Spring 5-2018

Density Estimation Using Nonparametric Bayesian Methods

Yanyi Wang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics

Density Estimation Using Nonparametric Bayesian Methods

by

Yanyi Wang

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

May 2018
St. Louis, Missouri

Table of Contents

List of Figures

Figure Page

List of Tables

Acknowledgments

On the very outset of this thesis, I would like to acknowledge following important people who have help and support me during the process of writing this thesis and throughout my masters degree.

Firstly, I would like to express my sincere gratitude to my thesis advisor Prof. Todd Kuffner for his patience, enthusiasm and kindness. His timely guidance and prompt inspirations enable me to finish my thesis. It is a great honor to work with him.

Secondly, I am deeply grateful to the exceptional faculty in the Mathematical Department. They have made my time at school more positive and enjoyable. And thanks to other professors who help me acquaire a lot of knowledge during my master's degree.

Last of all, I would like to thank my friends, parents and other family members for keeping me company on long walks. They encouraged me and help me a lot.

Yanyi Wang

Washington University in St. Louis

May 2018

Dedicated to My Family.

ABSTRACT OF THE THESIS

Density Estimation Using Nonparametric Bayesian Methods

by

Wang, Yanyi

Master of Arts in Statistics,

Washington University in St. Louis, 2018.

Professor Todd Kuffner, Chair

In modern data analysis, nonparametric Bayesian methods have become increasingly popular. These methods can solve many important statistical inference problems, such as density estimation, regression and survival analysis. In this thesis, We utilize several nonparametric Bayesian methods for density estimation. In particular, we use mixtures of Dirichlet processes (MDP) and mixtures of Polya trees (MPT) priors to perform Bayesian density estimation based on simulated data. The target density is a mixture of normal distributions, which makes the estimation problem non-trivial. The performance of these methods with frequentist nonparametric kernel density estimators is assessed according to a mean-square error criterion. For the cases we consider, the nonparametric Bayesian methods outperform their frequentist counterpart.

1. Introduction

To understand the motivation for nonparametric Bayesian inference, it is helpful to first consider alternative paradigms. Frequentist methods constitute the core of classical statistics. In a frequentist approach, unknown parameters always have fixed (unknown) values. The properties of frequentist methods are typically assessed by appealing to notions of optimality from decision theory, or by studying their large-sample properties. A frequentist probability of an event is identified with a relative frequency of that event's occurrence under hypothetical repetitions of the experiment which produced the random sample. By contrast, the Bayesian notion of probability is a quantification of degrees of belief about the occurrence of events. Such a notion of probability can be applied to any unknown, and hence uncertain, component of an experiment. In particular, instead of viewing parameter values as fixed, Bayesians can assign probability distributions to the set of parameter values, which express degrees of belief about any of the possible unknown values of the parameters.

In Bayesian parametric methods, the priors and posteriors usually have a finite (often low) number of parameters, while nonparametric Bayesian models models assign a prior to infinite-dimensional parameters.

Why do we need to use the nonparametric Bayesian methods? We know that it is convenient to restrict inference to a family of distributions with a finite number of parameters, but the simplified model may lead to misleading inference if the assumed parametric family is incorrect. Therefore, we need a class of statistical methods which can flexibly adapt to unknown density or distribution functions without making restrictive parametric assumptions. Within the Bayesian paradigm, this leads us to consider nonparametric Bayesian procedures.

In this thesis, we will use several nonparametric Bayesian methods to estimate the unknown density function, f, given the observed data, $y_i \sim f(y_i)$, $i = 1, \ldots, n$.

There are many non-Bayesian methods which have been used to estimate the density function, such as histogram estimates, kernel estimates, estimates using Fourier series expansions and wavelet-based methods. Nonparametric Bayesian methods include mixtures of Dirichlet process (MDP), mixtures of Poyla tree process (MPT) and Bernstein polynomials. In this thesis, We will use MDP priors and MPT priors to perform nonparametric Bayesian density estimation.

1.1 Data

The main objective in this thesis is to compare the performances of MDP- and MPT-based nonparametric Bayesian estimation. To do so, we simulate data sets from a mixture of 3 normal densities,

$$0.2 \times N(1,1) + 0.6 \times N(3,6) + 0.2 \times N(10,2).$$

Figure 1.1 shows the true density function f, and we will compare the true density plot with the estimated density plots in Chapter 3.

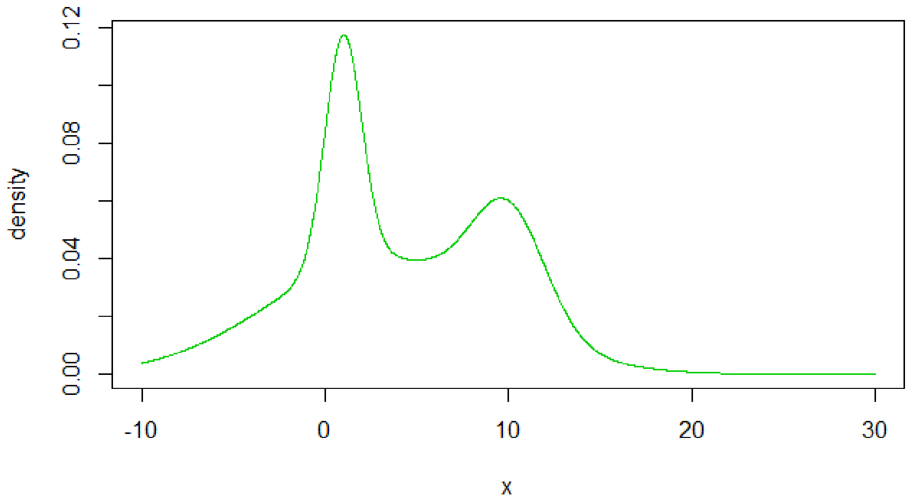

Figure 1.1. The true density function $f(y) = 0.2\times N(1, 1)+0.6\times N(3, 6)+ 0.2 \times N(10, 2)$

2. Statistical Models and Methods

Appropriate prior specification is crucial for good performance of Bayesian methods for parametric problems, and this equally true for nonparametric density estimation. When estimation is concerned with the density function, one must specify a prior distribution on the space of possible density functions. This requires that we first restrict attention to a particular set of functions. It is desired that this set is broad enough that it will contain the true density, or something very close to it. In a seminal paper on Bayesian nonparametric methods, Ferguson (1973) stipulated that prior distributions must satisfy two properties, which can be informally stated as:

(i) The support of the prior should be sufficiently large, i.e. there should be a large class of sets of densities functions with positive prior mass. By the support of the prior, we mean those elements of the parameter space (i.e. those sets of density functions) for which the prior probability is strictly positive. Any set of densities which are not in the support of the prior have prior mass of zero.

(ii) The posterior distribution should be analytically tractable. This means that it should be possible to exactly derive the posterior distribution via Bayes' rule, using only elementary calculus, and without resorting to computer-assisted approximations.

With the continued progress of Markov chain Monte Carlo (MCMC) methodology, the second requirement is now less important. The first property ensures that the prior assigns positive mass to a broad range of candidate density functions. This is desired because, *a priori*, we do not want to place restrictive assumptions on the true densities. This is because any density function not in the support of the prior will not be in the support of the posterior, either. This rather obvious observation is known as Cromwell's Rule in the Bayesian literature.

2.1 The Mixture of Dirichlet Process (MDP) Model

In order to introduce the MDP model, it is necessary to first introduce the Dirichlet process, which is the most widely-used prior in nonparametric Bayesian analysis.

2.1.1 Dirichlet Process

Let $\mathscr{X}$ be a complete, separable metric space (also known as a Polish space) and let $\mathscr{A}$ be a corresponding σ-field of subsets of $\mathscr{X}$. We define a random probability P on a measurable space $(\mathscr{X}, \mathscr{A})$ by defining the joint distribution of $(P(A_1), \ldots, P(A_k))$ for all measurable partitions $(A_1, \ldots, A_k)$. We say $(A_1, \ldots, A_k)$ is a measurable partition of $\mathscr{X}$ if $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\bigcup_{j=1}^{k} A_j = \mathscr{X}$, where $A_i, A_j \in \mathscr{A}$, for all $i, j = 1, \ldots, k$.

Definition 2.1.1 *(Ferguson, 1973) Let α be a positive real number, G_0 be a finite non-negative measure on $(\mathscr{X}, \mathscr{A})$. We say the stochastic process $P(A)$, $A \in \mathscr{A}$, is a Dirichlet process on $(\mathscr{X}, \mathscr{A})$ with parameter αG_0 if the distribution of the random vector $(P(A_1), \ldots, P(A_k))$ is Dirichlet, $\mathscr{D}(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$, where $(A_1, \ldots, A_k)$, $\forall k = 1, 2, \ldots,$ is a measurable partition of $\mathscr{A}$.*

We say the $(k-1)$-dimensional vector $(X_1, \ldots, X_{k-1})$ follows a Dirichlet distribution $\mathscr{D}(\beta_1, \ldots, \beta_k)$ if it has the density function

$$f(x_1, \ldots, x_{k-1}|\beta_1, \ldots, \beta_k) = \frac{\Gamma(\sum_{i=1}^k \beta_i)}{\prod_{i=1}^k \Gamma(\beta_i)} \left(\prod_{j=1}^{k-1} x_j^{\beta_j - 1}\right) \left(1 - \sum_{j=1}^{k-1} x_j\right)^{\beta_k - 1} I_{\mathbb{S}}(\mathbf{x}),$$

where $\beta_j > 0$ for $j = 1, \ldots, k$, $\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x} dx$, $\mathbf{x} = (x_1, \ldots, x_{k-1})$, $I_{\mathbb{S}}(\mathbf{x})$ is the indicator function for $\mathbf{x} \in \mathbb{S}$ and $\mathbb{S}$ is the simplex

$$\mathbb{S} = \left\{\mathbf{x} \in \mathbb{R}^{k-1} : x_1 + x_2 + \cdots + x_{k-1} \leq 1 \text{ and } x_j \geq 0, \text{ for } j = 1, \ldots, k-1\right\}.$$

For $k = 2$, the distribution becomes the Beta distribution, denoted by $\mathscr{B}e(\beta_1, \beta_2)$.

Since the Dirichlet process process is a discrete distribution, such a prior cannot be used to estimate continuous density functions unless we apply smoothing. Therefore, as is conventional, we utilize a mixture of Dirichlet processes as the prior on the space of density functions.

2.1.2 The MDP of Normal model

Ferguson (1983) specified the density $g(x)$ as a mixture of an infinite number of normal distributions,

$$g(x) = \sum_{i=1}^{\infty} w_i h(x|\mu_i, \sigma_i),$$

where $h(x|\mu_i, \sigma_i)$ is the density of normal distribution with mean μ_i and variance σ_i^2 and w_i denotes the weight of each normal distribution.

Let G be a probability measure on the half-plane $\{(\mu, \sigma) : \sigma > 0\}$ which can give us the mass w_i to the point (μ_i, σ_i), $i = 1, 2, \ldots$. Therefore, the previous formula can also be denoted by

$$g(x) = \int h(x|\mu, \sigma) dG(\mu, \sigma).$$

Ferguson (1983) notes that by using such mixtures of normals, density functions in the function space being considered can be estimated within a preselected accuracy in terms of the Lévy metric, and a similar result can be shown for the L_1 norm. We now define these metrics.

The Lévy metric is defined on the space $\mathcal{F}$ of cumulative distribution function (cdf) of one-dimensional random variables. The Lévy distance between two cdfs $F_1, F_2 \in \mathcal{F}$ is

$$L(F_1, F_2) = \inf\{\epsilon > 0 | F_1(y - \epsilon) - \epsilon \leq F_2(y) \leq F_1(y + \epsilon) + \epsilon \text{ for all } y \in \mathbb{R}\}.$$

The L_1 norm of the difference between function f_1 and f_2 is defined by

$$\|f_1 - f_2\|_1 = \int_R |f_1(x) - f_2(y)| dy.$$

The prior distribution for the countable infinite collection of parameters $(w_1, w_2, \ldots, \mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$, is specified as follows:

(i) $(w_1, w_2, \ldots)$ and $(\mu_1, \mu_2, \ldots, \sigma_1, \sigma_2, \ldots)$ are independent.

(ii) $w_1 = 1 - u_1$, $w_2 = u_1(1 - u_2)$, $\ldots$, $w_j = (\prod_{i=1}^{j-1} u_i)(1 - u_j)$, $\ldots$, where $u_1, u_2, \ldots$ are i.i.d. with beta distribution, $\mathscr{B}e(\alpha, 1)$.

(iii) (μ_1, σ_1), (μ_2, σ_2), $\ldots$ are i.i.d. with common gamma-normal conjugate prior. That is, $\rho_i = 1/\sigma_i^2$ follows a gamma distribution, $Gamma(a, 2/b)$, and given ρ_i, μ_i has the normal distribution, $N(\mu, \rho_i \tau)$. The parameters of the prior, α, a, b, μ and τ, are greater than zero.

The description of the prior shows that G, metioned at the begining of this section, is a Dirichlet process with parameter αG_0, and α measures how much you trust your prior 'guess'. A large value of α indicates that you place great trust in your prior guess, whereas a small α indicates a high degree of distrust in your prior guess. Such a G also follows from the Sethuraman representation (Sethuraman, 1994) of the Dirichlet process,

$$G = \sum_{i=1}^{\infty} w_i \delta_{\theta_i},$$

where θ_i are i.i.d. with distribution G_0, and w_i follows the distribution described in (ii).

We can also express the MDP model as a simple Bayes model given the likelihood $p_{\theta_i}(y_i)$ and the prior distribution G:

$$y_i \sim p_{\theta_i}(y_i), \quad i = 1, \ldots, n, \quad \theta_i | G \sim G, \quad G \sim DP(\alpha G_0).$$

Perhaps the most common variant is the MDP of normal model, due to its simplicity and analytic tractability. This model is given by

$$p_{\mu,\Sigma}(y_i) = N(y_i; \mu, \Sigma) \text{ and } G_0 \sim N(\mu | m_1, (1/k_0)\Sigma) IW(\Sigma | \nu_1, \psi_1),$$

where $IW(\Sigma|\nu_1, \psi_1)$ is the inverted Wishart distribution. Therefore, G_0 follows a conjugate normal-inverted-Wishart distribution..

2.1.3 Gibbs Sampling

We will use the `R` package `DPpackage` (Jara et al., 2018). The default MCMC sampler is a Gibbs sampler with auxiliary parameters. This is Algorithm 8 in Neal (2000).

Let G_0 be the continuous base measure in the MDP model, and let $\theta = (\theta_1, \ldots, \theta_n)$, where n is the number of observations. Let $\phi = \{\phi_1, \ldots, \phi_k\}$ be the set of distinct θ_i's, and k indicates the number of distinct values in θ. We define a new vector $c_i = (c_1, \ldots, c_n)$ as $c_i = j$ if and only if $\theta_i = \phi_j$, $i = 1, \ldots, n$, which indicates the "latent class" of observation y_i. Therefore, the distribution of $y_i|\theta_i$ can be expressed as $y_i|c_i, \phi \sim F(\phi_{c_i})$.

Also, we can rewrite the vector ϕ as $\phi = (\phi_c : c \in \{c_1, \ldots, c_n\})$ and n_j represents the number of elements in cluster j, $c_i = j$.

Next, we introduce auxiliary variables. Auxiliary variables are created and discarded within the MCMC procedure; they are used only to facilitate sampling from the posterior. The basic idea is that if we want to sample x from a distribution π_x, we can sample (x, y) from the distribution π_{xy}, which has marginal distribution for x equal to π_x. We define the permanent state variable as x, and the auxiliary variable is y. The algorithm proceeds as follows:

(i) Draw a value for $y^{(t+1)}$ from the conditional distribution of y given $x^{(t)}$.

(ii) Draw a value for $x^{(t+1)}$ from the conditional distribution of x given $y^{(t+1)}$.

(iii) Repeat (i) and (ii), the constructed chain $\left(x^{(t)}, y^{(t)}\right)$ will have a stationary distribution π_{xy}. Discard the auxiliary variable y.

The use of auxiliary variables necessitates some further notation and description of this method. For each i, let k^- indicate the number of different c_j for $j \neq i$ and label c_js with values in $1, \ldots, k^-$. Let m be the number of auxiliary variables. The conditional prior probability for c_i given $c_j, j \neq i$ can be formed as $n_{-i,c}/(n - 1 + \alpha)$ if $c_i = c \in \{1, \ldots, k^-\}$, where $n_{-i,c}$ denotes how many c_js are equal to c for $j \neq i$, and $\alpha/(n - 1 + \alpha)$ if c_i has some new value.

The Gibbs sampler proceeds according to the following steps (Neal, 2000):

(i) Repeat (ii) and (iii) for $i = 1, \ldots, n$. Then peroform step (iv).

(ii) Update c_i by drawing from the conditional distribution given the other states.

Compute k^- and $h = k^- + m$. We can sample c_i from the conditional distribution which is given above.

In summary, we can draw a new value for c_i from $\{1, \ldots, h\}$ as follows:

$$P(c_i = c | c_{-i}, y_i, \phi_1, \ldots, \phi_h) = \begin{cases} b\frac{n_{-i,c}}{n-1+\alpha} p_{\phi_c}(y_i) & \text{for } 1 \leq c \leq k^-, \\ \\ b\frac{\alpha/m}{n-1+\alpha} p_{\phi_c}(y_i) & \text{for } k^- \leq c \leq h, \end{cases}$$

where c_{-i} is a set of c_j for $j \neq i$, and b is an appropriate normalizing constant.

(iii) Sample ϕ according to the following way.

If $c_i = c_j$, for some $i \neq j$, since there is no connection between the auxiliary parameters and the rest of the states, we can simply draw the values of ϕ_c for which $k^- < c \leq h$ independently from G_0. If $c_i \neq c_j, \forall i \neq j$, we can let c_i become the first of the auxiliary components and drae the values of ϕ_c for which $k^- + 1 < c \leq h$ independently from G_0.

(iv) Sample the rest of ϕ.

For all $c \in c_1, \ldots, c_n$: sample ϕ_c from the posterior in the simple Bayesian model given by $y_i|\phi_c \sim p_{\phi_c}$ and $\phi_c \sim G_0$, for $i \in \{i : c_i = c\}$.

The use of Gibbs sampling is not restricted to conjugate priors, such as the normal-normal MDP model or conjugate normal-inverted-Wishart model. The Gibbs sampler can also be used with non-conjugate priors; for example, in the uniform-normal MDP model.

2.2 Mixtures of Polya Trees

2.2.1 Polya Tree Process

Before we define the Polya tree prior, we need to fix some notation. Define $U = \{0, 1\}$, $U^0 = \emptyset$, $U^m = U \times \cdots \times U$, which is an m-fold product, and $U^* = \bigcup_{m=0}^{\infty} U^m$. Let $\pi_0 = \Omega$, where Ω is a separable measurable space, and define $\Pi = \{\pi_m : m = 0, 1, \dots\}$ as a separating binary tree partition of Ω. Then we will have following relationships: for all $\epsilon = \epsilon_1 \dots \epsilon_m \in U^*$, $B_\epsilon \in \pi_m$, $B_{\epsilon 0}, B_{\epsilon 1} \in \pi_{m+1}$, and $B_{\epsilon 0} \cap B_{\epsilon 1} = B_\epsilon$, that is, $B_{\epsilon 0}, B_{\epsilon 1}$ split B_ϵ into two pieces. Moreover, degenerate splits are allowed, such as $B_\epsilon = B_{\epsilon 0} \cup \emptyset$. The general definition of Polya tree prior is as follows.

Definition 2.2.1 *(Lavine, 1992) We define a random probability measure $\mathscr{P}$ on Ω as a Polya tree prior $\mathscr{P} \sim PT(\Pi, \mathscr{A})$ if there exist some parameters $\mathscr{A} = \{\alpha_\epsilon : \epsilon \in U^* \text{ and } \alpha_\epsilon \geq 0\}$ and random variables $\mathscr{Y} = \{Y_\epsilon : \epsilon \in U^*\}$ that satisfy the conditions:*

(i) Y_ϵ are independent, for all $\epsilon \in U^*$;

(ii) Y_ϵ follows a Beta distribution, $Be(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$, for all $\epsilon \in U^*$;

(iii)

$$P(B_{\epsilon_1 \dots \epsilon_m}) = \prod_{j=1}^{m} \big(Y_{\epsilon_1 \dots \epsilon_{j-1}} \times I(\epsilon_j = 0) + (1 - Y_{\epsilon_1 \dots \epsilon_{j-1}}) \times I(\epsilon_j = 1)\big),$$

for every $m = 1, 2, \dots$, $\epsilon \in U^m$ and $I(A)$ is the indicator function, $I(A) = 1$ if A is true. If $j = 1$, we define $Y_{\epsilon_1 \dots \epsilon_{j-1}}$ as $Y_\emptyset$.

We can explain the Y_ϵ's mentioned above in the following way: $Y_\emptyset$ and $1 - Y_\emptyset$ are the probabilities that $\theta_i \in B_0$ and $\theta_i \in B_1$, respectively; Y_ϵ and $1 - Y_\epsilon$ are, respectively, the conditional probabilities that $\theta_i \in B_{\epsilon 0}$ and $\theta_i \in B_{\epsilon 1}$ given $\theta_i \in B_\epsilon$, for all $i = 1, 2, \dots$.

Compared to Dirichlet process priors, Polya tree priors have some advantages and disadvantages. In fact, Dirichlet processes are a special case of Polya trees if, for every $\epsilon \in E^*$, $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$ (Ferguson, 1974). This is because of the relationship between the Dirichlet distribution and Beta distribution. If $(Z_1/S, Z_2/S, \dots, Z_m/S) \sim \mathscr{D}(\beta_1, \beta_2, \dots, \beta_m)$, where $Z_i \sim Gamma(\beta_i, 1)$ and $S = \sum_{i=1}^m Z_i$, then the marginal distribution of Z_1/S follows the Beta distribution, $\mathscr{B}e(\beta_1, \sum_{i=2}^m \beta_i)$.

The advantage is that Polya tree priors can be constructed to assign probability 1 to the set of absolutely continuous random variables. The disadvantage is that binary tree partition Π plays an important role in Polya tree priors.

There is one particular class of finite Polya tree priors which is often used in the literature (Hanson, 2006). Assume there is a constraint on m, $m = 1, \ldots, M$. Let $e_m(k) = \epsilon_1 \ldots \epsilon_m$ represent the m-fold binary number representation of $k-1$, for instance, $e_4(3) = 0010$ and $e_5(7) = 00110$. At level m, define

$$B_\theta(e_m(k)) = \begin{cases} (G_\theta^{-1}((k-1)2^{-m}), G_\theta^{-1}(k2^{-m})] & \text{for } k = 1, \ldots, 2^m - 1, \\ (G_\theta^{-1}((2^m-1)2^{-m}), G_\theta^{-1}(1)) & \text{for } k = 2^m; \end{cases}$$

where G_θ is the cumulative distribution function of a continuous parametric distribution indexed by θ. Let $\Pi_\theta^m = \{B_\theta(e_m(k)) : k = 1, \ldots, 2^m\}$ partition Ω.

Definition 2.2.2 *(Hanson, 2006) Given* $\{\Pi_\theta^m\}_{m=1}^M$*, a random distribution* G *follows* $PT^M(c, \rho, G_\theta)$*,* $c > 0$*,* $\rho > 0$ *if there exist random vectors* $\mathcal{X} = \{(X_{e_m(k)0}, X_{e_m(k)1}),$ *where* $k = 1, \ldots, 2^{k-1}, m = 1, \ldots, M\}$ *satisfy:*

(i) The vectors $(X_{e_m(k)0}, X_{e_m(k)1})$ *are independent.*

(ii) $(X_{e_m(k)0}, X_{e_m(k)1})$ *follows Dirichlet distribution,* $\mathscr{D}(c\rho(m), c\rho(m))$*.*

(iii) $G\{B_\theta(\epsilon_1 \ldots \epsilon_m)\} = \prod_{k=1}^m X_{\epsilon_1 \ldots \epsilon_k}$*, for every* $B_\theta(\epsilon_1 \ldots \epsilon_m) \in \Pi_\theta^m$*.*

(iv) On sets in Π_θ^M*,* G *follows* G_θ*.*

Polya trees are conjugate and so we only need to update some parameters to obtain the posterior distribution. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the sample data from the unknown distribution. If $Y_1, \ldots, Y_n | G \overset{iid}{\sim} G$, $G \sim PT^M(c, \rho, G_\theta)$, then the posterior distribution of

G given θ and $\mathbf{Y}$ is updated through the following formula (update part (ii) in definition 2.2.2),

$$(X_{e_m(k)0}, X_{e_m(k)1})|\theta, \mathbf{Y} \sim \mathscr{D}(c\rho(m) + n_\theta(e_m(k)0; \mathbf{Y}), c\rho(m) + n_\theta(e_m(k)1; \mathbf{Y}),$$

where $n_\theta(\epsilon_1 \ldots \epsilon_m; \mathbf{Y})$ denote the number of $\{y_1, y_2, \ldots, y_n\}$ fall into $B_\theta(\epsilon_1 \ldots \epsilon_m)$.

2.2.2 Mixtures of Polya Trees

Since there are still some drawbacks of the Polya trees priors, for instance, the results can be strongly influenced by the specific sequence of partitions of the priors, and the lack of smoothness at the endpoints of the specific partitions, Lavine suggests the mixture of Polya trees processes which can allow the posterior distribution to be continuous.

Similar to the definition of the mixture of Dirichlet processes, we can index the parameter of the Polya trees processes with a random variable θ which have a parametric distribution.

The definition of the mixtures of Polya tree processes(MDP) is as follows:

Definition 2.2.3 *(Lavine, 1992) The distribution of a random probability measure $\mathscr{P}$ is a mixture of Polya trees if there exist a random variable θ, known as the index variable, with the mixing distribution H such that*

$$\mathscr{P}|\theta \sim PT(\Pi_\theta, \mathscr{A}_\theta),$$

$$\theta \sim H.$$

Or using the definition 2.2.2, we can also define the mixture of finite Polya trees (the finite MPT prior) on distribution G as

$$Y_1, \ldots, Y_n|G \overset{iid}{\sim} G, \quad G|\theta \sim PT^M(c, \rho, G_\theta), \quad \theta \sim dH(\theta).$$

Therefore, the prior on G can be written as $G \sim \int PT^M(c, \rho, G_\theta) dH(\theta)$.

2.2.3 Sampling $\mathcal{X}$

Thinking about the definition 2.2.2, Metropolis-Hastings steps are used to update the elements in $\mathcal{X}$. The concrete procedures are as follows (Hanson, 2006):

- Sample candidate $(X^*_{e_m(k)0}, X^*_{e_m(k)1})$ from the distribution $\mathscr{D}(q(X_{e_m(k)0}, X_{e_m(k)1}))$, where $q > 0$. In practice, we will set $q = 20$ or $q = 30$ for $c \geq 1$ since it works well. We can also use a more complicated choice of m, $q_m = h(m)$, where h is a decreasing function of m.

- Accept the candidate, or replace $(X_{e_m(k)0}, X_{e_m(k)1})$ by $(X^*_{e_m(k)0}, X^*_{e_m(k)1})$ with probability
$$min\left\{1, \frac{\Gamma(qX_{e_m(k)0})\Gamma(qX_{e_m(k)1}) \times A \times p(\mathbf{Y}, \beta|\mathcal{X}^*, \theta)}{\Gamma(qX^*_{e_m(k)0})\Gamma(qX^*_{e_m(k)1}) \times B \times p(\mathbf{Y}, \beta|\mathcal{X}, \theta)}\right\},$$
where
$$A = (X_{e_m(k)0})^{qX^*_{e_m(k)0} - c\rho(m)} \times (X_{e_m(k)1})^{qX^*_{e_m(k)1} - c\rho(m)},$$
$$B = (X^*_{e_m(k)0})^{qX_{e_m(k)0} - c\rho(m)} \times (X^*_{e_m(k)1})^{qX_{e_m(k)1} - c\rho(m)}.$$
And $\mathcal{X}^*$ is the set $\mathcal{X}$ with $(X_{e_m(k)0}, X_{e_m(k)1})$ replaced by $(X^*_{e_m(k)0}, X^*_{e_m(k)1})$, $\mathbf{Y}$ is the sample data set.

2.2.4 Sampling θ

In this section, we will introduce a special case called normal centering distribution, which is used in `DPpackage`. In this special case, the PT prior is centered around a normal distribution, $G_\theta = N_d(\mu, \Sigma)$, in multivariate cases. When $d = 1$, θ follows a univariate normal distribution $N(\mu, \sigma^2)$. Therefore, we can set $\theta = (\mu, \Sigma)$. And the priors we used for μ and Σ are
$$\mu \sim N(m_0, S_0), \quad \Sigma^{-1} \sim W\left(\nu_0, (\nu_0 T)^{-1}\right).$$

Note that $E(\Sigma^{-1}) = T^{-1}$ and so T can be considered as a "best guess" of Σ.

Let $p_{\mathcal{X}} = (p_{\mathcal{X}}(1), p_{\mathcal{X}}(2), \ldots, p_{\mathcal{X}}(2^M))$ denotes the vector of probability through $p_{\mathcal{X}}(k)$,

$$p_{\mathcal{X}}(k) = \prod_{j=1}^{M} X_{e_m(\lceil k2^{m-M} \rceil)},$$

where $\lceil x \rceil$ maps x to the least integer greater than or equal to x.

The method for sampling μ:

- Sample candidate μ^* from the full conditional distribution under normal

$$N\left([S_0^{-1} + n\Sigma^{-1}]^{-1}[S_0^{-1}m_0 + n\Sigma^{-1}\bar{y}], [S_0^{-1} + n\Sigma^{-1}]^{-1}\right).$$

- Accept the candidate with probability

$$min\left\{1, \prod_{i=1}^{n} \frac{p_{\mathcal{X}(K_{\mu^*,\Sigma}(M,y_i))}}{p_{\mathcal{X}(K_{\mu,\Sigma}(M,y_i))}}\right\},$$

  where $K_{\mu,\Sigma}(M, y_i)$ is a number of $k \in \{1, \ldots, 2^M\}$ such that the number (or the vector) y_i is in the set $B_{\mu,\Sigma}(e_M(k))$.

Similarly, the method for sampling Σ:

- Sample candidate Σ^{*-1} from the full conditional distribution under normal

$$W\left(n + \nu_0, \left[\nu_0 T + \sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)'\right]^{-1}\right).$$

- Accept the candidate with probability

$$min\left\{1, \prod_{i=1}^{n} \frac{p_{\mathcal{X}(K_{\mu,\Sigma^*}(M,y_i))}}{p_{\mathcal{X}(K_{\mu,\Sigma}(M,y_i))}}\right\}.$$

Hanson (2006) also provided an alternative method when the previous approaches give us an inaccurate results. When the data are very nonnormal, the acceptance probabilities are very small and the MCMC method will get stuck.

3. Data Analysis

3.1 Data Simulation

At first, we need to simulate the data from the true density function, $0.2 \times N(1,1) + 0.6 \times N(3,6) + 0.2 \times N(10,2)$. The size of data is n. And we will set n to be 50, 100, 200 and 1000 in this chapter. The simulation method is as follows (Chapter 3.3 of Jara et al. (2011)):

- for $i = 1, \ldots, n$, sample u from uniform distribution $U(0,1)$.
- $$data[i] \sim \begin{cases} N(1,1) & \text{if } 0 < u < 0.2, \\ N(3,6) & \text{if } 0.2 \leq u < 0.8 \\ N(10,2) & \text{if } 0.8 \leq u \leq 1 \end{cases}$$

3.2 Results Using MDP of Normal Model

Since we will use the `R` package, `DPpackage` (Jara et al., 2018), and the `R` function, `DPdensity` in this thesis, we should again state the independent hyperpriors which are used in this method and give the value of parameters in the priors.

$$y_i|\mu_i, \Sigma_i \sim N(\mu_i, \Sigma_i), i = 1, \ldots, n,$$

$$(\mu_i, \Sigma_i)|G \sim G,$$

$$G|\alpha, G_0 \sim DP(\alpha G_0);$$

we use the conjugate normal-inverted-Wishart:

$$G_0 \sim N(\mu|m_1, (1/k_0)\Sigma) IW(\Sigma|\nu_1, \psi_1);$$

We will use two priors with different value of parameters: 1) Fixed α, m_1 and ψ_1. 2) Randomize all the parameters.

MDP Prior 1: We set the value of paprameters as $\alpha = 1$, $m_1 = 1$, $\nu_1 = 4$ and $\psi_1^{-1} = 0.5$. And draw k_0 from the Gamma distribution. That is,

$$G_0 \sim N(\mu|1, (1/k_0)\Sigma) IW(\Sigma|4, 2),$$

$$k_0 \sim Gamma(0.5, 50)$$

MDP Prior 2: We set the value of paprameters as $\nu_1 = 4$, and randomize the rest of the parameters from the following distributions:

$$\alpha \sim Gamma(2, 1),$$

$$m_1 \sim N(0, 100000),$$

$$k_0 \sim Gamma(0.5, 50),$$

$$\psi_1 \sim IW(4, 0.5).$$

We can have a look at the simulated parameters (k_0 and the number of clusters) for MDP prior 1 when the sample size $n = 200$, see figure 3.1.The figure shows that we have 5 clusters and the posterior mean of k_0 is 0.0222. We can also see that MCMC is converge.

What's more, we can get the figure of predictive information about the means and covariance (see figure 3.2).

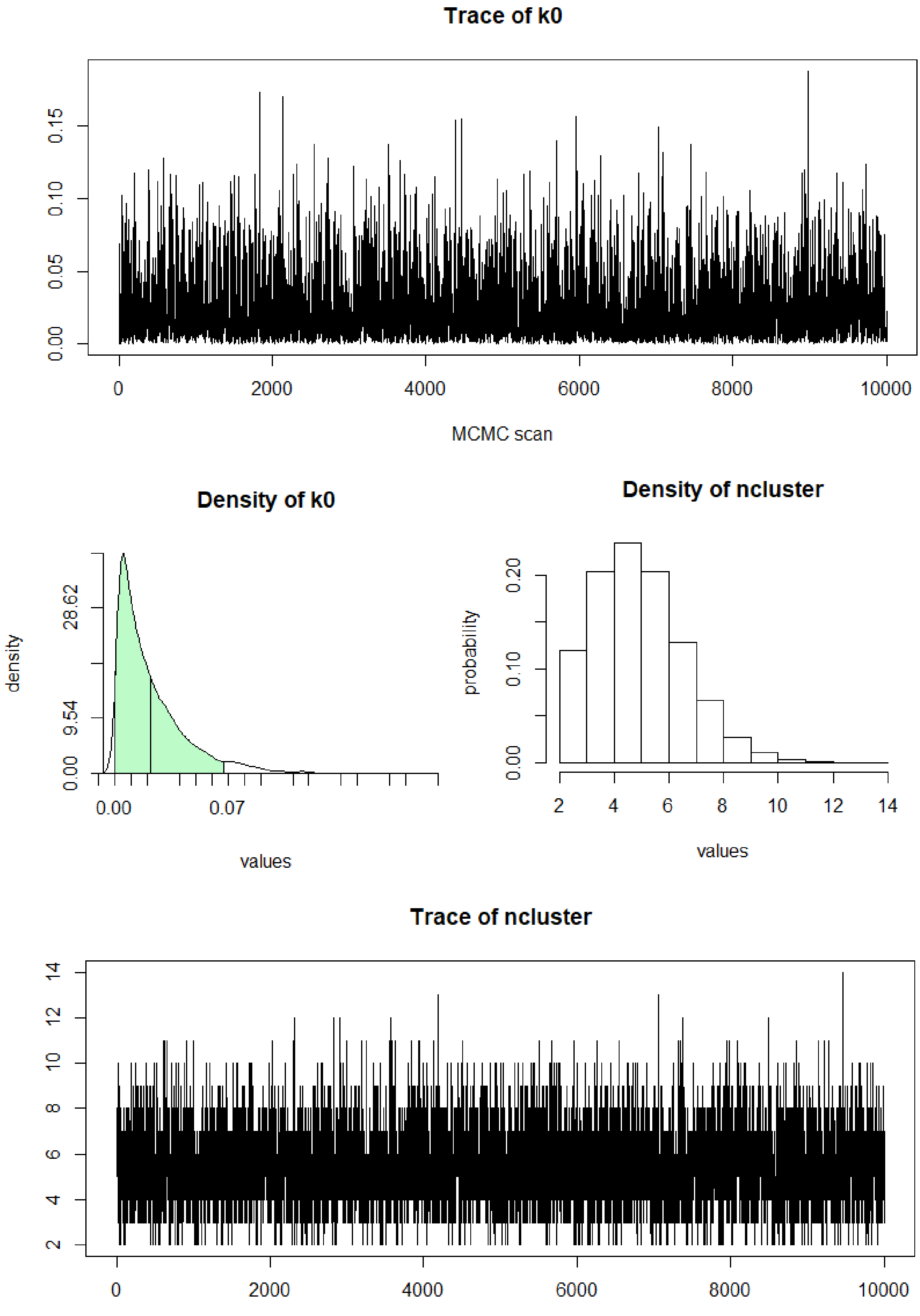

Figure 3.1. The simulated parameters we used in MDP prior 2 when $n = 200$.

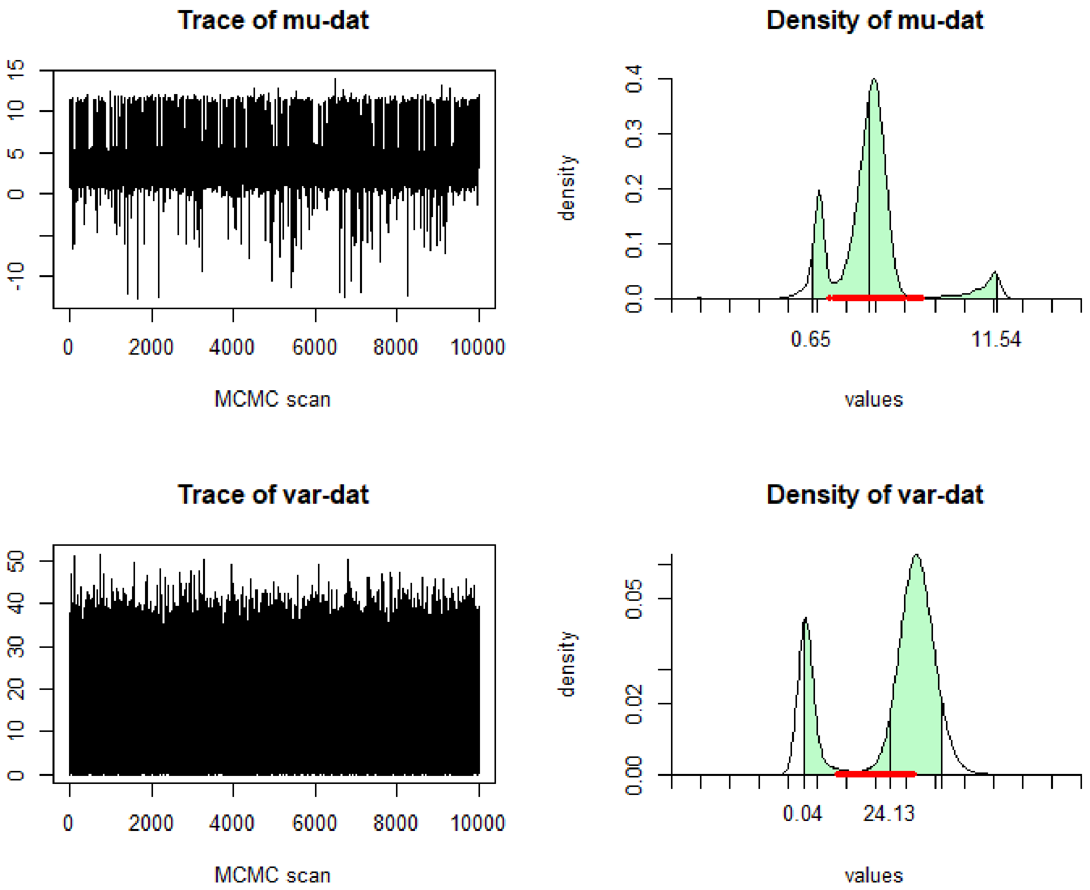

Figure 3.2. The prediction information about mean and covariance of MDP prior 1 when $n = 200$.

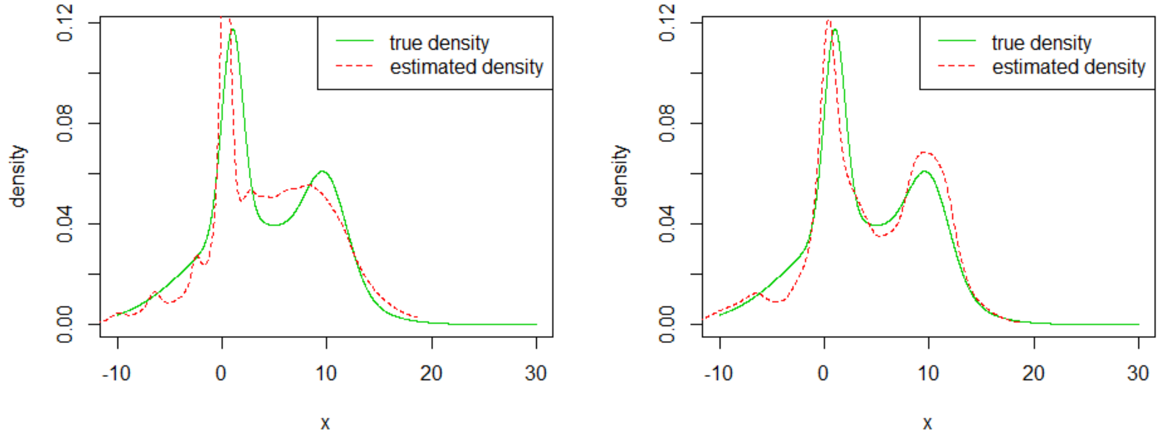

Figure 3.3. True density and estimated density of MDP priors when $n = 50$, the left figure shows the density estimation of MDP prior 1 and the right figure is of MDP prior 2.

3.2.1 True Density vs. Estimated Density

Since we want to evaluate the performance of different priors with different sample sizes, we plot the true density and the estimated density in the same figure. Figure 3.3 shows the estimated density of two different priors when $n = 50$, figure 3.4 shows the estimated density when $n = 100$, figure 3.5 shows the estimated density when $n = 200$ and figure 3.6 shows the estimated density when $n = 1000$.

3.3 Results Using MPT Model

We again state the MPT model as follows:

$$Y_1, \ldots, Y_n | G \sim G,$$

$$G | \alpha, \mu, \sigma \sim PT^M(\Pi^{\mu,\sigma^2}, A);$$

We also use two different priors in this thesis: 1) Fixed σ; 2) Randomize all parameters.

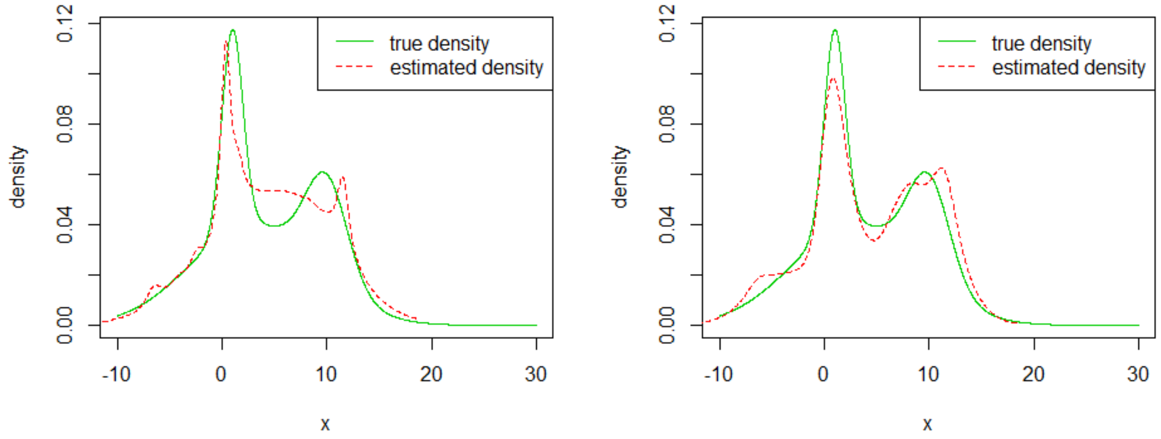

Figure 3.4. True density and estimated density of MDP priors when $n = 100$, the left figure shows the density estimation of MDP prior 1 and the right figure is of MDP prior 2.

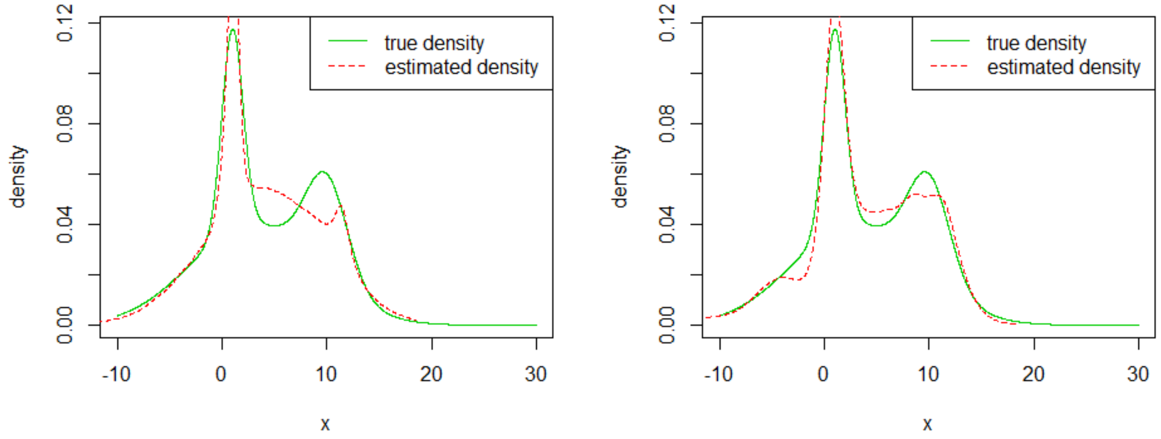

Figure 3.5. True density and estimated density of MDP priors when $n = 200$, the left figure shows the density estimation of MDP prior 1 and the right figure is of MDP prior 2.

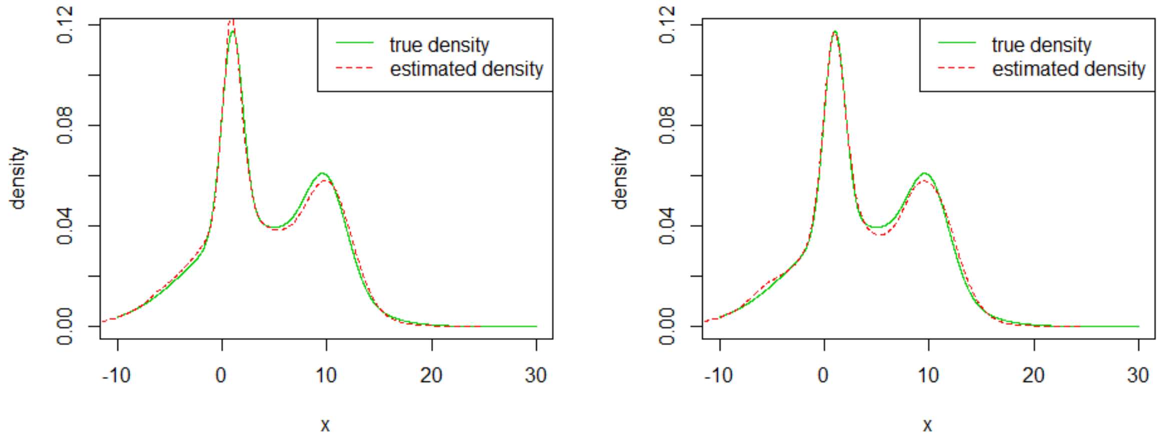

Figure 3.6. True density and estimated density of MDP priors when $n = 1000$, the left figure shows the density estimation of MDP prior 1 and the right figure is of MDP prior 2.

MPT Prior 1: Let $\sigma = 20$ and $M = 6$, and we define the rest parameters as following distributions.

$$\mu \sim N(21, 100),$$

$$\alpha \sim Gamma(1, 0.01).$$

MPT Prior 2: The distribution of μ and α is the same as which in MPT prior 1, and the distribution of σ is

$$\sigma^{-2} \sim Gamma(0.5, 50).$$

We use `R` function, `PTdensity` in this section.

3.3.1 True Density vs. Estimated Density

The same as the section 3.2.1, we plot the true density and the estimated density of different MPT priors with various sample sizes. Figure 3.7 shows the density of prior 1 and prior 2 when $n = 50$, figure 3.8 shows the density when $n = 100$, figure 3.9 shows the density when $n = 200$ and figure 3.10 shows the density when $n = 1000$.

3.4 Model Comparison

We also use frequentist nonparametric method, Gaussian kermel method, to estimate the density function. The kernel density estimator is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where $h > 0$ is the bandwidth and K is the kernel. We use the standard normal density function as the kernel in this method.

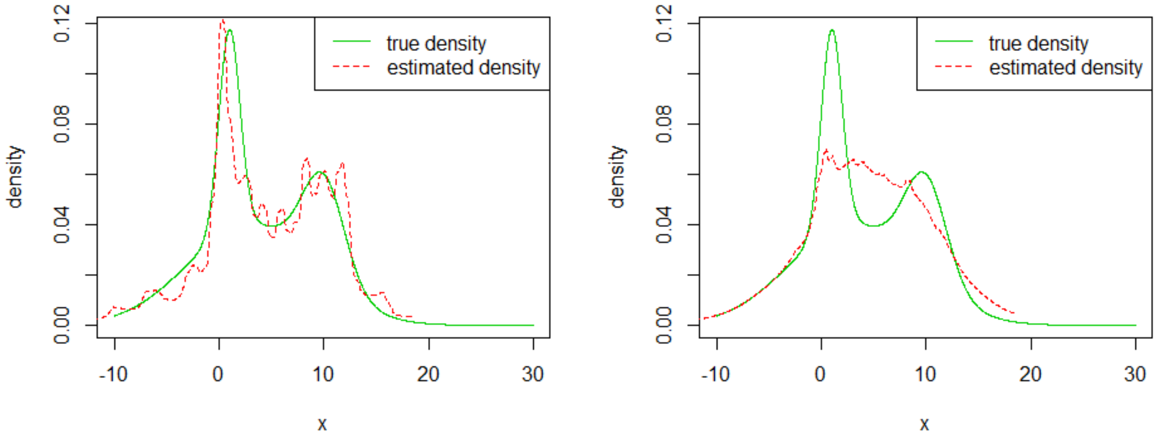

Figure 3.7. True density and estimated density of MPT priors when $n = 50$, the left figure shows the density estimation of MPT prior 1 and the right figure is of MPT prior 2.

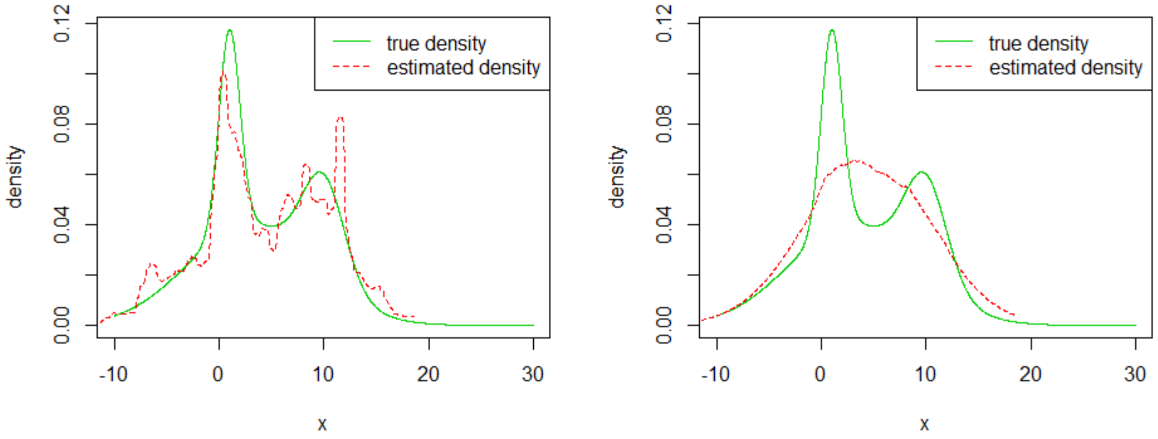

Figure 3.8. True density and estimated density of MPT priors when $n = 100$, the left figure shows the density estimation of MPT prior 1 and the right figure is of MPT prior 2.

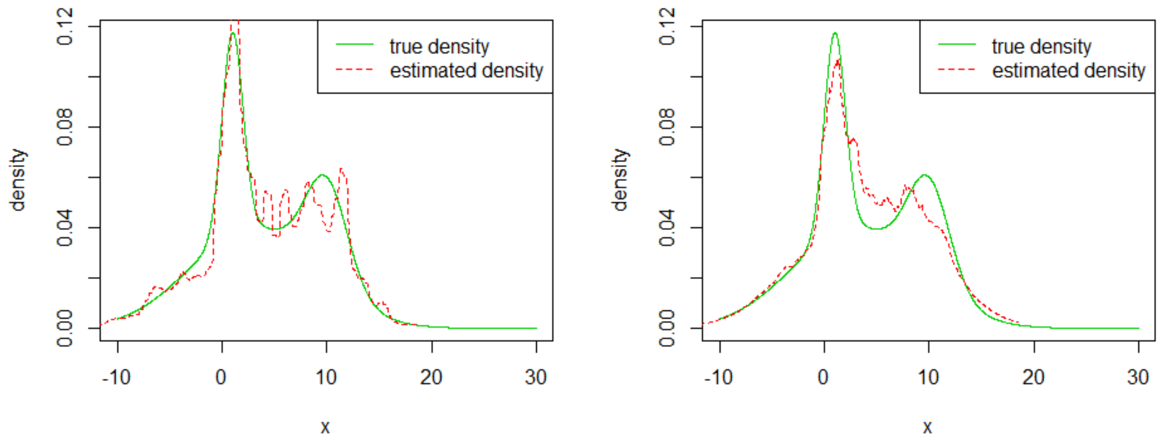

Figure 3.9. True density and estimated density of MPT priors when $n = 200$, the left figure shows the density estimation of MPT prior 1 and the right figure is of MPT prior 2.

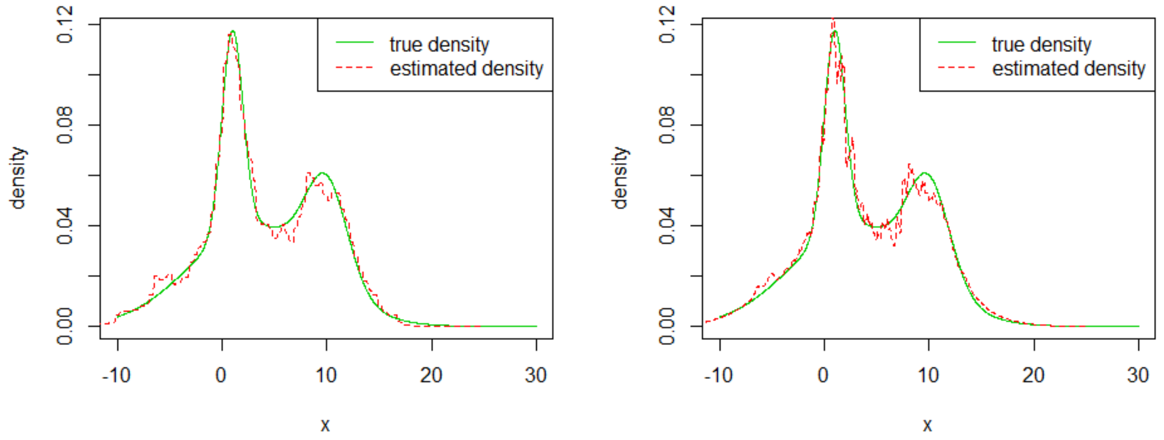

Figure 3.10. True density and estimated density of MPT priors when $n = 1000$, the left figure shows the density estimation of MPT prior 1 and the right figure is of MPT prior 2.

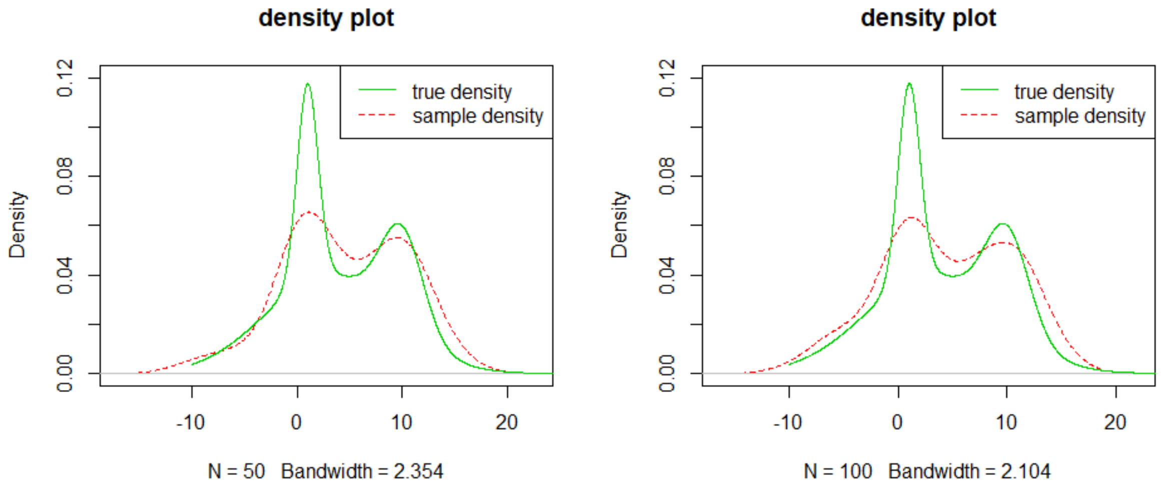

Figure 3.11. True density and estimated density using Gaussian kernel method when $n = 50$ (left) and $n = 100$ (right).

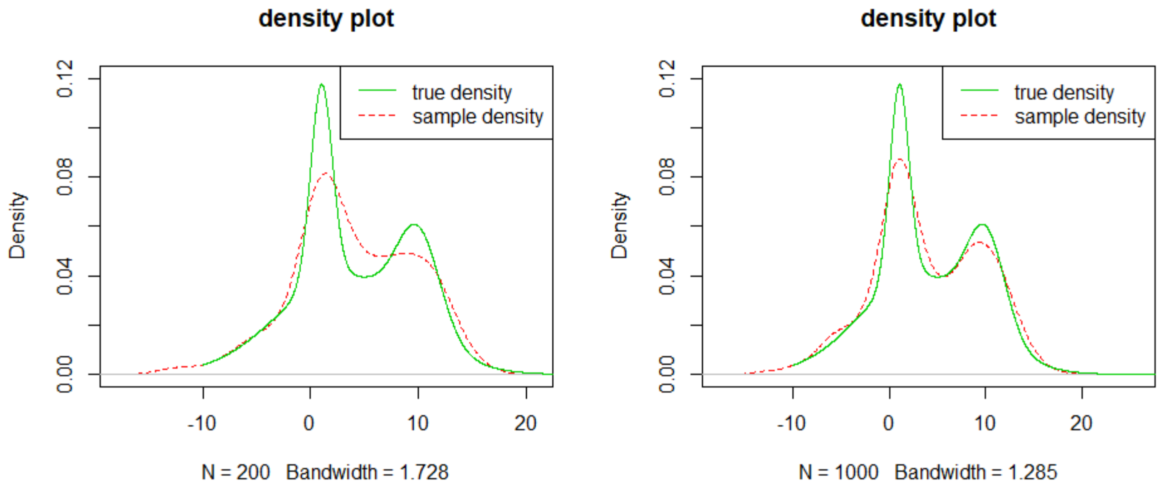

Figure 3.12. True density and estimated density using Gaussian kernel method when $n = 200$ (left) and $n = 1000$ (right).

Figure 3.11 and figure 3.12 shows the true density and the estimated density using Gaussian kernel method. From left to right, the sample size is $n = 50$, $n = 100$, $n = 200$ and $n = 1000$, respectively. We use `R` function `density` to implement this method.

Table 3.1
Mean-Square Error of Five Different Methods with Different Sample Sizes

Methods	**n = 50**	**n = 100**	**n = 200**	**n = 1000**
Gaussian Kernel Method	1.58×10^{-4}	1.01×10^{-4}	9.81×10^{-5}	5.48×10^{-5}
MDP Prior 1	2.71×10^{-4}	1.01×10^{-4}	8.37×10^{-5}	4.78×10^{-6}
MDP Prior 2	7.51×10^{-5}	4.81×10^{-5}	2.46×10^{-5}	3.43×10^{-6}
MPT Prior 1	1.09×10^{-4}	1.24×10^{-4}	5.87×10^{-5}	1.85×10^{-5}
MPT Prior 2	2.68×10^{-4}	2.40×10^{-4}	5.35×10^{-5}	2.14×10^{-5}

To have a look at the performance in a more mathematical way, we calculate the estimate of mean-square error. The mean-square error of an estimator $\hat{f}(x_i)$ is defined as

$$MSE = E_{\hat{f}}\left[\Big(f(x_i) - \hat{f}(x_i)\Big)^2\right].$$

We use the sample average to calculate the estimate of mean squared error.

$$\widehat{MSE} = \frac{1}{N}\sum_{i=1}^{N}\Big(f(x_i) - \hat{f}(x_i)\Big)^2,$$

where $f(x)$ is the true density and $\hat{f}(x)$ is the estimated density at point x_i. $N = 1000$ and $x_i, i = 1, \ldots, N$ is a sequences on $[-10, 20]$ with fixed increment and the length of the sequence is 1000.

Table 3.1 shows the mean-square error of Gaussian kernel method, two MDP methods and two MPT methods with different sample sizes, $n = 50, 100, 200, 1000$.

4. Conclusion

In this thesis, we use mixtures of Dirichlet process (MDP) and mixtures of Polya trees priors (MPT) to perform Bayesian density estimation based on simulated data with different sizes. The data is simulated from a mixture of normal distribution. Moreover, to compare the performance between Bayesian methods and frequentist methods, we also use Gaussian kernel method.

According to the figures of density plot using five different methods, we can see that MDP methods perform better than other methods. And the estimated density plots of MPT methods are less smoother than the density plots of MDP methods.

For MDP methods, prior 2 (randomize all the parameters) performs better than prior 1 (fix α, m_1 and ψ_1). When the sample size n is large enough ($n = 1000$), the estimated density plot is almost the same as the true density plot. As for MPT methods, prior 1 (fixed σ) is smoother than prior 2 (randomize all the parameters) since the number of levels of the finite polya tree in prior 2 ($M = 8$) is larger than the levels ($M = 6$) in prior 1.

Looking at the table 3.1, we have a more mathematical conclusion. We can conclude that MDP method using prior 1 is the best estimation method in these five methods. Compare nonparametric frequentist method to nonparametric Bayesian methods, we can see that they all perform well when n is large. And in most cases, nonparametric Bayesian outperform their frequentist counterpart.

Bibliography

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Thomas S. Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, pages 615–629, 1974.

Thomas S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Elsevier, 1983.

Timothy E. Hanson. Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(476):1548–1565, 2006.

Alejandro Jara, Timothy E. Hanson, Fernando A. Quintana, Peter Müller, and Gary L. Rosner. DPpackage: Bayesian semi-and nonparametric modeling in `R`. *Journal of Statistical Software*, 40(5):1, 2011.

Alejandro Jara, Timothy Hanson, Fernando Quintana, Peter Mueller, and Gary Rosner. Package 'DPpackage', 2018.

Michael Lavine. Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, pages 1222–1235, 1992.

Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.