008). This means that for $i = 1, 2, ..., J$ and for $k = 1, ..., 2^{i-1}$, we can define a prior as follow:

$$\theta_{i,2k-1} \sim Beta(\alpha_{i,2k-1}, \alpha_{i,2k})$$

for the $2k-1$-th θ value at level i of the Polya Tree. And for all the α_{ik} of the prior distribution, the article suggest choosing $\alpha_{ik} = cp(i)$, where c represent the amount of "belief" on the prior, and $p(x)$ often chosen to be x^2, so that the average of probability of being in a range (across different θ) would equal the same probability on the original distribution (Christensen et al., 2008). Based on this, we can make the Polya Tree to be a prior, with parameters c, J, μ, σ that we can use to "draw" the actual distribution of the data from it. If we reformulate the problem as the problem of 1 sample, where the data is assumed to be drawn from some distribution G with priors on the θs, then we will have:

$$X_1, X_2, ..., X_n \overset{i.i.d}{\sim} G$$

$$G \sim PT(c, J, N(\mu, \sigma))$$

where $PT(c, J, D)$ is the notation of a PT with J level, the "belief" on the beta priors to be c, and is an expansion from the original parametric distribution D

3 Mixture of Polya Tree

After we look at Polya Tree, notice that we are currently only consider our sampling distribution to have θ_{ik} to be drawn from a distribution, while μ, σ are considered to be constant. But in reality, we can also assume μ, σ to be random and conduct Bayesian analysis based on this. But before we go into any further detail, we need to look at why we have "Mixture" in this method.

We first define a finite mixture of a distribution, by first considering several distribution $f_1(x), f_2(x), ..., f_n(x)$, and we can get a linear combination of these distribution by first defining a list of weight $w_1, ..., w_n \geq 0$ where $w_1 + ... + w_n = 1$. Then we finite mixture of these distribution will be defined to be:

$$f(x) = w_1 f_1(x) + ... + w_n f_n(x)$$

Here we can look at this like the probability that we use distribution $f_k(x)$ would be w_k, and then the expression above is just the mean of those n original distribution given the probability of "picking" them,

Now we consider the infinite mixture instead, which will related directly to the mixture of Polya Tree. If we consider the p.d.f $f(x|\theta)$ of the distribution of x based on θ, and assuming θ following some distribution with p.d.f $f(\theta)$. Then based on the same idea of calculating the mean from a "set" of p.d.f as in finite mixture, if we consider at every point x, we have a random variable $g(\theta) = f(X|\theta)$ where the pdf of $g(\theta)$ is also the pdf of θ. Then we can define a mixture for $f(x|\theta)$ now to be:

$$\int f(x|\theta) f(\theta) d\theta$$

From here, we now can look at the mixture of Polya Tree (MPT). Here, we will consider a sampling distribution to be drawn from Mixture of Polya Tree prior, if we randomly draw θ_{ik} and then we calculate the mean of the "set" of p.d.f based on different μ, σ from the prior of μ, θ. If our sampling distribution is G, then we can write (Christensen et al., 2008):

$$G \sim \int \int PT(c, J, N(\mu, \sigma^2)) p(\mu) p(\sigma^2) d\mu d\sigma^2$$

From a paper by Hanson in 2006, it has been shown that the mixture of Polya Tree would be smooth, while the Polya Tree is not and also some other mixture of Polya Tree prior (not take the mean based μ, σ^2).

From here, we have had a method to extend the original distribution (normal distribution in our example) using extra parameters, while also incorporate prior belief about the original distribution. One remaining problem exist, how can we inference from the posterior distribution. It turns out that we development in MCMC methods, we can draw sample for the posterior distribution. We would need to use a combination of Gibb Sampings and Metropolis-Hastings algorithm in order to achieve our goal. In order to work with this, the article suggested writing the problem as follow:

$$X_1, X_2, ..., X_n \overset{i.i.d}{\sim} G$$

$$G \sim PT(c, J, N(\mu, \sigma))$$

$$\mu \sim p(\mu)$$

$$\sigma^2 \sim p(\sigma^2)$$

where $p(\mu), p(\sigma^2)$ is the p.d.f of μ, σ^2

From here, the Gibbs sampling process will be as follow:

- Let n_{ik} be the number of samples in the k part created at the i step of the Polya Tree, then we will draw θ_{ik} based on:

  $$Beta(n_{ik} + cp(i), n_{i,k+1} + cp(i))$$

  for $k = 1, 3, ..., 2^{i-1}$

- For μ, notice that the conditional posterior is in the form:

  $$f(X_1, ..., X_n | N(\mu, \sigma^2))p(\mu) \prod_{i=1}^{J} \prod_{k=1}^{2^i} \theta_{ik}^{n_{ik}}$$

  which is a hard distribution to draw from, so we include the Metropolis-Hasting algorithm with candidate generating distribution $N(\mu, t_\mu \frac{\sigma^2}{n})$ (Christensene et al., 2008)

- For σ^2, notice that the conditional posterior is in the form:

  $$f(X_1, ..., X_n | N(\mu, \sigma^2))p(\sigma^2) \prod_{i=1}^{J} \prod_{k=1}^{2^i} \theta_{ik}^{n_{ik}}$$

  which is a hard distribution to draw from, so we include the Metropolis-Hasting algorithm with candidate generating distribution $LN(log(\sigma^2), t_\sigma)$ (Christensen et al., 2008)

(Here, t_μ and t_σ are chosen for good MCMC mixings)

In addition to this, we can also make our choice for c, J, as described in the paper by Hanson in 2006 on inference of Mixture of Polya Tree, to fit our specific case. From here, we have constructed a new extension of parametric distribution that incorporates more parameters, as well as a process for inferencing the posterior distribution. But no methods will actually means anything until we saw its performance in real scenarios. Here, we will look at how it performs in a one-sample case and the case where it is used to model the error terms in hierarchical models.

4 PT and MPT in one sample case

In this section, we will see how the Polya Tree and Mixture of Polya Tree would allow us to model data distribution that might not follow a normal shape. In this example, we will use the famous mtcar dataset in R, and we will focus specifically on the miles per gallon data of different car (the simplified implementation can be found on Github). Note that in the next few plots for demonstration, we drew 10 distributions from the random variable of the sampling distribution. We first look at how the mpg of the car are distributed, and as a simple assumption, we can think that the miles per gallon of different car might follow a normal distribution, but it turns out that assumption does not work that well, since we have two mode of the distribution of miles per gallon. Due to this, we are hoping that a Polya Tree would allow us to extend on this distribution. And that holds true when we tried to draw some sampling distribution from a Polya Tree using the same mean and standard deviation of the previous normal distribution, but with $1, 2$ level Polya Tree with $c = 1$ now. We can see that for a 1 level Polya Tree, we seem to capture the difference in number of observations around the mean, as well as the decrease in density in the rightmost half of the mean. For a 2 level PT (the difference in histogram is caused due to change of scale in R in order to fit both plots), we can see that it could explain more in the difference between two modes in the data distribution of miles per gallon, as well as the drop between the 2 modes of the data to the point near the mean of the sample. We can see that a 2-level tree would explain more variance in shape of the data comparing to the normal distribution.

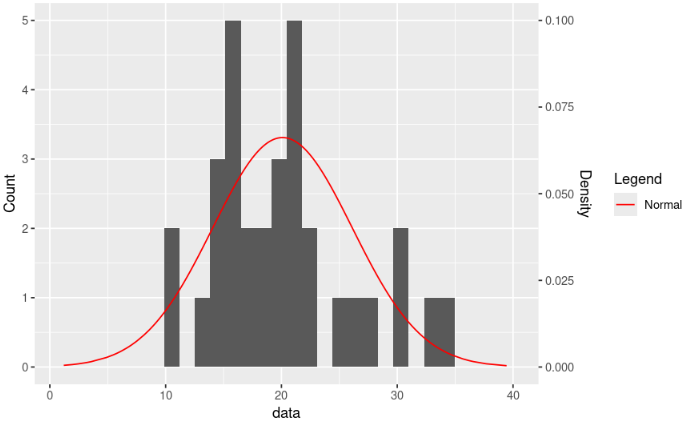


Figure 2: Miles per gallon of different car and the normal distribution

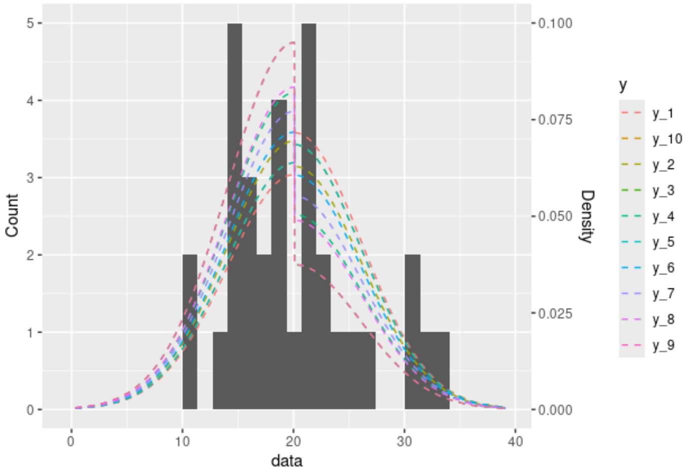


Figure 3: Miles per gallon of different car and the 1 level PT

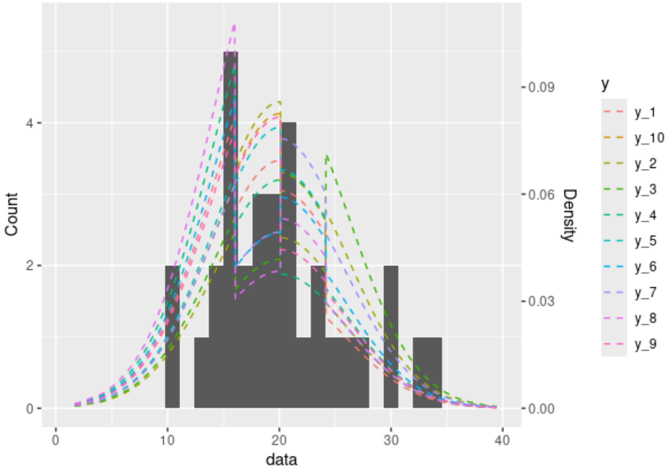


Figure 4: Miles per gallon of different car and the 2 level PT

But as we mentioned before, as well as the authors mentioned in the paper, in the 1- and 2-level Polya Tree, there were some sudden drops caused by the θ value to be different across each of the 2^J parts of the number line. In order to solve this issue, we can take the mean with respect to a prior for the parameter of the normal distribution. Here, we examine the MPT prior with μ to follow a $N(\mu_0, \sigma_0^2)$ and σ follow a $N(\sigma_0, \frac{\sigma_0^2}{10})$, as demonstrated in the example from the article by Christensen et al. (μ_0, σ_0 can be different thing, here for simplicity, we decide to make it similar to data distribution, when we examine this with less informed μ_0, σ_0, we have similar result). From here, we can see that the distribution is much more smooth now, while the property of a Polya Tree is still maintained, and different density at different parts of the data were accounted for. From here, it is clear that using a Mixture of Polya Tree prior for the sampling distribution, we can describe our data with a more smooth and more generalized distribution, allowing for more flexibility in the data. Notice here that for the two level PT, the curve can follow more closely to the actual histogram due to the fact that we take the mean over possible value of the normal distribution that it begin (we can imagine as we combine different extension of normal distribution at different center μ and different width σ^2.

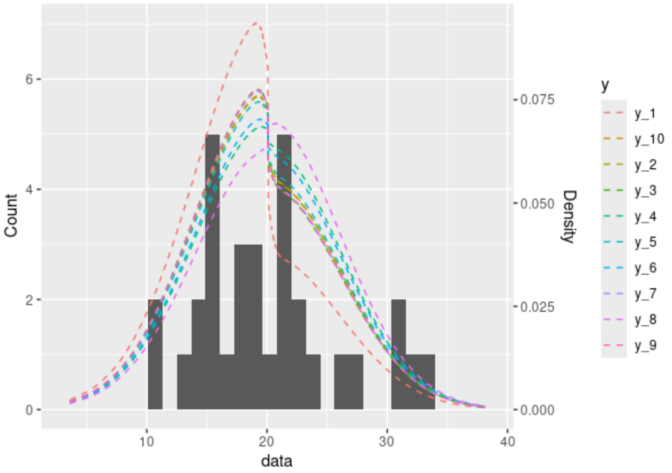


Figure 5: Miles per gallon of different car and the 1 level MPT

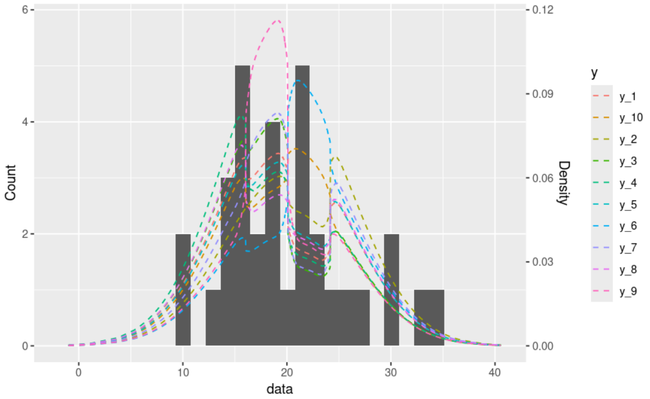
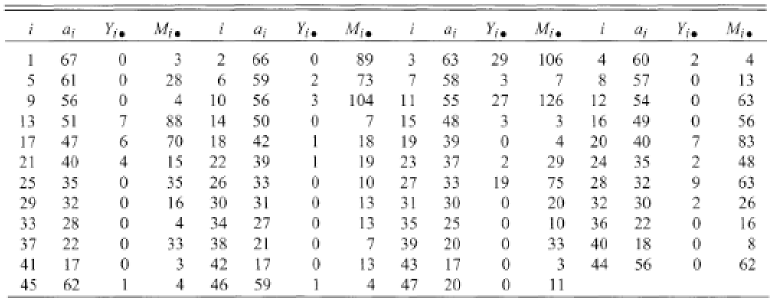


Figure 6: Miles per gallon of different car and the 2 level MPT

5 PT and MPT in Hierarchical Model case

In the original article by Christensen et al., the authors have looked at the use of MPT to model error terms in GLMM models. The first example is based on data on monkey hunters of a tribe and the result of different hunting trips of each hunters.

Table 1. Monkey hunting data from McMillan (2001). $Y_{i\bullet} = \sum_{j=1}^{N_i} Y_{ij}$ and $M_{i\bullet} = \sum_{j=1}^{N_i} M_{ij}$. The hunter's age in years is denoted a_i.



i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$	i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$	i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$	i	a_i	$Y_{i\bullet}$	$M_{i\bullet}$
1	67	0	3	2	66	0	89	3	63	29	106	4	60	2	4
5	61	0	28	6	59	2	73	7	58	3	7	8	57	0	13
9	56	0	4	10	56	3	104	11	55	27	126	12	54	0	63
13	51	7	88	14	50	0	7	15	48	3	3	16	49	0	56
17	47	6	70	18	42	1	18	19	39	0	4	20	40	7	83
21	40	4	15	22	39	1	19	23	37	2	29	24	35	2	48
25	35	0	35	26	33	0	10	27	33	19	75	28	32	9	63
29	32	0	16	30	31	0	13	31	30	0	20	32	30	2	26
33	28	0	4	34	27	0	13	35	25	0	10	36	22	0	16
37	22	0	33	38	21	0	7	39	20	0	33	40	18	0	8
41	17	0	3	42	17	0	13	43	17	0	3	44	56	0	62
45	62	1	4	46	59	1	4	47	20	0	11				

In this example, we are interested in modeling the result of hunts (Y_{ij}: result for trip j of hunter i) using a Poisson distribution with the rate is different based on hunter's age (a_i), and the time of the trip M_{ij}. The model would be as follow:

$$Y_{ij} \sim Pois(\lambda_i M_{ij})$$

$$log(\lambda_i) = \beta_0 + \beta_1(\alpha_i - 45) + \beta_2(\alpha_i - 45)^2 + \epsilon_i$$

And the authors have test the models by considering two cases: ϵ_i is drawn from a normal distribution ($\epsilon_i \sim N(\mu, \sigma^2)$) or an MPT ($\epsilon_i \sim G, G \sim PT(c, p(x) = x^2, N(\mu, \sigma^2))$), with specified normal prior for β, μ and inverse-gamma prior for σ^2. The reason for using MPT here is that instead of the errors follow a bell-shaped curve, there might be one group of "good" and one group of "bad" hunter, which will cause the distribution of error to be different. The result in the paper shown that the MPT model outperform the normal model on two main metrics: DIC (hierarchical model generalization of AIC), and LPML (Logarithm of the Pseudo Marginal Likelihood) by a little bit. The reason for this small improvement is that even though the model is better, when we plot the predictive densities of each model, we can see how there is no significance evidence of groups of "good" and "bad" hunter.

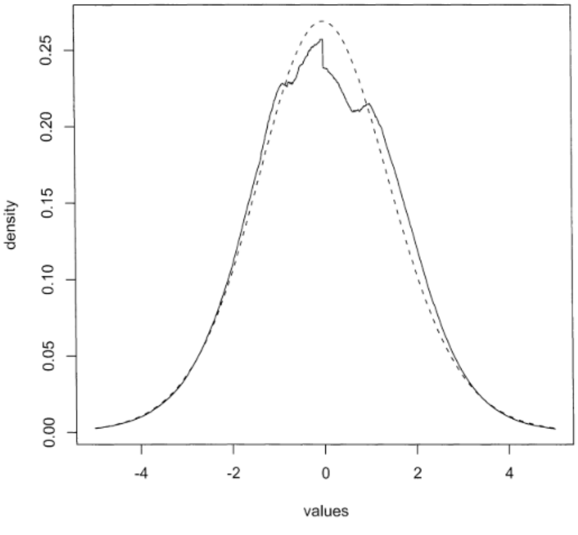


Figure 2. Monkey hunting: Predictive densities from the normal (dashed) and Polya tree (solid) models.

Another example that the author also look into is from a clinical trial related to toenail fungus. Here, there were 294 patients, each being treated with intraconazole (146 people) or terbinafine (148 people). And on weeks $0, 4, 8, 12, 24, 36, 48$, they were measured on the level of toenail fungus. Based on this, the researchers want to look at what affect the toenail fungus level. The model created on this problem was a logistic regression model on the probability that a person has moderate or severe amount of toenail fungus ($Y_{ij} = 1$). The model would be as follow:

$$logit(P(Y_{ij} = 1)) = \epsilon_i + \beta_1 TRT_i + \beta_2 Time_{ij} + \beta_3 TRT_i Timeij$$

where TRT is the indicator of what group the patient in ($TRT_i = 0$ for intraconazole and $TRT_i = 1$ otherwise), and $Time$ is the time that we are measuring the toenail fungus. Here, we will also consider both the normal model and the MPT model, with a similar form of distribution as in the previous example. But this time, the model improved much more comparing to the normal model. This can be also seen when we plot the predictive densities of each model:

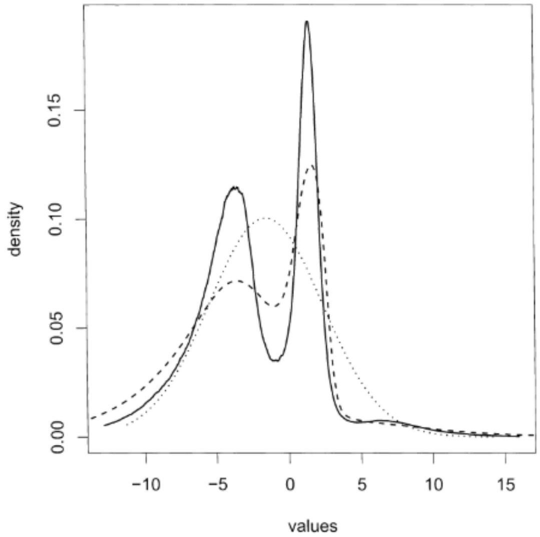


Figure 3. Toenail data estimated random effects distribution under the MPT ($c = 0.1$, solid line), MPT ($c = 1$, dashed line), and normal ($c \to \infty$, dotted line) models.

Here, we can see that for model with lower c, as mentioned in the paper, the plot clearly shows deviation from normality in such a way that the patients could be divided into two or three groups, something that the normal model would fail to do. This has shown that the Mixture of Polya Tree model could actually align with the cases where distribution might not follow a bell-shaped curve, specifically in this case it allowed us to look at different groups of patients probably based on resistance against infection and accompanying toenail separation (Christensen et al., 2008).

6 Conclusion

Overall, we have seen a new method of extending on parametric distribution to handle more nonparametric cases. With the flexibility of this type of prior, it has been used in many different cases, with one major usage being modeling the error distribution, or modeling random effect. Some interesting future research on how to work with multivariate cases (similar to Hanson's paper), or working with dependent Polya Tree across different covariates, etc. could bring great value to the landscape of nonparametric statistic and statistical modeling (or even modeling things like growth curve data, as the article suggest.)

Appendix

Christensen, Ronald, et al. “Parametric Nonparametric Statistics: An Introduction to Mixtures of Finite Polya Trees.” The American Statistician, vol. 62, no. 4, 2008, pp. 296–306, https://doi.org/10.1198/000313008X366983.

Hanson, Timothy E. “Inference for Mixtures of Finite Polya Tree Models.” Journal of the American Statistical Association, vol. 101, no. 476, 2006, pp. 1548–65, https://doi.org/10.1198/016214506000000384.

Lavine, Michael. “Some Aspects of Polya Tree Distributions for Statistical Modelling.” The Annals of Statistics, vol. 20, no. 3, 1992, pp. 1222–35, https://doi.org/10.1214/aos/1176348767.