

DENSITY IDENTIFICATION OF DISTRIBUTIONS ON  
BOUNDED DOMAIN

An abstract of

a Thesis

Presented to the faculty of  
the Department of Mathematics  
Western Illinois University

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science

By

INDRAJITH WASALA MUDIYANSELAGE

December 2018

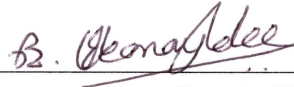
## ABSTRACT

In parametric density estimation methods, we assume that we know the shape of the distribution, but not the parameters. We estimate the parameters using a method such as maximum likelihood estimation. One may choose the parametric family by pre-modeling or a model selection criterion. If the distribution has an unknown truncation, then the parameter estimations may fail to produce correct answers. Also, most of the existing parametric methods are computationally demanding for large data sets. In this thesis, I propose a new parametric density estimation method when data are drawn from a distribution with unknown truncation. I introduce two density estimation approaches based on the estimated derivative of the empirical distribution. These new approaches also improve the efficiency of the density estimation by both eliminating the need of model selection procedures and possibly reducing the complexity of the necessary algorithms. The first method introduced in this thesis is to estimate densities using a polynomial approximation of the empirical distribution function. The second method is a mode-based density identification using a four-parameter family of functions. The principal assumption of the second method is that the mode is contained in the data range. I demonstrate the applications of the methods for ecological data sets, where many data ranges are restricted by geographical or ecological constraints.

*Keywords:* Density estimation, Parametric density, Truncated distribution, Empirical distribution, Ecological constraints.

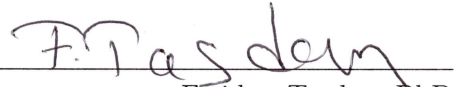
## APPROVAL PAGE

This thesis by Indrajith Wasala Mudiyanse is accepted in its present form by the Department of Mathematics of Western Illinois University as satisfying the thesis requirements for the degree Master of Science.



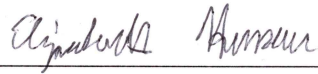
Dinesh Ekanayake, PhD

Chairperson, Examining Committee



Feridun Tasdan, PhD

Member, Examining Committee



Beth Hansen, PhD

Member, Examining Committee

---

11/30/2018



DENSITY IDENTIFICATION OF DISTRIBUTIONS ON  
BOUNDED DOMAIN

A Thesis

Presented to the faculty of  
the Department of Mathematics  
Western Illinois University

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science

By

INDRAJITH WASALA MUDIYANSELAGE

December 2018

ProQuest Number: 13421586

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 13421586

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346

## Acknowledgments

First, I wish to extend my sincere gratitude and deep appreciation to my thesis advisor Dr. Dinesh Ekanayake, for his tireless efforts and unfailing encouragements. This thesis would not nearly be what it is without his guidance and suggestions. I would also like to thank the other committee members, Dr. Feridun Tasdan and Dr. Beth Hansen, for reviewing my work and their helpful suggestions. I thank the WIU math department for the excellent teaching and learning opportunities they have provided me over the years. Finally, but not least I express my love and gratitude to my beloved wife, for her understanding and continued support through the duration of my studies.

# TABLE OF CONTENTS

Acknowledgments . . . . .	ii
List of figures . . . . .	iv
Chapter	
I Introduction . . . . .	1
II Background . . . . .	5
2.1 Preliminary material . . . . .	5
2.2 Equations of some known continuous distribution functions. . .	10
2.3 Histogram . . . . .	11
2.3.1 Mean square error (MSE) for histograms . . . . .	12
III Parametric density estimation . . . . .	13
3.1 Range restrictions . . . . .	15
IV Method . . . . .	19
4.1 Estimating the derivative of $F_n(x)$ . . . . .	20
4.2 Method of polynomial approximation . . . . .	23
4.3 Mode based density identification . . . . .	24
V Examples . . . . .	26
5.1 Example 1 . . . . .	26
5.2 Example 2 . . . . .	29
VI Conclusion . . . . .	32
VII Appendix . . . . .	33
7.1 R function to implement derivative construction algorithm. . .	33
Bibliography . . . . .	36



## LIST OF FIGURES

1.1	The impact of range restriction on mean and median . . . . .	3
3.1	Example distribution for range restrictions . . . . .	16
4.1	Empirical distribution function . . . . .	19
4.2	Flowchart of R function . . . . .	22
4.3	Approximations of distributions using the derivative of the polynomial	24
4.4	Approximations of distributions in the exponential family distribution	25
5.1	Polynomial approximations of distributions for example 1 . . . . .	27
5.2	Approximations of distributions in the exponential family distribution for example 1 . . . . .	28
5.3	Approximations of distributions by histograms for example 1 . . . . .	29
5.4	Polynomial approximations of distributions for example 2 . . . . .	30
5.5	Approximations of distributions in the exponential family distribution for example 2 . . . . .	30
5.6	Approximations of distributions by histograms for example 2 . . . . .	31

## CHAPTER I

### Introduction

In the field of statistics, one of the fundamental concepts of probability is *density estimation* using observed data [24, 21, 5]. A probability density function (pdf) of a continuous random variable  $X$  is a function  $f$  that describes the distribution of  $X$ . By utilizing  $f$ , we can find the probabilities associated with  $X$  in any interval  $(a, b)$  from the relationship

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx.$$

A random sample of the population can be considered as observed data and construction of the underlying unknown probability density function is called probability density estimation.

There are two approaches to density estimation; parametric and nonparametric. In parametric density approximation, we assume that the data are drawn from a known parametric family of distributions. Then the underlying density can be estimated based on sample statistics. There are many well-known parametric families of continuous distributions. The normal distribution, uniform distribution, students t distribution, gamma distribution, chi-square distribution, and Beta distribution are some of the most popular continuous distributions. These are distributions with 1–2 parameters that completely characterize the mean and variance. If either a model selection process justifies it, or it can be logically assumed that the observed data follows a well-known distribution, then parametric density estimation is possibly the simplest method for density approximation. The main motivation is to replace a potentially complex distribution by a simpler, mathematically tractable approximation. But often it is not appropriate to make such rigid assumptions about the underlying density functions. In such situations, nonparametric approaches are more appropriate since they do not make any assumptions about the underlying form of the distribution. Here, the density estimation will be constructed without imposing any

parametric constraints. These nonparametric methods make few assumptions and allow the data to drive the estimation process more directly in determining the estimate of the underlying distribution. There are many popular nonparametric density estimation methods. The histogram approach, naive estimator [29], kernel density estimation [24, 21], locally adaptive estimators [25], and nearest neighborhood method [29] are some of the most popular nonparametric density estimation methods.

Many diverse fields of study, such as astronomy [11], ecology [4], engineering [7], forestry [17], and public health [3] implement various density estimation methods to identify probability density functions. Histograms, Parzen windows, and Gaussian mixture models are some common example approaches [23]. Despite their simplicity, each approach has its own disadvantages. For example, histograms require an effective method of data binning (that is grouping the observations into a smaller number of “bins”), Parzen windows require careful selection of the kernel function and its parameters, and Gaussian mixture models are computationally inefficient and require the estimation of a large number of parameters [23]. Therefore, it is important to select the most suitable density estimation method depending on the knowledge of the origin and the restrictions of the data set. In this thesis, I present a parametric density estimation approach for when data are drawn from a truncated distribution.

One may choose a parametric density estimation method based on the identification of the population mean or median [30, 14]. However, the range of the observed data may be restricted and not available for a large portion of the domain. If the observed data is insufficient for accurate estimation of the mean or median, the accuracy of the density estimation may drop with the accuracy of the estimated mean or median (See Figure 1.1).

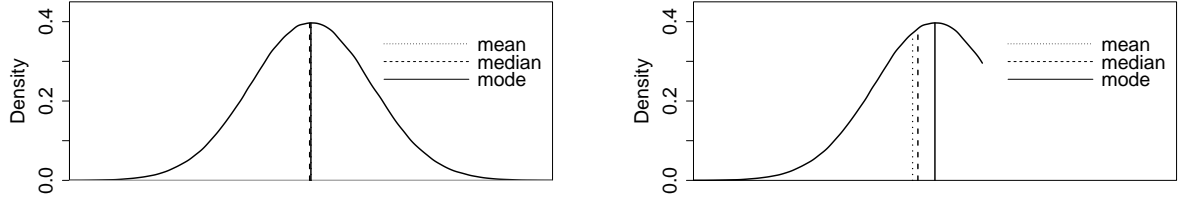


Figure 1.1: The impact of range restriction on mean and median.

Some variables may have well-defined bounded domains. For such data, density estimation methods (such as kernel density estimation) are frequently biased near the boundaries. If a smoothing technique is used to compensate for the errors near the boundaries, important features of the main regions in the distribution may disappear due to over-smoothing [32]. Moreover, present days, datasets are becoming more complex and larger, with multiple variables. Processing such large, complex datasets to estimate robust densities in a short time is challenging [18], since existing methods are computationally demanding for large datasets. Thus, the demand for new approaches to overcome such problems with less computational cost is upsurging.

The density function of a continuous random variable can be obtained by the derivative of the cumulative distribution function (CDF). The empirical distribution function generated by the sample data is an estimate for the cumulative distribution. In this thesis, I introduce two approaches based on the estimated derivative of the empirical distribution. The derivatives are estimated using the inverse function theorem and linear regression. The first method introduced in this thesis is to estimate densities using a polynomial approximation of the empirical distribution function. The corresponding assumption is that the data range is the same as the range of the bounded distribution. The distribution with bounded support is identified by finding the derivative of the estimated polynomial to the empirical distribution. To reduce unnecessary oscillations, we fixed the order of the approximated polynomial. The second method I introduce in this thesis is a mode-based density identification when the data range is restricted. There are many approaches in literature to identify the mode

[1, 31]. Here we use the normalized density function (estimated by the derivative of the empirical distribution function) to estimate the mode. The principal assumption of the proposed method is that the mode is contained in the data range (see section 4.3). A new approach is introduced to improve the efficiency of the density estimation by eliminating the model selection procedures. These methods may be faster, less complex and more accurate than some existing methods for range-restricted data sets.

Ecology is an interdisciplinary science that includes biology and earth science. It is the scientific analysis and study of interactions among organisms and their environment [16]. Distribution models have become a popular tool in ecology, allowing researchers to predict the underlying distributions of observed data. In ecological data analysis, there are many modeling approaches, each applying a distinct set of assumptions [22, 6]. But each approach has been the subject of much debate and currently there is no consensus as to the best single approach [10, 6, 27]. When it comes to ecological data, many data ranges are restricted by geographical or ecological constraints, such as measures of distances, which are restricted by landscape patterns, and tree characteristic measures, which are restricted by tree species. On the other hand, some variables have well-defined bounded domains, like percentage measures. Subsequently, the methods proposed here will be useful in ecological data analysis. As examples, we demonstrate application of the methods for bald eagle and Indiana bat nesting habitats.

## CHAPTER II

### Background

This chapter presents some preliminary and background material for the research presented in this thesis. Section 2.1 covers important definitions, lemmas, and theorems necessary for the theoretical work. Section 2.2 provides some examples of continuous distributions utilized in this thesis. Section 2.3 provides a review of histograms, as they will be used to compare our density approximations. The background for parametric density estimation is presented in Chapters 3.

#### 2.1 Preliminary material

**Definition 2.1.1. (*Probability space*)** An ordered triple  $(\Omega, \mathcal{F}, \mathbb{P})$  where;

1.  $\Omega$  is a set (possible outcomes); elements are  $\omega$  called elementary outcomes.
2.  $\mathcal{F}$  is a family of subsets (events) of  $\Omega$  where each event is a set containing zero or more outcomes:
  - Empty set  $\emptyset$  and  $\Omega$  are members of  $\mathcal{F}$ .
  - $A \in \mathcal{F}$  implies  $A^c = \{\omega \in \Omega : \omega \notin A\} \in \mathcal{F}$ .
  - $A_1, A_2 \dots$  in  $\mathcal{F}$  implies  $A = \cup_{i=1}^{\infty} A_i \in \mathcal{F}$ .
3.  $\mathbb{P}$  is a function, domain  $\mathcal{F}$ , range a subset of  $[1, 0]$ ; that is, a function  $\mathbb{P}$  from events to probabilities.

**Definition 2.1.2. (*Random variable*)** Consider a random experiment with a sample space  $C$ . A function  $X$ , which assigns to each element  $c \in C$  one and only one number  $X(c) = x$  is called a random variable. The space or range of  $X$  is the set of real numbers  $\mathcal{D} = \{x : x = X(c), c \in C\}$ . Where,  $\mathcal{D}$  generally is a countable set or an interval of real numbers.

**Definition 2.1.3. (Discrete Random Variable)** We say a random variable is a discrete random variable if its space is either finite or countable.

**Definition 2.1.4. (Continuous Random Variables)** We say a random variable is a continuous random variable if its cumulative distribution function  $F_X(x)$  is a continuous function for all  $x \in \mathbb{R}$ .

**Definition 2.1.5. (Indicator function)** Let  $\Omega$  be a sample space and  $\mathcal{F} \subset \Omega$  be an event the indicator function of event  $\mathcal{F}$ , denoted by  $1_{\mathcal{F}}$  is a random variable defined as follows:

$$1_{\mathcal{F}}(x) = \begin{cases} 1 & ; \text{ if } \omega \in \mathcal{F} \\ 0 & ; \text{ if } \omega \notin \mathcal{F} \end{cases}$$

**Definition 2.1.6. (Probability Mass Function (pmf))** Let  $X$  be a discrete random variable with space  $\mathcal{D}$ . The probability mass function (pmf) of  $X$  is given by

$$p_X(x) = P[X = x], \text{ for } x \in \mathcal{D}.$$

Properties of probability mass function:  $P_X(x)$

- $0 \leq P_X(x) \leq 1$  for any  $x \in \mathcal{D}$  (Sample space).
- $\sum_{x \in \mathcal{D}} P_X(x) = 1$ .
- for any  $A \subset \mathcal{D}$ ,  $P_X(X \in A) = \sum_{x \in A} P_X(x)$ .

**Definition 2.1.7. (Cumulative Distribution Function)** Let  $X$  be a random variable. Then its cumulative distribution function (cdf) is defined by  $F_X(x)$ , where

$$F_X(x) = P_X((-\infty, x]) = P(\{c \in C : X(c) \leq x\}).$$

The cumulative distribution function (CDF) for a real-valued random variable  $X$  is a function which gives the probability that the random variable takes a value less than

or equal to  $x$ . The cumulative distribution function for a discrete random variable is defined as:

$$F_X(X \leq x) = \sum_{x \leq t} P_X(t).$$

On the other hand, the cumulative distribution function for a real valued continuous random variable is the area under the curve of the probability density function upto  $x$ ; where  $-\infty < x < \infty$ .

$$F_X(X \leq x) = \int_{-\infty}^x f_X(t) dt.$$

Most continuous random variables are **absolutely continuous**; that is,

$$F_X(x) = \int_{-\infty}^x f(t) dt,$$

for some function  $f_X(t)$ . The function  $f_X(t)$  is called a **probability density function** (pdf) of  $X$ .

Properties of probability density function:  $f_X(x)$ .

- $f_X(x) \geq 0$  for all  $x \in \mathbb{R}$ .
- $\int_{-\infty}^{\infty} f_X(t) dt = 1$ .
- $P(a \leq X \leq b) = \int_a^b f_X(t) dt$ .

**Definition 2.1.8. (Order statistic)** The order statistics of a random sample  $X_1, X_2, \dots, X_n$  from a continuous distribution (with support  $a < b$ ) are the sample values placed in ascending order. They are denoted by  $X_{\{1\}}, X_{\{2\}}, \dots, X_{\{n\}}$  and

$$X_{\{1\}} \leq X_{\{2\}} \leq \dots \leq X_{\{n\}}.$$

**Definition 2.1.9. (Expectation)** Let  $X$  be a random variable. If  $X$  is a continuous random variable with pdf  $f(x)$  and

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty,$$



then the expectation of  $X$  is

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

If  $X$  is a discrete random variable with pmf  $p(x)$  and

$$\sum_x |x|p(x) < \infty,$$

then the expectation of  $X$  is

$$E(X) = \sum_x xp(x).$$

**Definition 2.1.10. (Bias of  $\hat{\theta}$  with respect to  $\theta$ )**

$$Bias_{\theta}[\hat{\theta}] = E_{x|\theta}[\hat{\theta}] - \theta = E_{x|\theta}[\hat{\theta} - \theta].$$

Where  $E_{x|\theta}$  denotes expected value over the distribution  $P(x|\theta)$ .

**Definition 2.1.11. (Mean squared error (MSE))** The MSE of an estimator  $\hat{\theta}$  with respect to an unknown parameter  $\theta$  is defined as

$$MSE(\hat{\theta}) = E_{\hat{\theta}}[(\hat{\theta} - \theta)^2].$$

Variance and bias relationship.

$$MSE(\hat{\theta}) = Var_{\hat{\theta}}(\hat{\theta}) + Bias_{\hat{\theta}}(\hat{\theta}, \theta)^2.$$

**Definition 2.1.12. (Likelihood function)** Let  $X_1, X_2, \dots, X_n$  denote a sample on a random variable  $X$  with pdf  $f(x; \cdot)$ . The likelihood function of the random sample is given by

$$L(\theta) = L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta).$$

**Definition 2.1.13. (Convergence in probability)** Let  $\{X_n\}$  be a sequence of random variables and let  $X$  be a random variable defined on a sample space. We say that  $X_n$  convergence in probability to  $X$  if, for all  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P[|X_n - X| \geq \epsilon] = 0$$

or equivalently,

$$\lim_{n \rightarrow \infty} P[|X_n - X| < \epsilon] = 1$$

if so, we write

$$X_n \xrightarrow{P} X.$$

**Definition 2.1.14. (*Almost surely*)** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space. An event  $E \in \mathcal{F}$  happens almost surely if  $\mathbb{P}(E^c) = 0$ .

**Definition 2.1.15. (*arg max*)** Given an arbitrary set  $X$ , a totally ordered set  $Y$ , and a function,  $f: X \rightarrow Y$ , the  $\arg \max$  over some subset,  $S$ , of  $X$  is defined by

$$\arg \max_{x \in S \subseteq X} : = \{x | x \in S \wedge \forall y \in S: f(y) \leq f(x)\}.$$

**Definition 2.1.16. (*arg min*)** Given an arbitrary set  $X$ , a totally ordered set  $Y$ , and a function,  $f: X \rightarrow Y$ , the  $\arg \min$  over some subset,  $S$ , of  $X$  is defined by

$$\arg \min_{x \in S \subseteq X} : = \{x | x \in S \wedge \forall y \in S: f(y) \geq f(x)\}.$$

**Definition 2.1.17. (*Lipschitz continuity*)** A function  $g$  is continuous in interval  $B$  if there is a constant  $L > 0$  such that for all  $x, y \in B$

$$|g(x) - g(y)| \leq L|x - y|.$$

**Lemma 2.1.1. (*Uniform law of large numbers*)** If the data are i.i.d.,  $\Theta$  is compact,  $a(z_i, \theta)$  is continuous at each  $\theta \in \Theta$  with probability one, and there is  $d(z)$  with  $\|a(z, \theta)\| \leq d(z)$  for all  $\theta \in \Theta$  and  $E[d(z)] < \infty$ , then  $E[a(z, \theta)]$  is continuous and  $\sup_{\theta \in \Theta} \|n^{-1} \sum_{i=1}^n a(z_i, \theta) - E[a(z, \theta)]\| \xrightarrow{P} 0$ .

**Theorem 2.1.1. (*Inverse function theorem dimension one*)** If  $f: (a, b) \rightarrow (c, d)$  is a differentiable surjection and  $f'(x)$  is never zero then  $f$  is a homeomorphism, and its inverse is differentiable with derivative

$$(f^{-1})'(y) = \frac{1}{f'(x)}$$

Where  $y = f(x)$ .

## 2.2 Equations of some known continuous distribution functions.

**Normal distribution,**  $Normal(\mu, \sigma^2)$ .

Probability density function of normal distribution is

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} ; \text{ where } -\infty \leq x \leq \infty.$$

Here  $\mu$  and  $\sigma^2$  are the population mean and variance respectively.

**Gamma distribution,**  $Gamma(\alpha, \beta)$ .

Probability density function of gamma distribution is

$$f(x, \alpha, \beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & ; \text{ if } x > 0 \\ 0 & ; \text{ otherwise.} \end{cases}$$

Here  $\alpha > 0$  and  $\beta > 0$  are the shape and scale parameters respectively and

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

**Weibull distribution,**  $Weibull(\lambda, k)$ .

Probability density function of weibull distribution is

$$f(x, \lambda, k) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & ; \text{ if } x \geq 0 \\ 0 & ; \text{ if } x < 0. \end{cases}$$

Here  $k > 0$  is the shape parameter and  $\lambda > 0$  is the scale parameter.

**Uniform distribution,**  $Uniform(a, b)$ .

Probability density function of the continuous uniform distribution is

$$f(x, a, b) = \begin{cases} \frac{1}{b-a} & ; \text{ if } a \leq x \leq b \\ 0 & ; \text{ otherwise.} \end{cases}$$

Here  $a$  and  $b$  are the two constant boundaries.

**Exponential distribution,  $Exponential(\lambda)$ .**

Probability density function of exponential distribution is

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & ; \text{ if } x \geq 0 \\ 0 & ; \text{ if } x < 0. \end{cases}$$

Here  $\lambda > 0$  is the rate parameter.

**Beta distribution,  $Beta(\alpha, \beta)$ .**

Probability density function of Beta distribution on the interval  $[0, 1]$  is

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} ; \text{ where } 0 \leq x \leq 1$$

Here  $\alpha$  and  $\beta$  are shape parameters and

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy.$$

## 2.3 Histogram

A histogram is an estimate for the probability distribution of a random variable using observed data values. Histograms are constructed using non-overlapping intervals, called “bins”, versus the relative frequency of the data in each bin. Let  $X_1, \dots, X_n$  be the observations. Suppose  $r_i \in \mathbb{R}, i \in \mathbb{Z}$ , with  $\dots r_{-1} < r_0 < r_1 \dots$ , defines a partition of the real line. Define  $v_k = \sum_{i=1}^n 1\{X_i \in (r_k, r_{k+1}]\}$  ( $v_k$  has a binomial distribution). Then we can define histogram by

$$\hat{f}_n(x) = \frac{v_k}{n(r_{k+1} - r_k)} = \frac{1}{n(r_{k+1} - r_k)} \sum_{i=1}^n 1\{X_i \in (r_k, r_{k+1}]\} \text{ for } x \in (r_k, r_{k+1}].$$

Let  $r_{k+1} - r_k = h$  and  $r_0$  be the origin. Define  $\lfloor x \rfloor = \sup \{k \in \mathbb{Z} : x > k\}$ . Then

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n 1\{X_i - r_0 \in (hk_x, h(k_x + 1)]\} , \text{ where } k_x = \lfloor (x - r_0/h) \rfloor.$$

Note that the histogram can be written as a function of the empirical distribution function:  $\hat{f}_n(x) = \frac{F_n(r_{k+1}) - F_n(r_k)}{h}$ , where  $F_n(x)$  is the empirical distribution function.

### 2.3.1 Mean square error (MSE) for histograms

Since  $v_k \sim \text{Binomial}(n, p_k)$  with  $p_k = P(r_k < X < r_k + h) = \int_{r_k}^{r_k+h} f(t)dt$ ,  $E[\hat{f}(x)] = p_k/h$  and  $\text{Var}[\hat{f}(x)] = p_k(1 - p_k)/(nh^2)$ . Let us assume the underlying density is Lipschitz continuous with Lipschitz constant  $L > 0$ . Let  $x \in (r_0 + hk, r_0 + h(k+1)]$ . Then

$$\begin{aligned}
 |E[\hat{f}_n(x)] - f(x)| &= |(p_k/h) - f(x)| \\
 &= \left| \frac{1}{h} \int_{r_0+hk}^{r_0+h(k+1)} f(t)dt - f(x) \right| \\
 &= \left| \frac{1}{h} \int_{r_0+hk}^{r_0+h(k+1)} (f(t) - f(x))dt \right| \\
 &\leq \frac{1}{h} \int_{r_0+hk}^{r_0+h(k+1)} |f(t) - f(x)|dt \\
 &\leq \frac{1}{h} \int_{r_0+hk}^{r_0+h(k+1)} Lhdt \\
 &= Lh.
 \end{aligned}$$

We choose  $h \equiv h_n \rightarrow 0$  as  $n \rightarrow \infty$ . Then  $\text{Bias}(\hat{f}_n(x)) = E[\hat{f}_n(x)] - f(x) \rightarrow 0$  as  $n \rightarrow \infty$ . Similarly,

$$\text{Var}(\hat{f}_n(x)) = \frac{p_k(1 - p_k)}{nh^2} \leq \frac{p_k}{nh^2} \leq \frac{h(|f(x)| + Lh)}{nh^2} \text{ (under the above assumptions).}$$

Therefore,  $\text{Var}(\hat{f}_n(x)) \leq \frac{|f(x)|}{nh} + \frac{L}{n}$ . By these results, the MSE of the histogram estimator is

$$\text{MSE}(\hat{f}_n(x)) \leq \frac{|f(x)|}{nh} + \frac{L}{n} + (Lh)^2.$$

We want to choose  $h$  to be small enough to make the bias small, but also large enough to ensure the variance is also small.

## CHAPTER III

### Parametric density estimation

Consider a sample  $X_1, X_2, \dots, X_n$  of independent identically distributed (iid) univariate continuous random variables. Suppose  $\hat{f}$  is the underlying common density function and we seek to identify  $\hat{f}$  from the data set. Let  $\hat{f}$  belong to the parametric family  $\mathcal{F}$ , where

$$\mathcal{F} \triangleq \{f(\cdot, \theta) : \theta \in S \subset \mathbb{R}^k\},$$

and  $k$  is a sufficiently small positive integer. Here  $\theta$  represents the parameter set. For example, for the family of normal densities, we have  $\theta = (\mu, \sigma)$  and  $f(x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ . One may choose the parametric family by pre-modeling (such as using the shape of the histogram), assuming a location-scale family of probability distributions, or using a model selection criterion. In parametric density estimation, we assume that there exists  $\theta_0 \in S$  such that  $\hat{f}(x) = f(x, \theta_0)$ ,  $-\infty < x < \infty$ .

The standard method of estimating  $\theta_0$  is maximum likelihood estimation [9, 12]. We define the likelihood function with parameter  $\theta$ ,  $L_n(\mathbf{x}, \theta)$ , as the joint density function of  $X_1, X_2, \dots, X_n$ . Since  $X_1, X_2, \dots, X_n$  are iid, we have  $L_n(\mathbf{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$ . Then the corresponding parameter estimation problem can be formulated as

$$\begin{aligned} & \text{maximize} && \prod_{i=1}^n f(X_i, \theta), \\ & \text{subject to} && \theta \in S. \end{aligned} \tag{3.1}$$

$$\tag{3.2}$$

If the maximum exists and equals  $\hat{\theta}$  (that is  $\hat{\theta} = \arg \max L_n(\mathbf{x}, \theta)$ ), we call  $\hat{\theta}$  the maximum likelihood estimator (MLE). Since maximizing a product is a bit challenging and tedious, we often maximize the average log likelihood function,

$$l_n(\mathbf{x}, \theta) = \frac{1}{n} \log(L_n(\mathbf{x}, \theta)) = \frac{1}{n} \sum_{i=1}^n \log f(x_i, \theta).$$

Since  $L_n(\mathbf{x}, \theta) \geq 0$  and the logarithmic function is monotone increasing on  $(0, \infty)$ , maximizing the likelihood function is equivalent to maximizing the average log likeli-

hood function. Maximum likelihood estimators do not have optimum properties with finite sample size (note that  $L_n(\mathbf{x}, \theta)$  depends on sample size), but when the sample size increases to infinity, the sequence of MLEs possesses consistency and efficiency (see page 292 and 332 of [15]). To show the consistency of the MLE, let  $L(\theta)$  be the expectation of function,  $\log f(X, \theta)$ , with respect to the true unknown parameter  $\theta_0$ . Then,

$$L(\theta) = E_{\theta_0}[\log f(X, \theta)] = \int \log f(x, \theta) f(x, \theta_0) dx.$$

Assume that there is a function  $d(x)$  such that for all  $\theta \in S$ ,  $\|\log f(X, \theta)\| \leq d(X)$  and  $E[d(X)] < \infty$ . By the uniform law of large numbers (Lemma 2.4 of [20]), for all  $\theta$ ,  $l_n(\theta) \rightarrow L(\theta)$  almost surely. Now consider the difference  $L(\theta) - L(\theta_0)$ ,

$$E_{\theta_0}[\log f(X, \theta) - \log f(X, \theta_0)] = E_{\theta_0}\left[\log \frac{f(X, \theta)}{f(X, \theta_0)}\right].$$

Since  $\log(t) \leq t - 1$ ,

$$\begin{aligned} E_{\theta_0}\left[\log \frac{f(X, \theta)}{f(X, \theta_0)}\right] &\leq E_{\theta_0}\left[\frac{f(X, \theta)}{f(X, \theta_0)} - 1\right] = \int \left(\frac{f(x, \theta)}{f(x, \theta_0)} - 1\right) f(x, \theta_0) dx \\ &= \int f(x, \theta) dx - \int f(x, \theta_0) dx = 1 - 1 = 0. \end{aligned}$$

Therefore,  $L(\theta) \leq L(\theta_0)$ . This means  $\theta_0$  is the maximizer of  $L(\theta)$ . Now we have

- $\hat{\theta}$  is the maximizer of  $l_n(\theta)$  (by definition).
- For all  $\theta$ ,  $l_n(\theta) \rightarrow L(\theta)$  almost surely.
- $\theta_0$  is the maximizer of  $L(\theta)$ .

Based on the above  $\hat{\theta} \rightarrow \theta_0$  as  $n \rightarrow \infty$ .

As an example, consider a sample data set from a normal distribution. Let  $X_1, X_2, \dots, X_n$  be a independent, identically distributed (i.i.d.), random variables such that  $X_i \sim \text{Normal}(\mu, \sigma^2)$  for all  $i$ . Then  $f(X, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$  and

$L_n(X, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$ . Therefore, the average log likelihood estimator is given by

$$l_n(X, \mu, \sigma^2) = \frac{1}{n} \sum_{i=1}^n \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right) = \frac{1}{n} \sum_{i=1}^n \left( -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

Let  $a(X, \mu, \sigma) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x - \mu)^2}{2\sigma^2}$ , which is a continuous function of  $\mu$  and  $\sigma$ .

With the substitution  $u = \frac{x - \mu}{\sigma}$ , we have

$$\begin{aligned} E[a(X, \mu, \sigma)] &= \int_{-\infty}^{\infty} \left( -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x - \mu)^2}{2\sigma^2} \right) \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}} \right) dx \\ &= \int_{-\infty}^{\infty} \left( -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} u^2 \right) \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \right) du = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2}. \end{aligned}$$

Here we have utilized the fact that, for  $X \sim \text{Normal}(0, 1)$ ,  $E[X^0] = E[X^2] = 1$ .

If  $d(X) = C_1 + C_2|x| + C_3x^2$  for some positive constants  $C_1, C_2$  and  $C_3$ , we have

$\|a(X, \mu, \sigma)\| \leq d(X)$  and  $E[d(x)] < \infty$ . By the uniform law of large numbers, the average log likelihood estimator

$$l_n(X, \mu, \sigma^2) = \frac{1}{n} \sum_{i=1}^n \left( -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{n}$$

converges almost surely to  $-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2}$ . We have  $\sum_{i=1}^n \frac{(x_i - \mu)^2}{n} \rightarrow \sigma^2$  as desired.

### 3.1 Range restrictions

Suppose  $X_1, X_2, \dots, X_n$  are independent, identically distributed random variables, with common density  $f(x, \theta)$ , where  $-\infty < x < \infty$ . Let  $F(x, \theta)$  be the corresponding cumulative distribution function. The truncated distribution of  $f(X, \theta)$  that results from restricting the domain to  $a \leq x \leq b$  (for some positive constants  $a$  and  $b$ ) is given by

$$f_T(x, \theta) = \begin{cases} \frac{f(x, \theta)}{F(b, \theta) - F(a, \theta)}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

We will show that the maximum likelihood estimation method fails to converge to the true parameter values when the distribution is truncated. For simplicity, assume



that  $f(a, \theta) = f(b, \theta) = \alpha$  and  $f(x, \theta) > \alpha$  for all  $x \in (a, b)$  and some positive constant  $\alpha$  (see Figure 3.1).

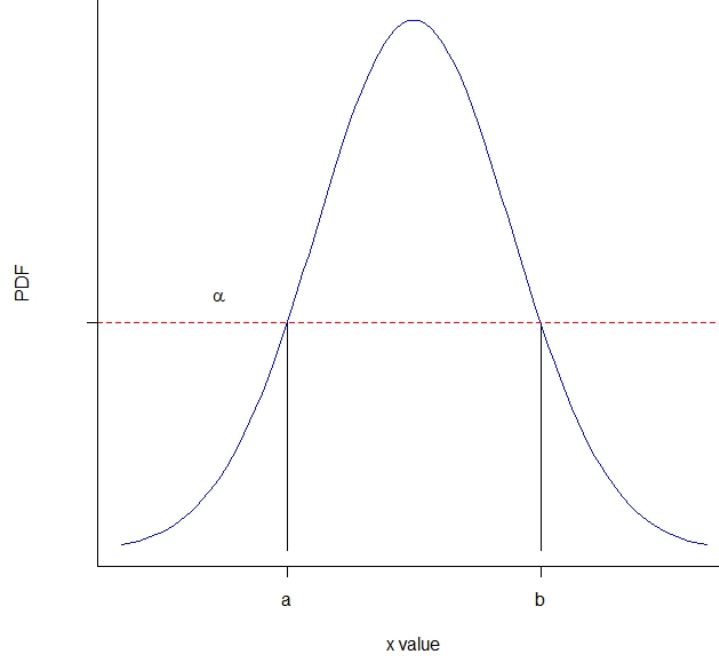


Figure 3.1: Example distribution for which  $f(a, \theta) = f(b, \theta) = \alpha$  and  $f(x, \theta) > \alpha$  for all  $x \in (a, b)$  and some positive constant  $\alpha$ . If  $x < a$  or  $x > b$  the range of the pdf is below  $\alpha$ .

Let  $\tilde{L}(\theta)$  be the expectation of  $\ln f(X, \theta)$  with respect to  $f_T(X, \theta_0)$  and  $L(\theta)$  be the expectation of  $\ln f(X, \theta)$  with respect to  $f(X, \theta_0)$ . Then  $\tilde{L}(\theta) = \int_a^b \ln f(x, \theta) f_T(x, \theta_0) dx$  and  $L(\theta) = \int_{-\infty}^{\infty} \ln f(x, \theta) f(x, \theta_0) dx$ . We have

$$\begin{aligned} \|\tilde{L}(\theta) - L(\theta)\| &= \left\| \int_a^b \ln f(x, \theta) f_T(x, \theta_0) dx - \int_{-\infty}^{\infty} \ln f(x, \theta) f(x, \theta_0) dx \right\| \\ &= \left\| \int_a^b \ln f(x, \theta) f(x, \theta_0) \left( \frac{1}{F(b) - F(a)} - 1 \right) dx - \int_{-\infty}^a \ln f(x, \theta) f(x, \theta_0) dx \right. \\ &\quad \left. - \int_b^{\infty} \ln f(x, \theta) f(x, \theta_0) dx \right\|. \end{aligned}$$

For  $\alpha > 0$ ,

$$\begin{aligned}
\|\tilde{L}(\theta) - \mathbb{L}(\theta)\| &= \left\| \int_a^b \left( \ln \alpha + \ln \frac{f(x, \theta)}{\alpha} \right) f(x, \theta_0) \left( \frac{1}{F(b) - F(a)} - 1 \right) dx \right. \\
&\quad \left. - \int_{-\infty}^a \left( \ln \alpha + \ln \frac{f(x, \theta)}{\alpha} \right) f(x, \theta_0) dx - \int_b^{\infty} \left( \ln \alpha + \ln \frac{f(x, \theta)}{\alpha} \right) f(x, \theta_0) dx \right\| \\
&= \left\| \int_a^b \ln \frac{f(x, \theta)}{\alpha} f(x, \theta_0) \left( \frac{1}{F(b) - F(a)} - 1 \right) dx \right. \\
&\quad \left. - \int_{-\infty}^a \ln \frac{f(x, \theta)}{\alpha} f(x, \theta_0) dx - \int_b^{\infty} \ln \frac{f(x, \theta)}{\alpha} f(x, \theta_0) dx \right. \\
&\quad \left. + \ln \alpha \int_a^b f(x, \theta_0) \left( \frac{1}{F(b) - F(a)} \right) dx - \ln \alpha \int_{-\infty}^{\infty} f(x, \theta_0) dx \right\|.
\end{aligned}$$

Since  $\int_a^b f(x, \theta_0) \left( \frac{1}{F(b) - F(a)} \right) dx = \int_{-\infty}^{\infty} f(x, \theta_0) dx = 1$ ,

$$\begin{aligned}
\|\tilde{L}(\theta) - \mathbb{L}(\theta)\| &= \left\| \int_a^b \ln \frac{f(x, \theta)}{\alpha} f(x, \theta_0) \left( \frac{1}{F(b) - F(a)} - 1 \right) dx + \int_{-\infty}^a \ln \frac{\alpha}{f(x, \theta)} f(x, \theta_0) dx \right. \\
&\quad \left. + \int_b^{\infty} \ln \frac{\alpha}{f(x, \theta)} f(x, \theta_0) dx \right\|
\end{aligned}$$

Here we have three integral terms.

- In the first term,  $\ln \frac{f(x, \theta)}{\alpha} > 0$  in the interval  $(a, b)$  and  $\left( \frac{1}{F(b) - F(a)} \right) > 1$  for any distinct  $a, b$  values where  $a < b$ . Therefore, the first integral term is positive.
- In the middle term,  $\ln \frac{\alpha}{f(x, \theta)} > 0$  in the interval  $(-\infty, a)$ . Therefore, the second integral term is positive.
- In the last term,  $\ln \frac{\alpha}{f(x, \theta)} > 0$  in the interval  $(b, \infty)$ . Therefore, the final integral term is also positive.

Then the error in convergence is  $\|\tilde{L}(\theta) - \mathbb{L}(\theta)\| > 0$ .

For example, consider the truncated normal distribution  $N(\mu, \sigma^2)$  with the domain is

constrained by restricted to  $[0, b]$ . Let  $a(X, \theta) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$ . Then

$$\begin{aligned}\tilde{L}(\theta) &= E[a(X, \theta)] = \int_0^b a(x, \theta) f_T(x, \theta_0) dx \\ &= \int_{-\infty}^{\infty} a(x, \theta) f_T(x, \theta_0) dx - \int_{-\infty}^0 a(x, \theta) f_T(x, \theta_0) dx - \int_b^{\infty} a(x, \theta) f_T(x, \theta_0) dx \\ &= -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} - \frac{\alpha_0}{2}\end{aligned}$$

for some positive constant  $\alpha_0$ . By the uniform law of large numbers, the average log likelihood estimator

$$l_n(X, \mu, \sigma^2) = \frac{1}{n} \sum_{i=1}^n \left( -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{n}$$

converges almost surely to  $-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} - \frac{\alpha_0}{2}$ . Then  $\sum_{i=1}^n \frac{(x_i - \mu)^2}{n}$  converges to  $(1 - \alpha_0)\sigma^2$ , resulting in errors in estimation. Here we introduce a method to solve above mentioned issue in identifying truncated distributions.

## CHAPTER IV

### Method

In this chapter we propose several methods to improve the density identification when the data range is restricted. The proposed methods are based on the empirical distribution function. For a set of i.i.d. random variables  $X_1, X_2, \dots, X_n$  with distribution function  $F(x) = \mathbb{P}(X_i \leq x)$ , the empirical distribution function is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\},$$

where  $\mathbf{1}$  is the indicator function. This is the distribution function of a discrete random variable with values  $X_1, X_2, \dots, X_n$  (provided that they are distinct). If we define an order statistic such that  $X_{\{i\}} < X_{\{j\}}$  for  $i < j$ , then  $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_{\{i\}} \leq x\}$  will be the staircase function with a jump of  $1/n$  at each  $X_{\{i\}}$ .

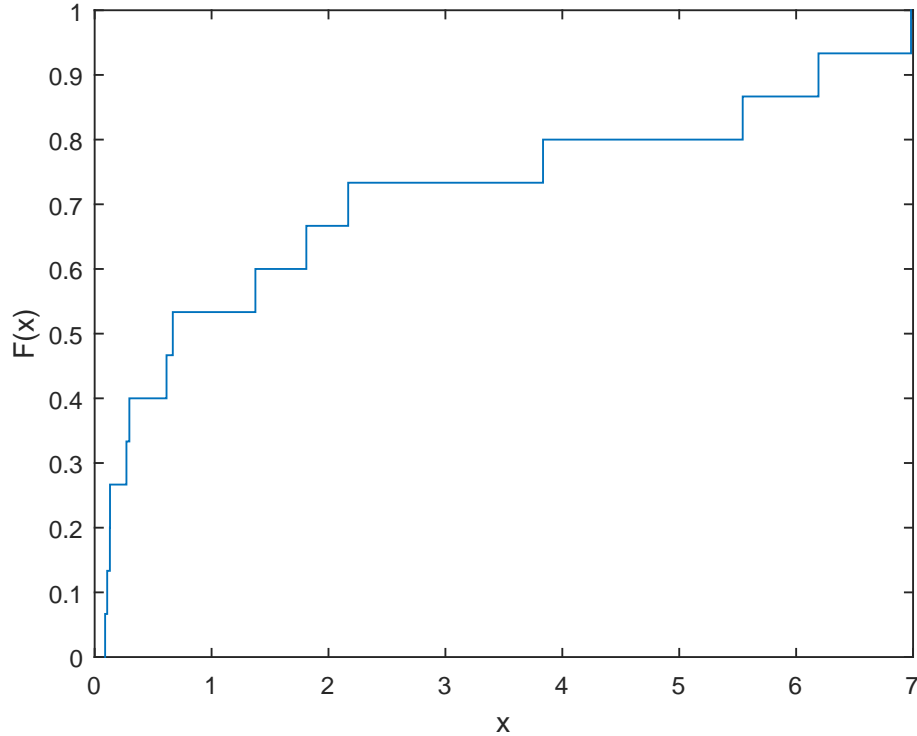


Figure 4.1: Empirical distribution function

Subsequently, we may view  $F_n(x)$  as the sum of  $n$  independent Bernoulli random variables with success probability  $F(x)$ . Then  $nF_n(x)$  is binomial with  $\mathbb{E}(F_n(x)) = F(x)$  and  $\text{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$ . It is easy to see that, as  $n \rightarrow \infty$ ,  $F_n(x)$  converges to  $F(x)$  in probability. We even have a stronger convergence result.

**Theorem 4.0.1. (*Glivenko-Cantelli* [13, 2])**

*For any distribution,  $F_n(x)$  converges uniformly to  $F(x)$  almost surely. That is,  $\|F_n(x) - F(x)\|_\infty \rightarrow 0$  almost surely.*

In light of this result, we use  $F_n(x)$  as the foundation to construct the density approximations.

#### 4.1 Estimating the derivative of $F_n(x)$

The underlying density function  $f_X(x)$  of a continuous distribution  $F(x)$  can be obtained by the derivative  $F'(x)$ . Below we discuss a method to construct the derivative at each observed value using the empirical distribution.

Let  $\{x_k\}_{k=1}^n \subset \mathbb{R}$  be a data set arranged in ascending order and  $S_i = \{x_{i-3}, x_{i-2}, \dots, x_{i+3}\} \subset \{x_k\}_{k=1}^n$  having  $x_{i+3} > x_{i-3}$  for all  $i \geq 4$ . Assume that  $S$  consists of distinct data points (if not, we can choose arbitrarily small perturbations to make the data points distinct). Suppose  $\mathfrak{F} \in C^1(\Omega)$  is a strictly increasing smoothing approximation for the empirical distribution  $F_n(x)$  on  $S$  such that  $\mathfrak{F}(x_i) = i/n$ , where  $\Omega$  is an open set containing  $[x_{i-3}, x_{i+3}]$ . Then we have  $\mathfrak{F}'(x) > 0$  for all  $x \in \Omega$ . Let  $J$  be the image of  $\Omega$  under  $\mathfrak{F}$ . From the inverse function theorem, function  $\mathfrak{F}$  has an inverse  $\mathfrak{G} : J \rightarrow \Omega$  such that  $\mathfrak{G} \in C^1(J)$  with derivative  $\mathfrak{G}'(s) = \frac{1}{\mathfrak{F}'(\mathfrak{G}(s))}$ ,  $s \in J$ . Since  $\mathfrak{F}'(x_i) = \frac{1}{\mathfrak{G}'(i/n)}$ , we may approximate  $\mathfrak{F}'(x_i)$  using the inverse function. Here we let  $\mathfrak{F}(x)$  be linear and, using the slope of the simple linear regression equation  $\mathfrak{G}(s) = ms + b$ ,

$$\mathfrak{F}'(x_i) = \frac{1}{m} = \frac{\sum_{j=-3}^3 (j/n)^2}{\sum_{j=-3}^3 j(x_{i+j} - \bar{x})/n} = \frac{28/n}{3x_{i+3} + 2x_{i+2} + x_{i+1} - x_{i-1} - 2x_{i-2} - 3x_{i-3}},$$

where  $\bar{x}$  is the mean of the data subset  $S$ .

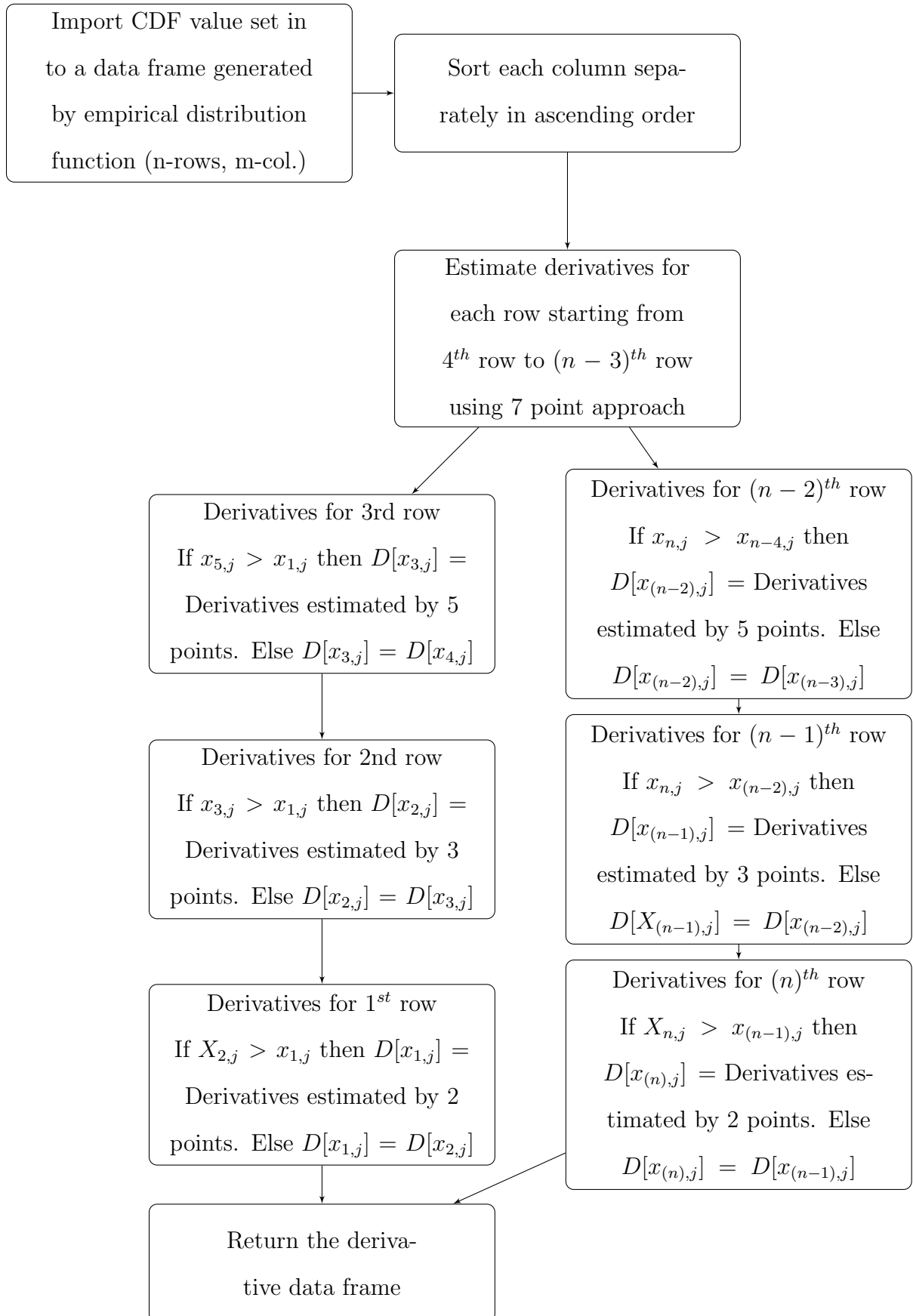
To obtain the derivative at range boundary, we assume that the derivatives near the boundary can be approximated by an exponential function. Here we explain only the estimation of  $\mathfrak{F}'(x_i)$  for  $i = 1, 2, 3$ ; the estimation of  $\mathfrak{F}'(x_i)$  for  $i = n - 2, n - 1, n$  is similar. Consider the regression model  $\mathfrak{F}'(x) \approx e^{a(x-x_1)+b}$  so that  $\mathfrak{F}'(x_1) = e^b$ . We estimate  $a$  and  $b$  using the estimated derivatives at the first four data points,  $\{x_1, x_2, x_3, x_4\}$ , similar to the construction above. Suppose  $S_0 = \{x_1, x_2, \dots, x_7\}$ .

- If  $x_5 > x_1$ , then  $\mathfrak{F}'(x_3) \approx \frac{\sum_{j=-2}^2 (j/n)^2}{\sum_{j=-2}^2 j (x_{3+j} - \bar{x}) / n} = \frac{10/n}{2x_5 + x_4 - x_2 - 2x_1}$ . Otherwise  $\mathfrak{F}'(x_3) \approx \mathfrak{F}'(x_4)$ .
- If  $x_3 > x_1$ , then  $\mathfrak{F}'(x_2) \approx \frac{\sum_{j=-1}^1 (j/n)^2}{\sum_{j=-1}^1 j (x_{1+j} - \bar{x}) / n} = \frac{2/n}{x_3 - x_1}$ . Otherwise  $\mathfrak{F}'(x_2) \approx \mathfrak{F}'(x_3)$ .
- If  $x_2 > x_1$ , then  $\mathfrak{F}'(x_1) \approx \frac{1/n}{x_2 - x_1}$  (forward difference approximation). Otherwise  $\mathfrak{F}'(x_1) \approx \mathfrak{F}'(x_2)$ .

We find  $a$  and  $b$  using least squares regression.

We created an R function, “derivative”, to estimate the derivatives. Let rows represent the observations ( $n$  number of rows) and columns represent the variables ( $m$  number of columns). The flowchart below explains the algorithm. Section 7.1 provides the code for the function.

Figure 4.2: Flowchart of R function.



## 4.2 Method of polynomial approximation

One way to identify the distribution  $F(x)$  with bounded support is by fitting a polynomial to the empirical distribution. The principal assumption is that the data range is the same as the range of the bounded distribution (such as a truncated distribution). To reduce any unnecessary oscillations resulting from higher degree polynomials, it is appropriate to choose a lower degree approximation. Assuming that the density can have at most two peaks (a bimodal distribution), the degree of the fitting polynomial can be reduced to as small as four, such that the derivative can contain two peaks. Here, we use a fifth order polynomial with constraints to construct the approximation, as demonstrated below.

Consider the polynomial

$$\begin{aligned} p(x) = & x^3 + c_0 (x - x^3) + (3 - 2c_0 - c_1) (x^2 - x^3) \\ & + c_2 (x^2 + x^4 - 2x^3) + c_3 (2x^2 + x^5 - 3x^3), \end{aligned} \quad (4.1)$$

where  $x \in [0, 1]$  and  $c_0, \dots, c_3$ , are model parameters. Then,  $p(0) = 0$ ,  $p(1) = 1$ ,  $p'(0) = c_0$ , and  $p'(1) = c_1$ . Suppose  $X$  is a random variable that can take values in  $[0, 1]$  and  $c_0$  and  $c_1$  are density estimations at  $x = 0$  and  $x = 1$  (see Section 4.1). Then  $p(x)$  describes a cumulative density function on  $[0, 1]$  having a corresponding probability density function that is approximated by  $p'(x)$ , with boundaries of  $p'(x)$  equal those of the probability density function. We obtain  $c_2$  and  $c_3$  by least squares regression on the empirical distribution function. Density function  $p'(x)$  can be at most bimodal. Using transformation  $x = \frac{y - a}{b - a}$ ,  $p(x)$  is well suited to approximate any unimodal distribution on  $y \in [a, b]$  (see Figure 4.3 (a) – (f) for some examples).



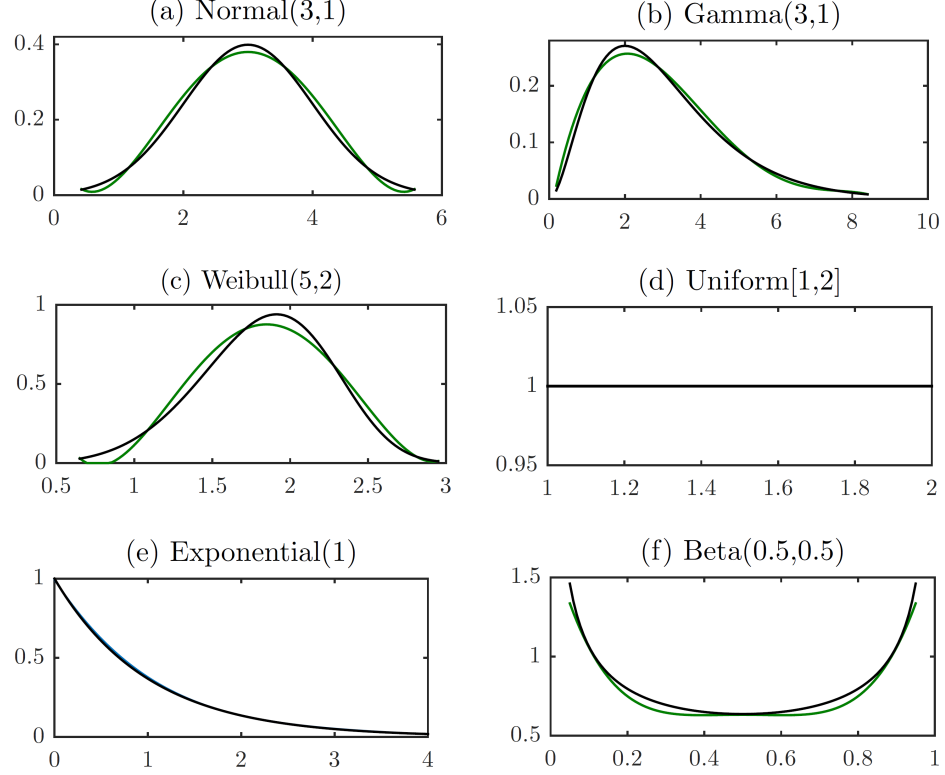


Figure 4.3: Approximations of distributions on  $y \in [a, b]$  using the derivative of the polynomial  $p(x) = x^3 + c_0(x - x^3) + (3 - 2c_0 - c_1)(x^2 - x^3) + c_2(x^2 + x^4 - 2x^3) + c_3(2x^2 + x^5 - 3x^3)$ , where  $x = \frac{y - a}{b - a}$ ,  $c_0, c_1 \geq 0$ ,  $c_2, c_3 \in \mathbb{R}$ . Black curves represent the distributions and green curves represent the approximations.

### 4.3 Mode based density identification

For a continuous random variable, data in a neighborhood containing the global maximum of the underlying density function are likely to be sampled more frequently. Subsequently, an approximation of a distribution based on the location of the global maximum (the mode for a unimodal density) provides a robust density estimation method when the data range is restricted. The principle assumption for the proposed method is that the mode is contained in the data range. Consider estimating the density function  $f(x, \theta)$ ,  $x \in (-\infty, \infty)$ . Suppose that the data range is restricted to  $[a, b]$  and the corresponding truncated density function is given by  $f_T(x, \theta) = \frac{f(x, \theta)}{F(b) - F(a)}$ ,

$x \in [a, b]$ . Let  $m = \arg \max_{x \in [a, b]} f(x, \theta)$  be the mode. Normalized density functions at  $m$  satisfy  $\frac{f_T(x, \theta)}{f_T(m, \theta)} = \frac{f(x, \theta)}{f(m, \theta)}$  and  $\max_{x \in [a, b]} \frac{f_T(x, \theta)}{f_T(m, \theta)} = 1$ . We estimate the normalized density function using the estimated derivatives of the empirical distribution function. We use the numerical estimate of the area to scale the normalized function to be the probability density function of the distribution.

While one can use the proposed method for any parametric distribution, we utilize the function  $\Psi(x_i) \triangleq \frac{\zeta_i e^{-4\alpha_i(x_i - \gamma_i)}}{(1 + e^{-\beta_i(x_i - \gamma_i)})^4}$  to estimate the density, where  $\zeta_i > 0$ ,  $\alpha_i, \beta_i \geq 0$ , and  $\gamma_i \in \mathbb{R}$ . For  $\eta_i \triangleq \alpha_i/\beta_i \in (0, 1)$ ,  $\max \Psi(x_i) = \zeta_i (1 - \eta_i)^{4(1-\eta_i)} \eta_i^{4\eta_i}$  at  $x = \gamma + \ln \left( \frac{1 - \eta_i}{\eta_i} \right)^{1/\beta_i}$ . Function  $\Psi(x)$  approximates well many distributions within the exponential family (see Figure 4.4).

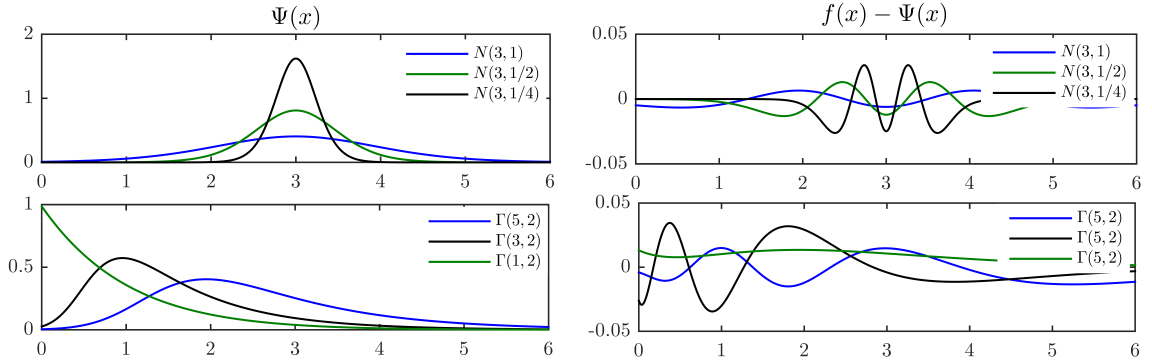


Figure 4.4: Approximations of distributions in the exponential family by  $\Psi(x)$ . In the left column,  $N(\mu, \sigma)$  and  $\Gamma(k, \theta)$  represent the approximations  $\Psi(x)$  for the normal and gamma distributions, respectively; in the right column, they represent the error  $f(x) - \Psi(x)$ , where  $f(x)$  denotes the true distribution. The absolute error in the approximations is less than 0.05 for all the distributions.

## CHAPTER V

### Examples

In this chapter we discuss applications of the introduced density identification methods for ecological data analysis.

#### 5.1 Example 1

In the first example, we estimate densities for an Indiana bat maternity roost selection dataset (given in [28]). From this dataset, we only consider 8 selected continuous variables: tree height (m), distance to forest edge (m), distance to water (km), tree diameter (cm), percentage of peeling bark, canopy opening (percentage variable), distance between maternity colonies (km) and potential maternity colony habitat ( $m^2$ ). The estimated densities can be further used for statistical analysis, such as hypothesis testing.

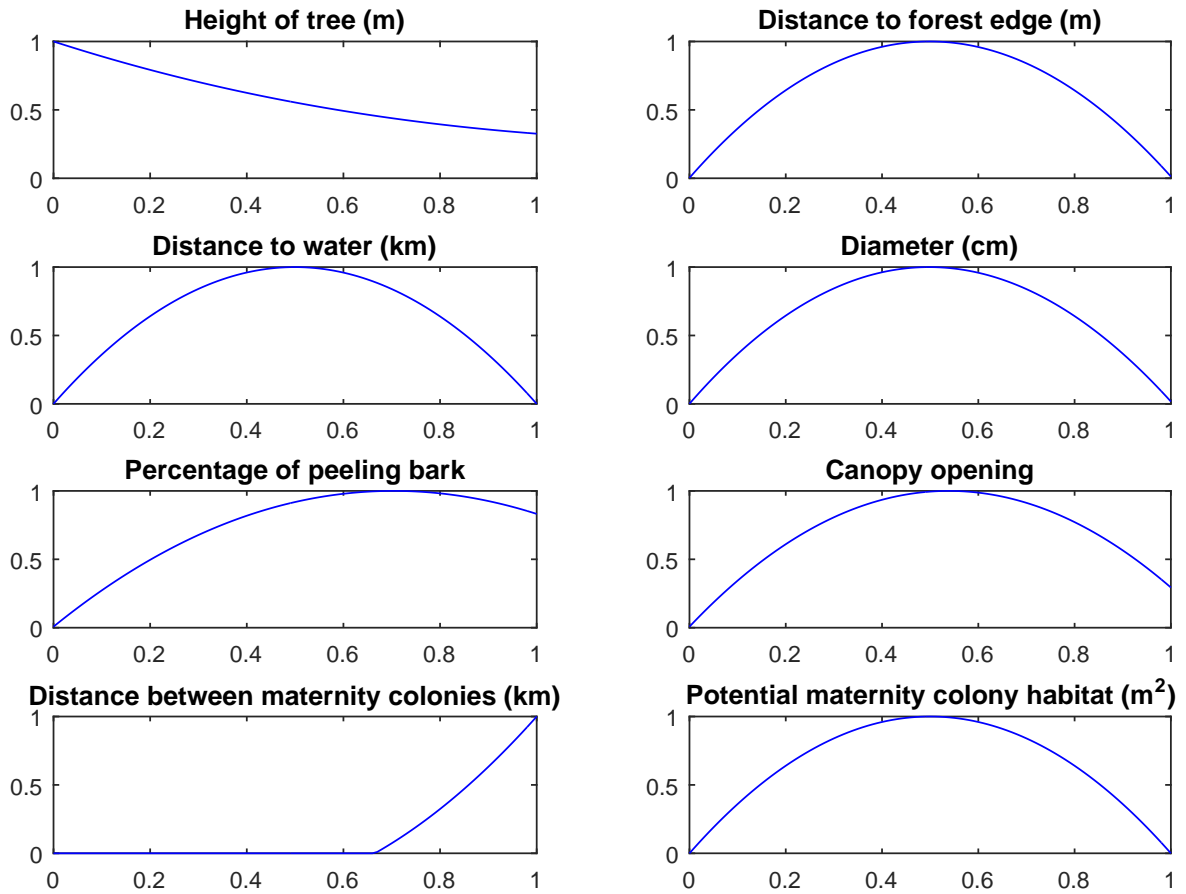


Figure 5.1: Polynomial approximations of distributions on  $y \in [a, b]$  using the derivative of the polynomial  $p(x) = x^3 + c_0(x - x^3) + (3 - 2c_0 - c_1)(x^2 - x^3) + c_2(x^2 + x^4 - 2x^3) + c_3(2x^2 + x^5 - 3x^3)$ , where  $x = \frac{y - a}{b - a}$ ,  $c_0, c_1 \geq 0$ ,  $c_2, c_3 \in \mathbb{R}$ .

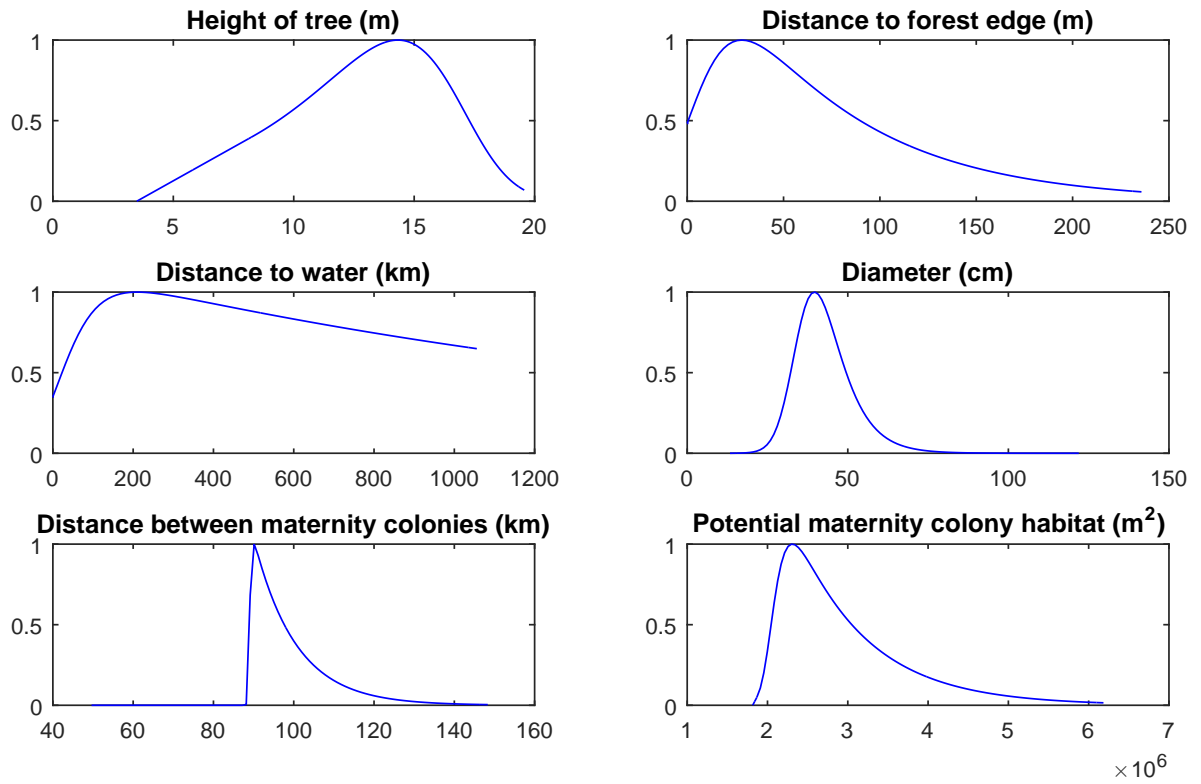


Figure 5.2: Approximations of distributions in the exponential family by  $\Psi(x)$  for every variable that is not a percentage.

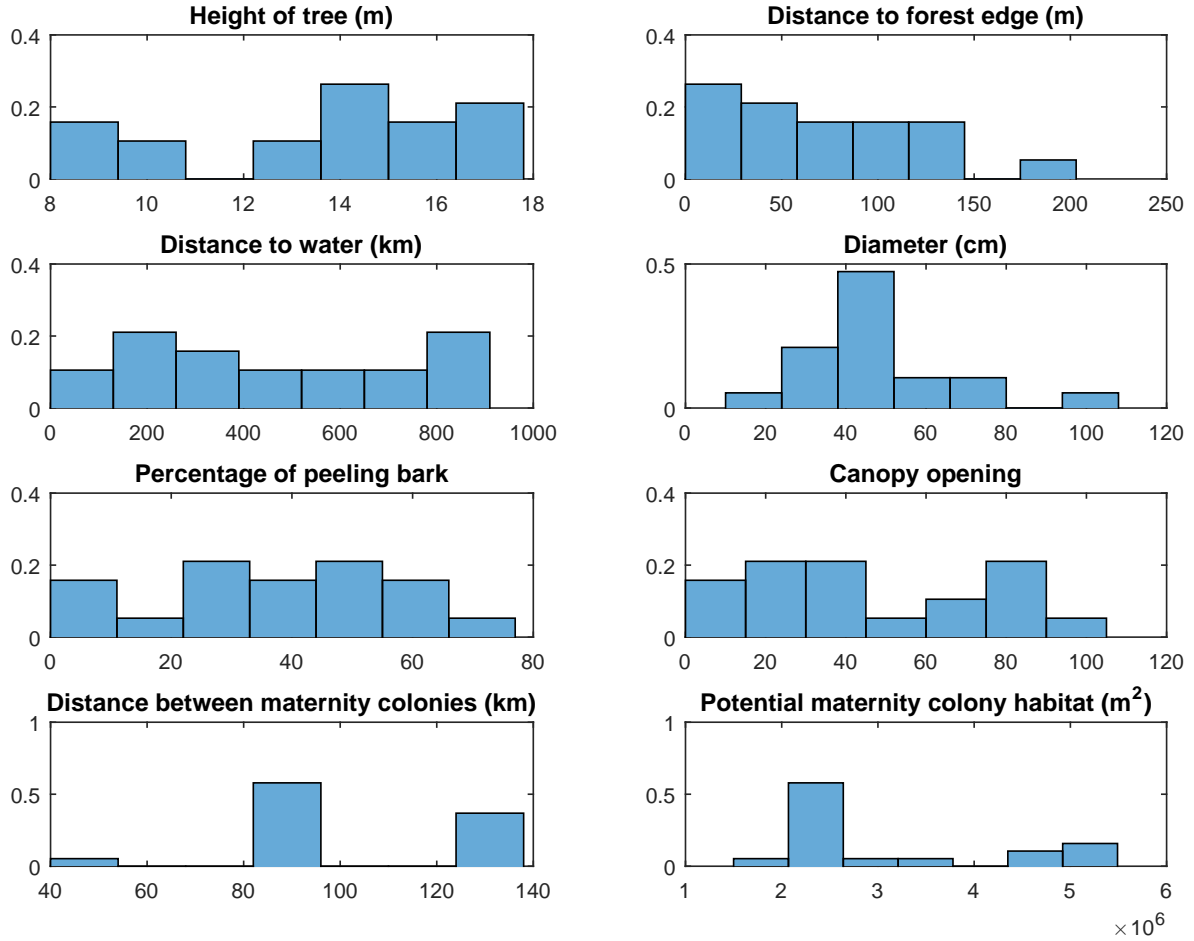


Figure 5.3: Approximations of distributions by histograms.

From Figures 5.2 and 5.3, the approximations of the distributions by  $\Psi(x)$  have similar shapes to the histogram approximations. Also, the positions of most of the modes in Figure 5.2 belong to the bins with the largest relative frequencies in Figure 5.3. The polynomial approximations for percentage data are similar to the shape of the histogram approximation.

## 5.2 Example 2

In the second example, we construct the densities for bald eagle nesting habitats in the Upper Mississippi River National Wildlife and Fish Refuge data (given in [19]).

The dataset is comprised of four variables: tree diameter at breast height (inches) , tree height (ft), nest height (ft), and distance to water(m). The estimated densities can be further used for statistical analysis, such as hypothesis testing.

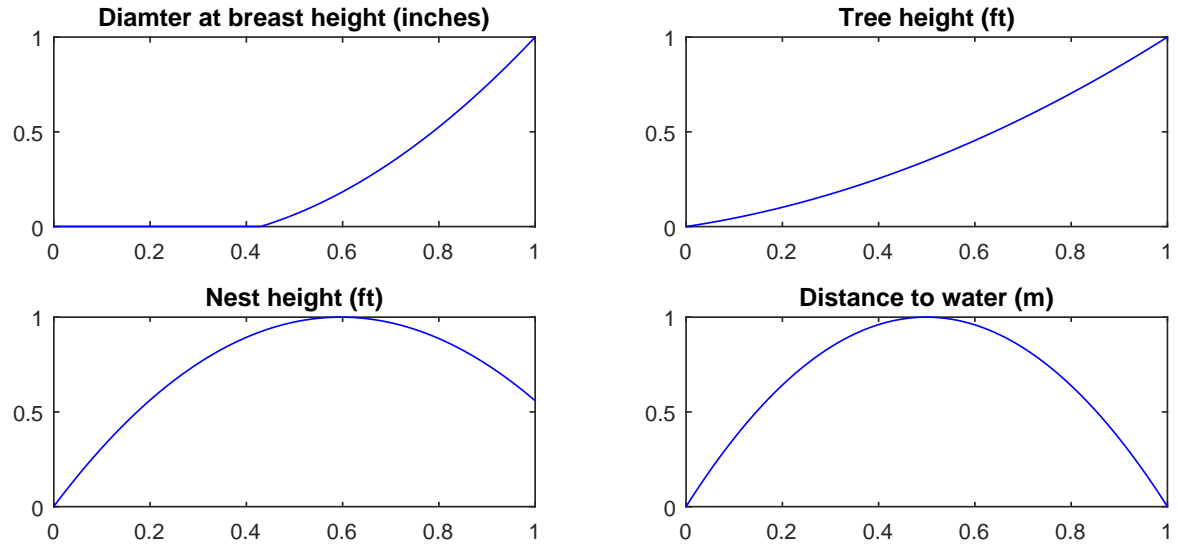


Figure 5.4: Polynomial approximations of distributions on  $y \in [a, b]$  using the derivative of the polynomial  $p(x) = x^3 + c_0(x - x^3) + (3 - 2c_0 - c_1)(x^2 - x^3) + c_2(x^2 + x^4 - 2x^3) + c_3(2x^2 + x^5 - 3x^3)$ , where  $x = \frac{y - a}{b - a}$ ,  $c_0, c_1 \geq 0$ ,  $c_2, c_3 \in \mathbb{R}$ .

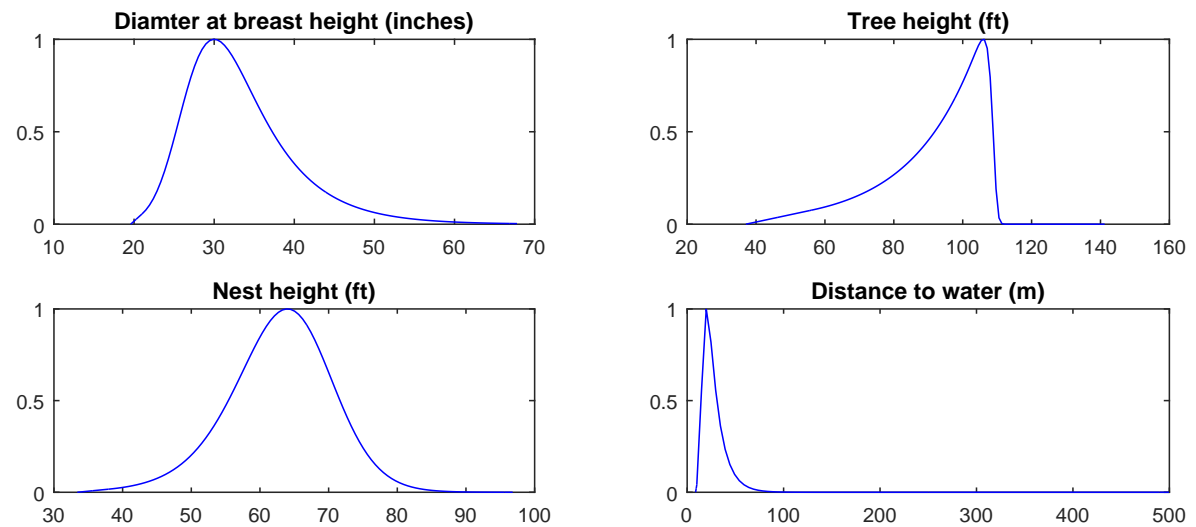


Figure 5.5: Approximations of distributions in the exponential family by  $\Psi(x)$ .

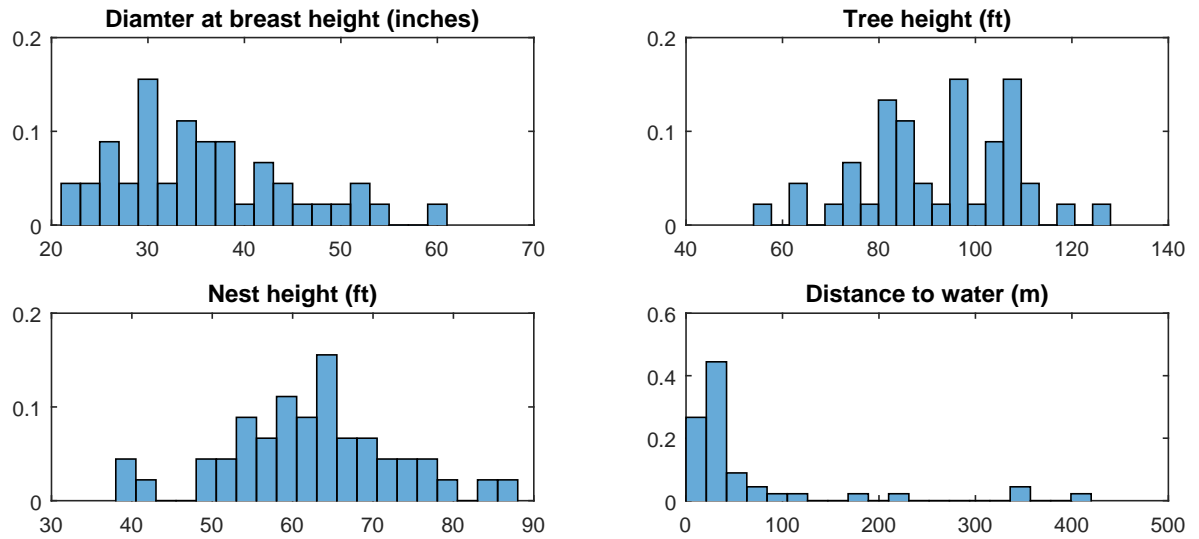


Figure 5.6: Approximations of distributions by histograms.

As before, Figures 5.5 and 5.6 are very comparable to each other.



## CHAPTER VI

### Conclusion

The introduced approximations of distributions in the exponential family by  $\Psi(x)$  is appropriate to model various unimodal density functions for data with restricted ranges. For a continuous random variable, data in a neighborhood containing the global maximum of the underlying density function are likely to be sampled more frequently. Therefore, this method provides a robust density estimation. By using fourth order polynomials, the continuous approximations can be at most bimodal. By using the transformation  $x_i = \frac{y_i - a_i}{b_i - a_i}$ ,  $p_i(x_i)$  is well suited to approximate any unimodal distribution on  $y_i \in [a_i, b_i]$ . But, since the polynomial approximation for the empirical CDF is constrained on the boundary, the density function estimated by the derivative cannot appropriately approximate some bimodal distributions. Unlike traditional sample-based methods, such as histograms, the proposed density estimation methods are well suited when data are drawn from a truncated distribution. Since many ecological data ranges are restricted by geographical or ecological constraints, Therefore, the methods proposed here will be useful in ecological data analysis.

## CHAPTER VII

### Appendix

#### 7.1 R function to implement derivative construction algorithm.

```
DS<-read.csv(file.choose(),header = T) # Importing the dataset and save it as DS
derivative<-function(data){
  A<- data.frame(data)
  A<- apply(A,2,sort,decreasing=F) # Column sorting according to ascending order
  n=length(A[,1]) # Number of rows
  m=length(A[1,]) # Number of columns
  dt<-matrix(0,ncol = m, nrow=n)
  dt<-data.frame(dt) # Construction of data frame for derivatives
  # Derivative estimation using 7 points
  for (i in 4:(n-3)){
    f7<- (-3*A[i-3,] -2*A[i-2,]-A[i-1,]+A[i+1,]+2*A[i+2,]+3*A[i+3,])
    dt[i,]=(28/n)/f7
  }
  # Derivative estimation at boundaries
  # Construction of dt[3,] and dt[n-2,] using 5 points
  # Construction of dt[3,]
  f5=-2*A[1,]-A[2,]+A[4,]+2*A[5,]
  con<-(f5==0)
  c_f5<-sum(con)
  if(c_f5==0){
    dt[3,]=(10/n)/f5 # 5 points derivative estimation
  }else{
    dt[3,]=(10/n)/f5
    dt[3,con]=dt[4,con] # Constant approximation
  }
```

```

}
# Construction of dt[n-2,]
f5=-2*A[n-4,]-A[n-3,]+A[n-1,]+2*A[n,]
con<-(f5==0)
c_f5<-sum(con)
if(c_f5==0){
  dt[n-2,]=(10/n)/f5 # 5 points derivative estimation
}else{
  dt[n-2,]=(10/n)/f5
  dt[n-2,con]=dt[n-3,con] # Constant approximation
}
# Construction of dt[2,] and dt[n-1,] using 3 points
# Construction of dt[2,]
f3=A[3,]-A[1,]
con<-(f3==0)
c_f3<-sum(con)
if(c_f3==0){
  dt[2,]=(2/n)/f3 # 3 points derivative estimation
}else{
  dt[2,]=(2/n)/f3
  dt[2,con]=dt[3,con] # Constant approximation
}
# Construction of dt[n-1,]
f3=A[n,]-A[n-2,]
con<-(f3==0)
c_f3<-sum(con)
if(c_f3==0){
  dt[n-1,]=(2/n)/f3 # 3 points derivative estimation

```

```

}else{
  dt[2,]=(2/n)/f3
  dt[n-1,con]=dt[n-2,con] # Constant approximation
}

# Construction of dt[1,] and dt[n,] using forward difference approximation
# Construction of dt[1,]
f1=(A[2,]-A[1,])
con<-(f1==0)
c_f1<-sum(con)
if(c_f3==0){
  dt[1,]=(1/n)/f1 # 2 points derivative estimation
}else{
  dt[1,]=(1/n)/f1
  dt[1,con]=dt[2,con] # Constant approximation
}

# Construction of dt[n,]
f1=(A[n,]-A[n-1,])
con<-(f1==0)
c_f1<-sum(con)
if(c_f3==0){
  dt[n,]=(1/n)/f1 # 2 points derivative estimation
}else{
  dt[n,]=(1/n)/f1
  dt[n,con]=dt[n-1,con] # Constant approximation
}

return(list(derivatives=dt,Sorted_data=A))
}

derivatives<- derivative(DS)$derivatives # Running the function

```

## BIBLIOGRAPHY

- [1] Bickel, D.R., Fruehwirth R. (2006). On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications. *Computational Statistics and Data Analysis*, Vol.50(12), pp.3500-3530.
- [2] Cantelli, F.P. (1933).Sulla determinazione empirica della legge di probabilita.Giorn.Ist. Ital. Attuari, Vol.4, pp. 421424.
- [3] Carlos, H.A., Shi, X., Sargent, J., Tanski, S., and Berke, E. M. (2010). Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics*, pp. 9-39. URL <http://doi.org/10.1186/1476-072X-9-39>.
- [4] Charles J., Krebs, Rudy, B., Scott, G., Reid D., Kenney, A.j, Hofer, E.J. (2011).Density estimation for small mammals from livetrapping grids: rodents in northern Canada. *Journal of Mammalogy*, Vol 92(5), pp. 974-981.
- [5] Chentsov, N.N. (1962). Estimation of unknown probability density based on observations. In *Dokl. Akad. Nauk SSSR* (Vol. 147, pp. 45-48).
- [6] Croft, S., Chauvenet, A.L.M., Smith, G.C. (2017) A systematic approach to estimate the distribution and total abundance of British mammals. *PLoS ONE*, Vol.12(6), e0176339. URL <https://doi.org/10.1371/journal.pone.0176339>.
- [7] Desforges, M.J.,Jacob, P.J., Cooper, J.E. (1998). Applications of probability density estimation to the detection of abnormal conditions in engineering. *Journal of Mechanical Engineering Science*,Vol. 212(8), pp.687-703. URL <https://doi.org/10.1243/0954406981521448>.
- [8] Dethier, M.N., Graham, E.S., Cohen, S., Tear, L.M. (1993). Visual versus random-point percent cover estimations: 'objective' is not always better. *Marine Ecology Progress Series*, Vol.96(1), pp.93-100.

- [9] Edwards, A.W.F. (1997). Three Early Papers on Efficient Parametric Estimation. *Statistical Science*, Vol.12(1), pp.35-47.
- [10] Elith, J., Graham, C.H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, Vol.32(1), pp.66-77. URL <https://doi.org/10.1111/j.1600-0587.2008.05505.x>
- [11] Ferdosi B. J., Buddelmeijer, H., Trager, S., Wilkinson, M. H. F., and Roerdink, J. B. T. M. (2011). Comparison of Density Estimation Methods for Astronomical datasets. *ASTRON ASTROPHYS.* 531. 10.1051/0004-6361/201116878.
- [12] Hald, A. (1999). On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, Vol. 14(2), pp.214-222.
- [13] Glivenko, V. (1933). Sulla determinazione empirica della legge di probabilita. *Giorn. Ist. Ital. Attuari*, Vol.4, pp. 92-99.
- [14] Hall, P., Presnell, B. (1999). Density Estimation under Constraints. *Journal of Computational and Graphical Statistics*, 8(2), pp.259-277, DOI: 10.1080/10618600.1999.10474813.
- [15] Hogg, R., McKean, J.W., Joseph W., and Craig, A.T., (2013) "Introduction to Mathematical Statistics". Pearson.
- [16] Kumar, P., Mina, U. (2018). *FUNDAMENTALS OF ECOLOGY AND ENVIRONMENT*. Pathfinder Publication, New Delhi, India. ISBN 978-81-934655-0-9.
- [17] Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., Wasserman, L. (2011). Forest Density Estimation. *The Journal of Machine Learning Research*, Vol 12, pp.907-951.
- [18] Lopez-Rubio, E. (2014) A Histogram Transform for Probability Density Function Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36(4), pp.644-656.

- [19] Mundahl, N., Bilyeu, A., Maas, L. (2013). Bald eagle nesting habitats in the upper mississippi river national wildlife and fish refuge. *Journal of Fish and Wildlife Management* 4, 362-376. doi:10.3996/012012-JFWM-009.
- [20] Newey, W.K., McFadden D. (1994). *Handbook of Econometrics: Chapter 36 Large sample estimation and hypothesis testing*. Elsevier, Vol 4., pp. 2113-2245.
- [21] Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3), 1065-1076.
- [22] Phillips, S., Anderson, R., Schapire., R., (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, Vol.190, pp. 231-259, doi:10.1016/j.ecolmodel.2005.03.026.
- [23] Rajwade, A., Banerjee, A., Rangarajan, A. (2006). A New Method of Probability Density Estimation with Application to Mutual Information Based Image Registration. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2(1640968), pp.17691776. URL <http://doi.org/10.1109/CVPR.2006.206>.
- [24] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832-837.
- [25] Sain, S. R. and Scott, D.W. (1996). On Locally Adaptive Density Estimation. In *J. Ame. Stat. Asso.*, Vol. 91, pp.15251534.
- [26] Savitzky, A., Golay M.J.E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, Vol. 36(8), pp.16271639, DOI: 10.1021/ac60214a047.
- [27] Segurado P., Araujo M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, Vol.31(10):pp.1555-1568.

- [28] Schroder, E., Ekanayake, D., Romano, S. (2018). Data for: Indiana bat maternity roost habitat preference within midwestern united states upland oak-hickory forests. Mendeley Data. doi:10.17632/djh4r2m5xd.1.
- [29] Silverman, B.W. (1986). Density estimation for statistics and data analysis. Chapman and Hall, London. <http://dx.doi.org/10.1007/978-1-4899-3324-9>.
- [30] Vikram, V., Garg, Luis, Tenorio, Willcox, K. (2014). Minimum Local Distance Density Estimation. Minimum Local Distance Density Estimation, Vol 46(1).
- [31] Xu, L., Bedrick, E.J., Hanson, T., Restrepo, C. (2014). A Comparison of Statistical Tools for Identifying Modality in Body Mass Distributions. Journal of Data Science. Vol 12, pp.175-196.
- [32] Zambom, A., Dias, R. (2013). A review of kernel density estimation with applications to econometrics. International Econometric Review, Vol. 5(1), pp.20-42.