Advanced Nonparametric Statistics and Smoothing

Efficient and robust density estimation using Bernstein type polynomials

Youhee Kil (r0768512)

Fall 2021

Contents

1 Introduction

The motivation of this report was driven by the limitations of nonparametric method such as the boundary effect and slow convergence rate. The paper "Efficient and Robust density estimation using Bernstein Type Polynomial" written by Guan (2016) proposed a maximum Bernstein likelihood density estimate method which is an approximate parametric model for a nonparametric model to estimate the underlying density by using maximum likelihood method. It is a new approach proposed by Guan (2016) to overcome some limitations of nonparametric, because this approach can inference better on the population parameters with a better underlying population density estimator.

The advantages of using the proposed maximum Bernstein likelihood density estimate are explained in details in later sections, briefly, first of all, it is efficient since it achieves a nearly parametric rate of convergence. Second, it has the robustness of nonparametric method thanks to the features of the Bernstein approximation to a continuous function. Generally, the parametric methods are much more efficient but lack of robustness, on the other hand, the Nonparametric methods are usually robust but lack of efficiency. There is a Semiparametric methods to balance between these parametric and nonparametric methods to trade-off between the efficiency and the robustness, however, miss-specification is the issue of the Semiparametric method. Third, it approximates any continuous function with compact support. Therefore, based on the listed advantages of the the proposed method, it can be used not only for estimating a density function but also be used as a general model to solve the other statistical problems.

The review of Nonparametric and empirical likelihood is preliminary to discuss about the maximum Bernstein likelihood density estimate. Owen (2001) proved the distribution function which maximizes the likelihood of X_i, $\ell(\pi_1, ..) = \sum_{s=1} n_s \log \pi_s$, is empirical distribution function :

$$\hat{F_E}(x) = \frac{1}{n} \sum I(xX_i)$$

where $I(A)$ is the indicator of A. Therefore, the empirical distribution function is often called nonparametric maximum likelihood estimator. The random variable X where one-sample nonparametric is assumed,$(x_1, x_2, x_3, ..., x_n)$, is independent observation then it will allow us to form a population, such as an unknown cumulative distribution function F or probability density function f. The continuous cumulative distribution function F can always be approximated by non-decreasing step functions, and with the same approach, the nonparametric likelihood method actually assumes that F is a non-decreasing step function with jumps only at the n observation and uses $\theta_i = dF(x_i), i = 1, ..., n$, as parameters. An approximate parametric density estimation, $\hat{h}_m$, is expected to converge to f much faster than the kernel density as one example of kernel density.

It is step behind that there is a unique $\hat{f}_m$ that maximises likelihood $\ell(f_m) = \sum_{j=1}^{n} \log f_m(x_i)$ for each m, number of dimensions. This concept is equivalent to minimizing the measure of the difference between f_m (Kullback-Leibler Divergence). Therefore, the empirical distribution which is often called as "nonparametric maximum likelihood estimator" converges to f as fast as $m \,\&\, n \longrightarrow \infty$, then we can expect that an approximate parametric density estimation ($\hat{f}_m$) converges to f at a rate faster than that of the kernel density as an example of the nonparametric density.

The paper is organised as follows. In Section 2, the Bernstein likelihood based on the special beta mixture model and the maximum Bernstein likelihood estimates (MBLEs) are introduced. The model selection for choosing the optimal degree m of the Bernstein polynomial model is described with a comparison of asymptotic results of the proposed estimates in section 3. Two different cases of application are given in section 4.

2 Methodology

Maximum Bernstein Likelihood Density Estimation

Bernstein polynomials is known as a very smooth estimator with acceptable behavior at the boundaries (Leblanc (2010)). As Bernstein (1912) defined if f is any continuous density function in the closed unit interval [0,1], it can be approximated by the Bernstein polynomial of degree $m > 0$. The equation of the Bernstein polynomial is

$$f_m(x) = B_m f(x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) p_{m,k}(x), \qquad 0 \leq x \leq 1 \tag{1}$$

where $p_{m,k} = \binom{m}{k} x^k (1-x)^{(m-k)}, k = 1, 2, ..., m$ is the Bernstein basis polynomial and B_m is the Bernstein operator. Under some conditions, the Bernstein polynomial, $B_m f(x)$, converges to $f(x)$ with its the best convergence rate. As the theorem presented by Lorentz (2013), any continuous f is defined on [0,1],

$$\lim_{m \to +\infty} B_m f(x) = f(x) \tag{2}$$

the Bernstein polynomial, $B_m f(x)$, congerges to $f(x)$ with its best convergence rate under the conditions where f is iterated Bernstein polynomials.

Moreover, Ghosal et al. (2001) explained that as mixtures of beta densities form a very flexible model for a density on the unit interval, "the class of Bernstein density is much smaller subclass of the beta mixtures defined by Bernstein polynomial, which can approximate any continuous density".

The conditions to converge with best rate are that f posses bounded second or even higher order derivatives, and we call this as iterated Bernstein polynomials in this paper. The iterated Bernstein polynomial produce the better approximation. $B_m^{(i)} f(x) = \sum_{k=0}^{m} f_{m,k}^{(i)} p_{m,k}(x)$, where its used to estimate the pdf and cdf. Besides, Guan (2016) presented that the iterated Bernstein polynomial can be written as a linear combination of the density function of the beta distribution, $beta(k+1, m+1-k)$ where $k = 0, 1, ..., m$. The equation of the linear combination the above:

$$P_{m,k}(x) \equiv (m+1) p_{m,k}(x)$$

where $p_{m,k} = \binom{m}{k} x^k (1-x)^{(m-k)}, k = 1, 2, ..., m$. Additionally, Lorentz (2013) proved that density f can be approximated with the same rate as $B_m^{(i)} f(x)$ by a Bernstein type polynomial of the form $f_B(x; \theta_m) = \sum_{k=0}^{m} \theta_{m,k} P_{m,k}$ where $\theta_m = (\theta_{m,0}, ..., \theta_{m,m})^T$, and its positive, which is called a polynomial with positive coefficients. Therefore, the density f_x can be approximately modeled and parameterised by $f_B(x; \theta_m)$ as a mixture of the beta distribution and estimate the θ_m as parameters using the maximum likelihood method.

$$\begin{aligned} \hat{f}(x; \theta_m) &= \sum_{k=0}^{m} \theta_{m,k} P_{m,k}(x) \\ &= P_{m,0} + \theta_m^T P_m(x) \end{aligned} \tag{3}$$

where $\theta_m = (\theta_{1,m}, ..., \theta_{m,m})^T, P_m(x) = \{P_{m,1}(x) - P_{m,0}(x), ..., P_{m,m}(x) - P_{m,0}(x)\}^T$

The Bernstein likelihood

The Bernstein likelihood can be defined that the maximiser the set of estimated parameters of $\ell_B(\theta_m)$ is called Maximum Bernstein Likelihood Estimation (MBLE) of θ_m. The Bernstein log-likelihood function of the set of parameters (θ) given the observed data is

$$\ell_B(\theta_m) = \sum_{i=1}^{n} \log f_B(x_j, \theta_m) \tag{4}$$

In this following section, $\hat{f}_B(x) = f_B(x; \hat{\theta}_m)$ of $f(x)$ and $\hat{F}_B(x) = F_B(x; \hat{\theta}_m)$ of $F(x)$ will be identified as the Bernstein PDF and the Bernstein CDF, respectively.

There are two proposed means of finding the maximum likelihood estimate $\hat{\theta}_m$ of θ. First, the EM algorithm. The EM algorithm can be applied to estimate θ, iteratively to find the maximum, since the Bernstein model is actually a finite mixture of m+1

completely known beta distributions. It's simple but a bit slow algorithm. Second, the Quasi-Newton method which is an algorithm method of using the gradient works pretty well for searching the maximiser of the Bernstein log-likelihood function of the set of parameters by calculating the beta density, iteratively, and it gives stable and fast result compared to EM algorithm.

3 Model Selection

degree of the Bernstein polynomial, m, for each distribution is crucial to determine the optimal value because it determines model of the Bernstein polynomial. As we mentioned as an advantage of the maximum Bernstein likelihood density estimation is only one regularization parameter for each density to choose. Such as the choice of optimal bandwidth and kernel function is difficult, even the selecting tuning parameters of the Bayesian approach is more complicated. In this case, the Bernstein polynomial model is only determined by the each positive integer m and it is called the optimal degrees m. In order to find the optimal degree m, the method of change-point is used in this paper.

The empirical likelihood allows us to add estimating equations which determine θ and side information on the top of the feasible constraints while maximizing nonparametric log likelihood to obain profile likelihood function of θ and inference can be done based on the profile likelihood function. Besides, the underlying distribution F can be estimated with the improved efficiency because of the added useful side information. The Bernstein log-likelihood, $\ell_B(\theta_m) = \sum_{i=1}^{n} \log f_B(x_j, \theta_m)$, works similarly. In order to estimate the θ_m which is the positive coefficient with optimal degree m, the equation, $\hat{\theta}_m = argmax_{\theta_m} \ell_B(\theta_m)$, can be used but it is difficult to maximize directly. Therefore, in this case, the profile log-likelihood is used to estimate θ_m by maximizing $\ell_{\theta_{m-1}}(\theta_m)$ with respect to θ_m. As the profile Bernstein likelihood $\ell(m) = \ell_B(\hat{\theta}_m)$ is always increasing as m increases. Since the Bernstein polynomial model of degree m is nested in the model of degree m+1. The maximum likelihood is increasing in m, but when m's are bigger than the optimal degree, it causes over-fitting. So, the method of change-point is applied in order to over this over-fitting problem in this paper.

The quality of the approximation improves when m increases. Since the Bernstein polynomial model of degree m is nested in the model of degree $m + 1$, for example, a model with $m = m_1$ is nested in a model with $m = m_2$, for $m_1 < m_2$. the maximum likelihood is increasing in m. In this paper, an optimal model degrees m is chosen by using of the method of change-point. The more details of the method of change-point is described with Figure 1 which is a plot for profile log-likelihood and the likelihood ratio

for change point. Firstly, the data x_j, $j = 1, 2, ..., n$ is fitted with the Bernstein model of degree m to obtain the profile log-likelihood($\ell(m)$) as we have seen in Figure 1 (left). Second, the changes of log-likelihoods are denoted as $y_i = \ell(m_i) - \ell(m_{i-1}), i = 1, 2, ..., k$ and treated $y_1, ..., , y_\tau$ as exponential with mean μ_1 and treated $y_{\tau+1}, ..., y_k$ as exponentials with mean μ_0, where τ is a change point and μ_τ is the optimal degree. A change-point estimate, $\hat{\tau}$, can be found by using the change-point detection method for exponential mode. Figure 1 (right) performs the likelihood ratio of τ, $R(\tau)$ and choose the maximum one as $\hat{\tau}$.

Asymptotic results

The best rate of convergence of the mean integrated squared error (MISE) is compared with all different models in this paper. As noted earlier, the proposed model achieves a nearly parametric rate of convergence under some conditions. In this paper, authors listed all different conditions to prove the approach converge with a nearly parametric rate. Moreover, the other constructive approximations of the functions on an interval finite or not which have fast enough convergence rate could also be used as an approximate parametric model for densities on the interval. The conversion rate of the parametric methods is $O(n^{-1})$ which is going to be the standard conversion rate to compare with other methods. Since the kernel density, the empirical distribution based Bernstein polynomial estimate and the maximum bernestin likelihood density estimate are similar in the sense that they are all mixtures of a common basis of densities. The nonparametric kernel estimate is $O(n^{-4/5})$ under the assumption f has bounded second derivative. Leblanc (2010) has proposed that a bias-reduction method using the empirical distribution-based Bernstein polynomial estimate $\tilde{f}_B(x)$ is $O(n^{-8/9})$. The maximum Bernstein likelihood density estimate $\hat{f}_B(x)$ is $O(n^{-1} \log n)$ even in the case when the original support of the data is infinite, the support is transformed into $[\alpha, \beta]$ as the finite support of F where α and β are the minimum and the maximum order statistics, respectively. More details and proof of asymptotic results of the proposed model can be checked in the paper.

4 Application

In this section, the applications of the proposed method are presented by using *mable, mixtools, DIFdetect* R packages. Applications will show how the proposed approach performs in many different cases since the paper described the proposed approach works pretty well with bounded within unit interval. First application confirms the proposed method works well with bimodal case. A bimodal data we consider the density estimation of the duration (in minutes) of eruptions of the Old Faithful based on the data set $n = 272$ eruptions. In this case, the data is truncated by interval $[a, b] = [0, 6]$ and transformed

to $y_i = x_i/b, i = 1, 2, ..., n$ When the support of the pdf f is ranged between $[a, b]$ which is not $[0, 1]$, then we can use the linearly tranformed data $y_i = (x_i - a)/(b - a)$ in [0,1] to obtain estimates $\hat{G}$ and $\hat{g}$ of the cdf G and the pdf f of $y_i's$, respectively. Then $\hat{F}(x) = \hat{G}\{(x - a)/(b - a)\}$ and $\hat{f}(x) = \hat{g}\{(x - a)/(b - a)\}/(b - a)$

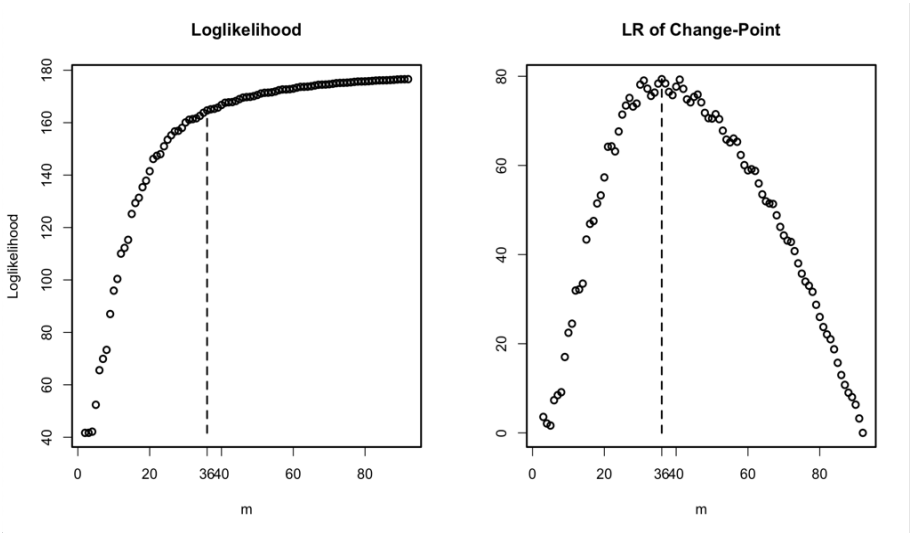

Figure 1: Profile log-likelihood (left) and the likelihood ratio for change-point (right) $[m \in M = \{2, .., 272(= n)\}]$ from the application of Old Faithful eruption duration

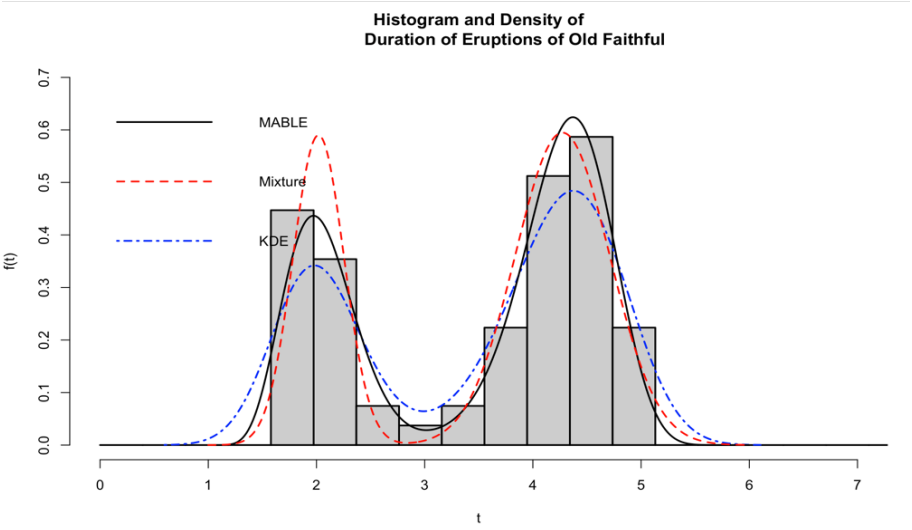

Figure 2: The density estimation with three different approaches, $\hat{f}_B, \hat{f}_p, \hat{f}_k$ as the proposed MBLE with $\hat{m} = 36$, the parametric density estimate, the kernel density estimate, respectively.

The MBLE, $\hat{g_B}$, the truncated density based on the data y_i's with the estimated optimal degree $\hat{m} = 36$ is transformed to give the MBLE $\hat{f}_B$ of f: $\hat{f}_B(x) = \hat{g}_B\{(x - 1)/5\}/5$ according to the listed equation earlier. The figure 2 shows the histogram of the bimodal data and the estimated densities using the kernel method, the parametric method from EM algorithm for nomal mixture model and the proposed method. As we can see the

proposed model fits even the bimodal data pretty well, but the parametric model from the normal mixture model doesn't fit the data perfectly.

The second application will confirm the proposed method work especially well with the boundary. A fictitious data set was selected randomly from the *DIFdetect* library, column 1 represents race. We will estimate the density of the column 22 which represent the total row sums for items. Generally, an approximate Bernstein polynomial model which is a mixture of certain beta distribution is used to fit data from a continuous population with a smooth density on finite interval. Next, the maximum Bernstein likelihood estimator of the unknown coefficients is found. In this case, the support of the density is not the unit interval. If the support of the density is not the unit interval then the transformation can be applied.

The Figure 3 represents the histogram of the data, we consider the density estimation of the total sum of the items of the fictitious data based on the $n = 405$ races. In this case, the data is truncated by interval $[a, b] = [0, 20]$ and transformed to $y_i = x_i/20, i = 1, 2, ..., n$.

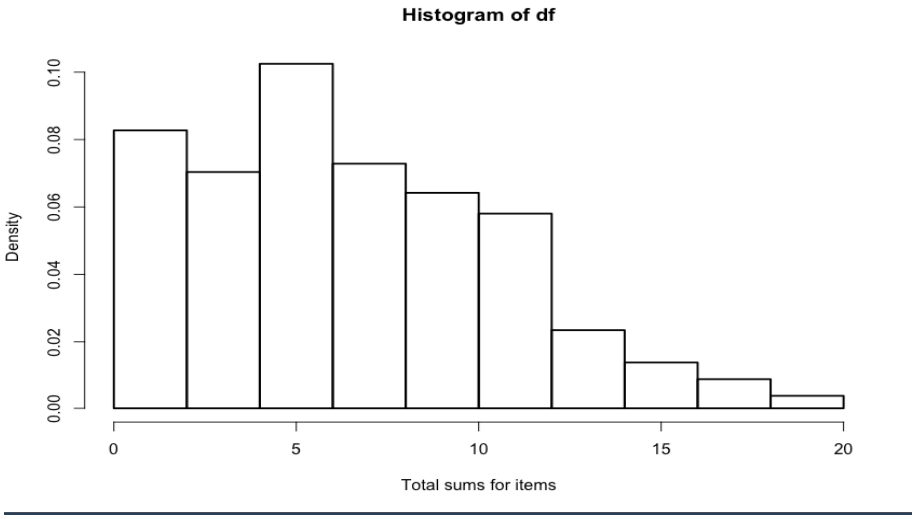

Figure 3: Histogram of total sum of items of the fictitious data

The upper bound and lower bound of m have to be found to perform the maximum Bernstein likelihood density estimation. According to the paper written by Guan (2016), there are still room for investigating the effect of the upper bound for m on the optimal degree and how to choose an appropriate upper bound. In this application, we took the writer's experience, $m_k \approx n$ was chosen for the upper bound. The writer suggested that a 1.5 bigger upper bound can be chosen if the log-likelihood changes a lot by checking the plot of log likelihood ratio y_i/i. The approximate lower bound for m is determined by mean and variance of the data, for example, $m_b = max\{1, \lceil \mu(1-\mu)/\sigma^2 - 3 \rceil\}$ where $\lceil x \rceil$ is the ceiling function of x. The MBLE $\hat{g}_B$ of the truncated density based on the

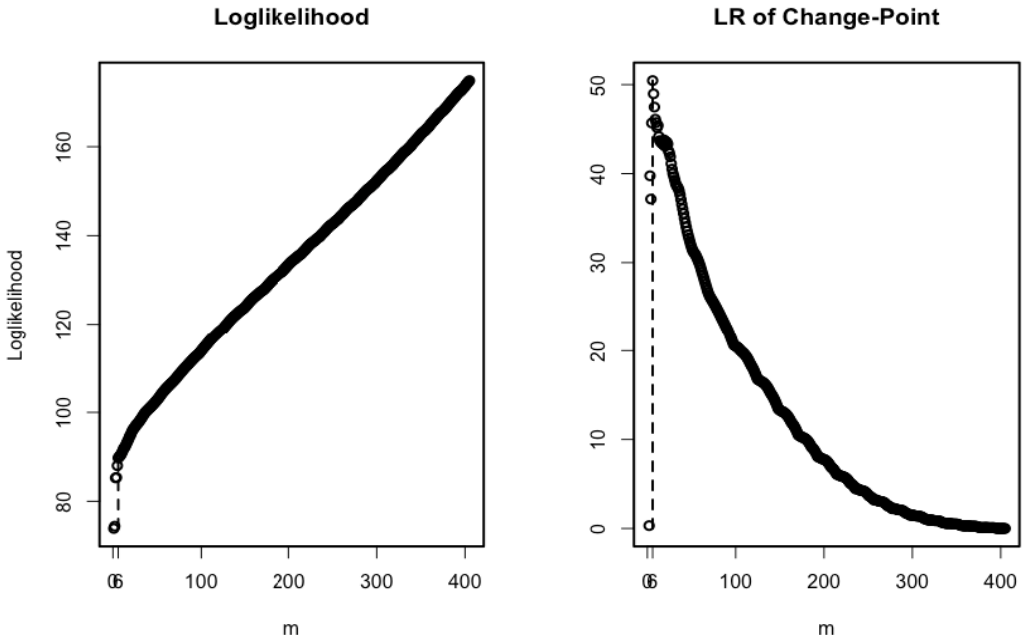

Figure 4: Profile log-likelihood (left) and the likelihood ratio for change-point (right) $[m \in M = \{1, .., 405\}]$

data y_i's with the estimated optimal degree $\hat{m} = 6$ is transformed to give the MBLE $\hat{f}_B$ of f: $\hat{f}_B(x) = \hat{g}_B\{x/20\}/20$ in this case.

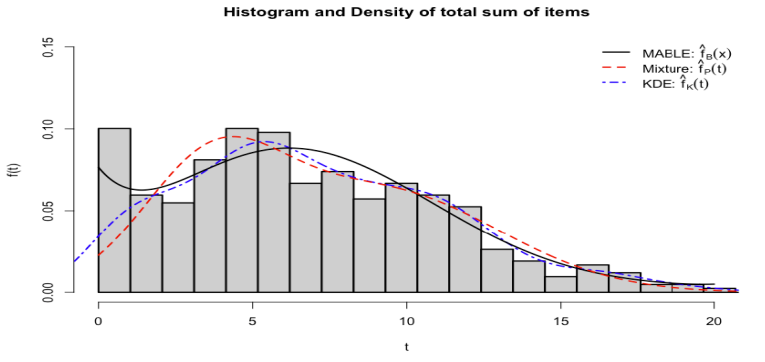

Figure 5: The density estimation with three different approaches, $\hat{f}_B, \hat{f}_p, \hat{f}_k$ as the proposed MBLE with $\hat{m} = 6$, the parametric density estimate, the kernel density estimate, respectively.

The figure 5 shows the histogram of the fictitious data and the estimated densities using the kernel method, the parametric method from EM algorithm for nomal mixture model and the proposed method. The proposed model captured the boundary of the data pretty well unlike other models, the parametric model and kernel density model, however, one thing is noticeable that the proposed method couldn't capture the exact trend between $t = 4$ to $t = 9$ in Figure 5.

5 Discussion

In this article, we proposed flexible parametric approach for nonparametric densities on the interval called maximum Bernstein Likelihood Density Estimation. The flexibility is obtained by approximating the density of the unobserved covariate by a Bernstein polynomial. The biggest advantage of this approach relies in its easiness of implementation with a single tuning parameter, which can be selected by using an change-point method, but it is a possible to obtain from an information criterion depending on characteristics of the data such as AIC or BIC. As the titled indicated, the proposed model is efficient and robust density estimation. The proposed model is efficient approach because it achieves a nearly parametric rate of convergence under some conditions which indicates it is one that produces the desired experimental results with less amount of resources compared to nonparametric models. Besides, the model shares the robustness of the nonparamtric ones due to the 'global' effectiviness of the Bernstein approximation to a continuous function. In a conclusions, it does not depend on any particular outcome model, therefore, it can suit any model or type of outcome. According to the results of simulation in the paper showed that the proposed estimator performance is quite well in many different situations with boundary unit interval, and it works particularly really good with Normal mixture model situation. In application section, normal mixture model was performed by transforming data into interval range and truncated density based on data with the optimal degree, it is confirmed that the proposed method estimates boundary of the density the best compared to other approaches. However, one thing is noticeable that there might be some case where leads to underfitting issue since the estimated model did some averaging between $t = 5$ and $t = 9$ in the second application. Moreover, the author hasn't implement the proposed method to multivariate distributions case yet. If there is a way to implement the proposed method to multivariate distribution, there would be lots of practical cases to apply. It would be interesting project along with investing to figure out how to construct confidence interval for θ based on $\hat{\theta}_B$. Because the author hasn't explained or researched yet how to construct the of confidence interval. For the future practical application of the proposed method, I would like to apply it for survival analysis which is used to investigate the time it takes for an event of interest to occur or estimating Area Under the Curve. As the proposed approach works well with unit

interval without boundary effect.

6 Appendix

```

#install.packages("remotes")
#remotes::install_github("trinker/DIFdetect")
library(DIFdetect)
library(difR)
library(mixtools)
library(ggplot2)


#Old Faithful Data
x<-faithful$eruptions
a<-1; b<-6
v<-seq(a, b,len = 512)
mu<-c(2,4.5); sig<-c(1,1)
mean<- mean(x)
sd<- sd(x)
lowerbound<- mean(1-mean)/(sd^2-3) # positive integer ->2
lowerbound
pmix<-normalmixEM(x,.5, mu, sig)
lam<-pmix$lambda; mu<-pmix$mu; sig<-pmix$sigma
y1<-lam[1]*dnorm(v,mu[1], sig[1])+lam[2]*dnorm(v, mu[2], sig[2])
res<-mable(x, M = c(2,272), interval = c(a,b), controls  =
             mable.ctrl(sig.level = 1e-8, maxit = 2000, eps = 1.0e-7))

par(mfrow = c(1,2))
plot(res, which = "likelihood")
plot(res, which = "change-point")
par(mfrow = c(1,1))
hist(x, breaks = seq(0,7.5,len = 20), xlim = c(0,7), ylim = c(0,.7),
     prob  = TRUE,xlab = "t", ylab = "f(t)", col  = "light grey",
     main = "Histogram and Density of
               Duration of Eruptions of Old Faithful")
lines(density(x, bw = "nrd0", adjust = 1), lty = 4, col = 4, lwd = 2)
plot(res, which = "density", add = TRUE)
lines(v, y1, lty = 2, col = 2, lwd = 2)
legend("topleft", lty = c(1,2,4), col = c(1,2,4), bty = "n",
       c(expression(paste("MABLE")),
         expression(paste("Mixture")),
         expression(paste("KDE"))))


#a fictitious data
data(dat)
df <- dat$tot
op<-par(mfrow = c(1,1),lwd = 2)
hist(df,xlab="Total sums for items",freq=F)
```

```
hist(df,xlab="tot",freq=F)
a<-0; b<-20
v<-seq(a, b,len = 405)
me <- 1.644*(sd(v)/sqrt(405))
mu<-c(mean(v)-me, mean(v)+me); sig<-c(1,1)
lower.m <- 1
upper.m <- 405
pmix<-normalmixEM(df,.5, mu, sig)
lam<-pmix$lambda; mu<-pmix$mu; sig<-pmix$sigma
y1<-lam[1]*dnorm(v,mu[1], sig[1])+lam[2]*dnorm(v, mu[2], sig[2])

res<-mable(df, M = c(1, 405), interval = c(a,b), controls  =
            mable.ctrl(sig.level = 1e-8, maxit = 2000, eps = 1.0e-7))
op<-par(mfrow = c(1,2),lwd = 2)
layout(rbind(c(1, 2), c(3, 3)))
plot(res, which = "likelihood")
plot(res, which = "change-point")
hist(df, breaks = seq(0, 30,len = 30), xlim = c(0,20), ylim = c(0,.15),
     prob  = TRUE,xlab = "t", ylab = "f(t)", col  = "light grey",
     main = "Histogram and Density of total sum of items")
lines(density(df, bw = "nrd0", adjust = 1), lty = 4, col = 4, lwd = 2)
plot(res, which = "density", add = TRUE)
lines(v, y1, lty = 2, col = 2, lwd = 2)
legend("topright", lty = c(1,2,4), col = c(1,2,4), lwd = 2, bty = "n",
       c(expression(paste("MABLE: ",hat(f)[B](x))),
         expression(paste("Mixture: ",hat(f)[P](t))),
         expression(paste("KDE: ",hat(f)[K](t)))))
```

References

Ghosal, S. et al. (2001). Convergence rates for density estimation with bernstein polynomials. *The Annals of Statistics*, 29(5):1264–1280.

Guan, Z. (2016). Efficient and robust density estimation using bernstein type polynomials. *Journal of Nonparametric Statistics*, 28(2):250–271.

Leblanc, A. (2010). A bias-reduced approach to density estimation using bernstein polynomials. *Journal of Nonparametric Statistics*, 22(4):459–475.

Lorentz, G. G. (2013). *Bernstein polynomials*. American Mathematical Soc.

Owen, A. B. (2001). *Empirical likelihood*. CRC press.