# On Nonparametric Bayesian Inference for Tukey Depth

Xuejun Han

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment
of the requirements for the degree of
Master of Science in Mathematics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

---

[1] The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

# Abstract

The Dirichlet process is perhaps the most popular prior used in the nonparametric Bayesian inference. This prior which is placed on the space of probability distributions has conjugacy property and asymptotic consistency. In this thesis, our concentration is on applying this nonparametric Bayesian inference on the Tukey depth and Tukey median. Due to the complexity of the distribution of Tukey median, we use this nonparametric Bayesian inference, namely the Lo's bootstrap, to approximate the distribution of the Tukey median. We also compare our results with the Efron's bootstrap and Rubin's bootstrap. Furthermore, the existing asymptotic theory for the Tukey median is reviewed. Based on these existing results, we conjecture that the bootstrap sample Tukey median converges to the same asymptotic distribution and our simulation supports the conjecture that the asymptotic consistency holds.

# Dedications

I dedicate my dissertation work to my parents.

Thank you deeply for the support throughout the process.

# Acknowledgement

I would first like to thank my thesis supervisor Prof. Mahmoud Zarepour, without whom the thesis was not possible to be completed. A sincere gratitude for his continual guidance and inspiration throughout the entire process, as well as his patience and kindness to be available anytime even on weekends. He also gives me much valuable advice on my future study, which encourages me a lot.

I wish to acknowledge and thank the Department of Mathematics and Statistics for providing me with a great academic environment and a variety of resources during my research process. The staff here and the colleagues in lab B03 are all very kind and warm-hearted and offered me a lot of help. It is really my good fortune to be in this big family.

Finally, I must express my profound gratitude to my parents for their unfailing support and continuous encouragement throughout my years of study and the research process. This accomplishment would not have been possible without them. Thank you deeply.

# Contents

# List of Figures

# Chapter 1

# Introduction

A statistical model on a sample space $\mathcal{S}$ means a set of probability distributions on $\mathcal{S}$, denoted by $\mathcal{P}$. The set $\mathcal{P}$ can be parameterized by $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, where $\mathcal{P}$ is indexed by the parameter $\theta$ and $\Theta$ is the parameter space of the model. If the parameter space $\Theta$ has finite dimensions, we call the model $\mathcal{P}$ is parametric. Otherwise, $\mathcal{P}$ is a nonparametric model. A probability distribution on an infinite dimensional space $\Theta$ is a stochastic process with paths in $\Theta$. In Bayesian statistics, we assume a prior on $\Theta$ based on some known ideas and make inferences from the posterior distribution. The nonparametric Bayesian model is a nonparametric model in the Bayesian framework. This is always involved with a question how to choose a prior. The choice of a prior has been discussed extensively in the nonparametric Bayesian inference. Typically, the priors are constructed to make sure that the posterior distribution is analytically tractable and we can draw samples from the prior and the posterior distributions. In practice, calculating the posterior distribution is difficult analytically, however the conjugate prior makes it tractable most of the time. The conjugacy means that the posterior distribution given the observed data is in the same distribution family as the prior distribution. The conjugate priors are well-used in the nonparametric Bayesian inference.

In this thesis, we give a detailed introduction to a classic prior which is the Dirichlet process, first introduced by Ferguson (1973). The Dirichlet process is a conjugate prior for estimating a completely unknown distribution based on a random sample. This tool is widely applied in some practical areas, such as data mining and machine leaning. The approximations and constructions of the Gamma process and Dirichlet process are discussed in Chapter 2. An example of how to estimate the probability distribution of a sample using the Dirichlet process is given in Chapter 2.

In Chapter 3, we introduce an important concept in multivariate analysis, namely, Data Depth, which is a useful tool in multivariate data analysis. It provides a possible way to order the multivariate data and obtain some characteristics of the given data. A variety of depth functions were introduced in the literature, including Mahalanobis depth (Mahalanobis, 1936), Tukey depth or half-space depth (Tukey, 1975), convex hull peeling depth (Barnett, 1976), simplicial depth (Liu, 1990), Oja depth (Oja, 1985), majority depth (Singh, 1991) and some other new notions. The difference between these depth functions are shown graphically by contours and bagplots. The bagplot was proposed by Rousseeuw (1999), which is considered as a bivariate generalization of the univariate boxplot.

In Chapter 4, we employ different types of bootstrap methods and Tukey depth to draw inference on the Tukey median. Four bootstrap methods are discussed here: Efron's bootstrap, Rubin's Bayesian bootstrap, Lo's Bayesian bootstrap and Muliere and Secchi's Bootstrap, with corresponding simulation study for the Tukey median. It is notable that the Dirichlet process prior is used in Lo's Bayesian bootstrap and Muliere and Secchi's Bootstrap.

Finally, we review some existing results on asymptotic properties of the Tukey Median in Chapter 5 mainly based on Nolan (1999) and Massé (2001) and conjecture that the bootstrap Tukey median converges to the same limiting distribution. The problem of proving the consistency is discussed in Chapter 5, which should be considered as the future work.

# Chapter 2

# Gamma Process and Dirichlet Process

## 2.1 Dirichlet Distribution

In probability and statistics, the Dirichlet distribution (named after Peter Gustav Lejeune Dirichlet) is a family of continuous multivariate probability distributions parameterized by a vector $\boldsymbol{\alpha}$ of positive real numbers. It is a multivariate generalization of Beta distribution. Dirichlet distribution is often used as a conjugate prior distribution for multinomial distribution in Bayesian inference.

**Definition 2.1.1.** *Let* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ *with* $\alpha_1, \ldots, \alpha_K > 0$, $\boldsymbol{p} = (p_1, \ldots, p_K)$. *Then,* $\boldsymbol{p}$ *is said to have Dirichlet distribution of order* $K \geq 2$ *with parameter* $\boldsymbol{\alpha}$, *denoted by* $\boldsymbol{p} \sim Dir\left(\alpha_1, \ldots, \alpha_K\right)$, *if it has a probability density function with respect to Lebesgue measure on the Euclidean space* $\mathbb{R}^{K-1}$ *given by*

$$f\left(p_1, \ldots, p_{K-1}; \alpha_1, \ldots, \alpha_K\right) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma\left(\alpha_i\right)} \prod_{i=1}^{K-1} p_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{K-1} p_i\right)^{\alpha_k - 1},$$

*on the open* $(K-1)$*-dimensional simplex defined by:*

$$p_1, \ldots, p_{K-1} \geq 0$$

$$p_1 + \cdots + p_{K-1} \leq 1$$

$$p_K = 1 - p_1 - \cdots - p_{K-1}$$

*and zero elsewhere.*

If any $\alpha_i = 0$, we still define a Dirichlet distribution by setting the corresponding $p_i = 0$ and interpreting the density above as a density on a lower-dimensional set.

The following are some main properties of the Dirichlet distribution.

1. Let $Z_1, \ldots, Z_K$ be independent gamma random variables with parameter $\alpha_i \geq 0$, then we have
$$\left( \frac{Z_1}{\sum_{i=1}^{K} Z_i}, \ldots, \frac{Z_k}{\sum_{i=1}^{K} Z_i} \right) \sim Dir\left(\alpha_1, \ldots, \alpha_K\right).$$

2. If $(p_1, \ldots, p_K) \sim Dir\left(\alpha_1, \ldots, \alpha_K\right)$ and $r_1, \ldots, r_l$ are integers such that $0 < r_1 < \ldots < r_l = K$, then
$$\left( \sum_{i=1}^{r_1} p_i, \sum_{i=r_1+1}^{r_2} p_i, \ldots, \sum_{i=r_{l-1}+1}^{r_l} p_i \right) \sim Dir\left( \sum_{i=1}^{r_1} \alpha_i, \sum_{i=r_1+1}^{r_2} \alpha_i, \ldots, \sum_{i=r_{l-1}+1}^{r_l} \alpha_i \right).$$

3. The marginal distribution of each $p_i$ for $i = 1, \ldots, K$ is Beta distribution:
$$p_i \sim Beta\left(\alpha_i, \alpha_0 - \alpha_i\right)$$
where $\alpha_0 = \sum_{j=1}^{K} \alpha_j$.

4. If $(p_1, \ldots, p_K) \sim Dir\left(\alpha_1, \ldots, \alpha_K\right)$, then
$$\begin{aligned} E[p_i] &= \frac{\alpha_i}{\alpha_0} \\ Var[p_i] &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$
Furthermore, if $i \neq j$,
$$Cov[p_i, p_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}.$$

5. If the prior distribution of $(p_1, \dots, p_K)$ is $Dir(\alpha_1, \dots, \alpha_K)$ and if $P\{X = j | p_1, \dots, p_K\} = p_j$ a.s. for $j = 1, \dots, K$, then the posterior distribution of $(p_1, \dots, p_K)$ given $X = j$ is $Dir(\alpha_1^{(j)}, \dots, \alpha_K^{(j)})$, where

$$\alpha_i^{(j)} = \begin{cases} \alpha_i & if \quad i \neq j \\ \alpha_j + 1 & if \quad i = j. \end{cases}$$

## 2.2 Gamma Process

### 2.2.1 Definition

**Definition 2.2.1.** *Let $(\mathbb{X}, \mathcal{X})$ be a Polish space equipped with the corresponding Borel $\sigma$-algebra. A measure $\mu$ on $\mathbb{X}$ is said to be boundedly finite if $\mu(A) < +\infty$ for any bounded measurable set $A \in \mathcal{X}$. Denote by $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ the space of boundedly finite measures equipped with the corresponding Borel $\sigma$-algebra. Let a random measure (RM) $\mu$ be a measurable mapping from $(\Omega, \mathcal{F}, P)$ into $(\mathbb{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$. If for any $A_1, \dots, A_n$ in $\mathcal{X}$, with $A_i \cap A_j = \emptyset$ for $i \neq j$, the random variables $\mu(A_1), \dots, \mu(A_n)$ are mutually independent, then $\mu$ is called completely random measure (CRM).*

**Definition 2.2.2.** *Let $\mathbb{E}$ be a locally compact space with a countable basis. A random measure $\mu$ on $\mathbb{E}$ of the form: $\mu = \sum_{n=1}^{N} \delta_{X_n}$, where $\delta$ is the Dirac measure, $X_n$ are $\mathbb{E}$-valued random variables and $N$ is a random variable with values in $\{0, 1, \dots, \infty\}$, is called a point process or random counting measure.*

**Proposition 2.2.3.** *The gamma distribution, Gamma$(\alpha, 1)$, $\alpha > 0$, has the characteristic function*

$$\varphi(t) = (1 - it)^{-\alpha} = \exp \int_0^\infty (e^{iux} - 1) dN(x),$$

*where the Gamma Lévy measure $N(x) = \alpha \int_x^\infty e^{-y} y^{-1} dy$, $x > 0$.*

**Definition 2.2.4.** *Let $H$ be a probability distribution on a measurable space $(\Theta, \Sigma)$ with corresponding Borel $\sigma$-algebra and $\alpha$ a positive scalar. A completely random measure $\mathcal{G}$ on*

$\Omega \times \Theta$ *is called a Gamma process, denoted by* $\mathcal{G} \sim GP(\alpha, H)$*, with Lévy measure*

$$\nu(dx, d\theta) = N(dx)dH(\theta),$$

*where* $N(dx) = \alpha x^{-1} e^{-x} dx$*, and* $\alpha$ *is called the concentration parameter and* $H$ *is the base measure.*

Gamma process is a pure-jump nondecreasing Lévy process with independent gamma distributed increments, here, defined as $Gamma(\alpha, 1)$. The jumps occurring in the interval $[x, x+dx]$ follow the distribution $H$ and the length of each jump follows $Gamma(\alpha, 1)$. From the proposition above, we have

$$N(x) = \alpha \int_x^\infty e^{-y} y^{-1} dy, \quad x > 0.$$

**Theorem 2.2.5.** *The Gamma process* $\mathcal{G} \sim GP(\alpha, H)$ *can be defined as a sum representation based on Gamma Lévy measure by:*

$$\mathcal{G}^{Ferg}(\cdot) = \sum_{i=1}^{\infty} N^{-1}(\Gamma_i) \delta_{\theta_i}(\cdot), \tag{2.2.1}$$

*where* $\{\theta_i, i = 1, 2, \ldots\}$ *is a sequence of i.i.d. random variables with distribution* $H$ *and the distributions of* $\Gamma_i$ *are defined as follows:*

$$\Gamma_i = E_1 + E_2 + \cdots + E_i \tag{2.2.2}$$

*where* $E_1, \ldots, E_i \overset{i.i.d.}{\sim} \mathcal{E}xp(1)$*.*

This representation in (2.2.1) contains the inverse of the Gamma Lévy measure which is difficult to calculate. However, Bondesson (1982) introduced an alternative sum representation avoiding to calculate the inverse of Gamma Lévy measure, which is

$$\mathcal{G}^{Bond}(\cdot) = \sum_{i=1}^{\infty} \exp\left(-\frac{\Gamma_i}{\alpha}\right) E_i \delta_{\theta_i}(\cdot). \tag{2.2.3}$$

### 2.2.2 Approximation of Gamma Process

The Bondesson (1982) sum representation of Gamma process can be approximated by truncating the higher order in (2.2.3) using the stopping rule

$$n = \inf\{i : \exp\left(-\frac{\Gamma_i}{\alpha}\right) E_i < \epsilon\},$$

The approximation of Bondesson (1982) sum representation is written as

$$\mathcal{G}_n^{Bond}(\cdot) = \sum_{i=1}^{n} \exp\left(-\frac{\Gamma_i}{\alpha}\right) E_i \delta_{\theta_i}(\cdot). \qquad (2.2.4)$$

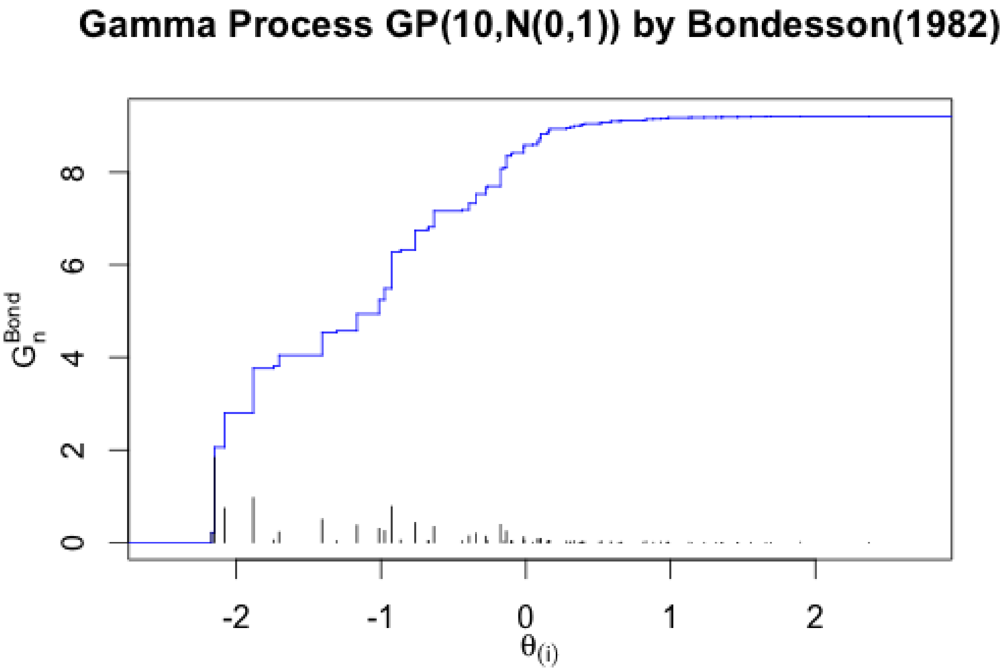


Figure 2.1: One sample path of the Gamma process approximated using the Bondesson (1982) approximation in (2.2.4). We choose $a = 10$, $H = N(0, 1)$. For $\epsilon = 0.0001$, we get $n = 74$. Vertical lines represent weights for each $\theta_{(i)}, i = 1, \ldots, n$.

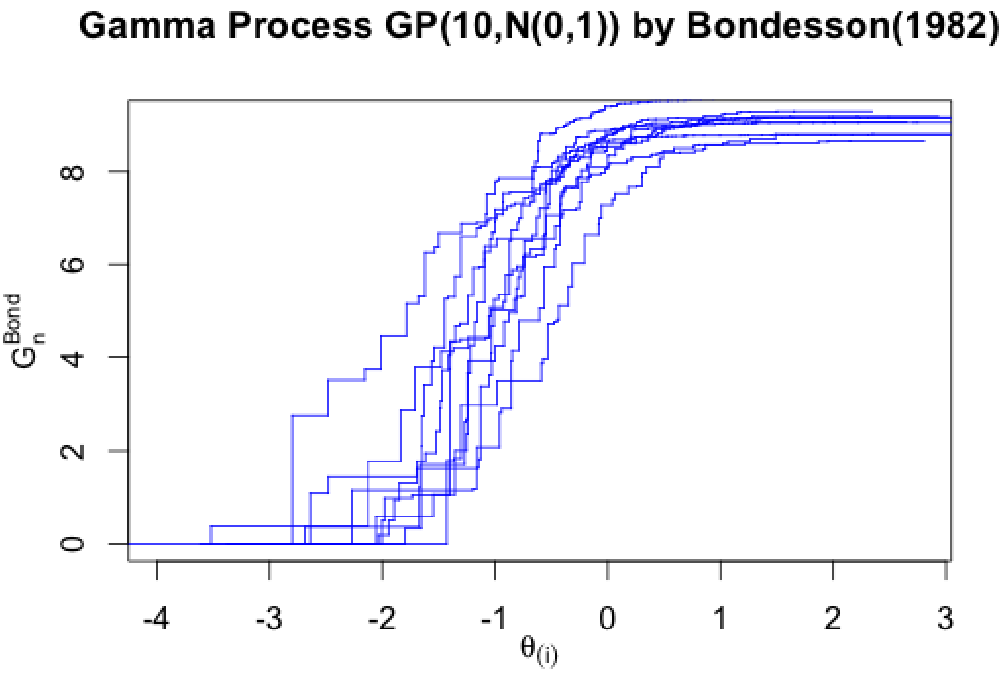


Figure 2.2: Ten sample paths of the Gamma process approximated by using the Bondesson(1982) Gamma process approximation in (2.2.4). We choose $a = 10$, $H = N(0,1)$. For $\epsilon = 0.0001$, we get $n = 60, 67, 60, 36, 51, 73, 94, 44, 63, 48$.

Zarepour and Al Labadi (2012) derive a finite sum representation of Gamma process $\mathcal{G}_n$, which almost surely converges to the Ferguson (1973) sum representation. To show this, let $X_n \sim Gamma(\frac{\alpha}{n}, 1)$, i.e.:

$$G_n(x) = P(X_n > x) = \int_x^\infty \frac{1}{\Gamma(\frac{\alpha}{n})} e^{-t} t^{\frac{\alpha}{n}-1} dt, \tag{2.2.5}$$

and

$$G_n^{-1}(y) = \inf\{x : G_n(x) \leq y\}, \quad 0 < y < 1. \tag{2.2.6}$$

Note that $\frac{n}{\Gamma(\frac{\alpha}{n})} = \frac{\alpha}{\Gamma(\frac{\alpha}{n}+1)}$ and let $n \to \infty$, we have $\frac{n}{\Gamma(\frac{\alpha}{n})} \to \alpha$. Thus, for $x > 0$,

$$nG_n(x) = \frac{n}{\Gamma(\frac{\alpha}{n})} \int_x^\infty e^{-t} t^{\frac{\alpha}{n}-1} dt \to \alpha \int_x^\infty e^{-t} t^{-1} dt = N(x).$$

For every $x > 0$, $nG_n(x)$ is a sequence of monotone functions and so is $N(x)$. Therefore,

$$G_n^{-1}\left(\frac{x}{n}\right) \to N^{-1}(x).$$

By the Strong Law of Large Numbers, we have $\frac{\Gamma_{n+1}}{n} \xrightarrow{a.s.} 1$ as $n \to \infty$. Taking $x = \Gamma_i$, we have

$$G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right) \xrightarrow{a.s.} N^{-1}(\Gamma_i).$$

Thus, the approximation of Gamma process can be defined as follows:

$$\mathcal{G}_n^{Zar\&Al-Lab}(\cdot) = \sum_{i=1}^{n} G_n^{-1}\left(\frac{\Gamma_i}{\Gamma_{n+1}}\right) \delta_{\theta_i}(\cdot) \xrightarrow{a.s.} \mathcal{G}(\cdot) = \sum_{i=1}^{\infty} N^{-1}(\Gamma_i)\delta_{\theta_i}(\cdot). \tag{2.2.7}$$

Note that the weights here, i.e. $G_n^{-1}(\frac{\Gamma_i}{\Gamma_{n+1}})$ are monotonically decreasing almost surely. The approximation in (2.2.7) introduced by Zarepour and Al Labadi (2012) is an asymptotic result which is not based on the stopping rule. However, we can also define the stopping rule as follows:

$$n = \inf\{i : G_i^{-1}\left(\frac{\Gamma_i}{\Gamma_{i+1}}\right) < \epsilon\}.$$

For comparison purpose, we will calculate the value of $n$ using Zarepour and Al Labadi (2012) approximation, which is smaller than that in Bondesson (1982) approximation. The following describes the algorithm to draw samples from Gamma process by using Zarepour and Al Labadi (2012) approximation method:

1. Computing $G_n^{-1}(p)$ is equivalent as computing the quantile of the gamma distribution $Gamma(a/n, 1)$ evaluated at $1 - p$.

2. Find the value of $n$ given a tolerance value $\epsilon$ by using the stopping rule: $n = \inf\{j : G_j^{-1}(\Gamma_j/\Gamma_{j+1}) < \epsilon\}$. Simultaneously, we get the values of $\Gamma_i$, for

$i = 1, \ldots, n+1$

3. Generate $\theta_i \overset{i.i.d.}{\sim} H$, for $i = 1, \ldots, n$
4. Give each $\theta_i$ a weight of $G_i^{-1}(\Gamma_i/\Gamma_{n+1})$, then draw a path by ordering $\theta_i, i = 1, \ldots, n$.

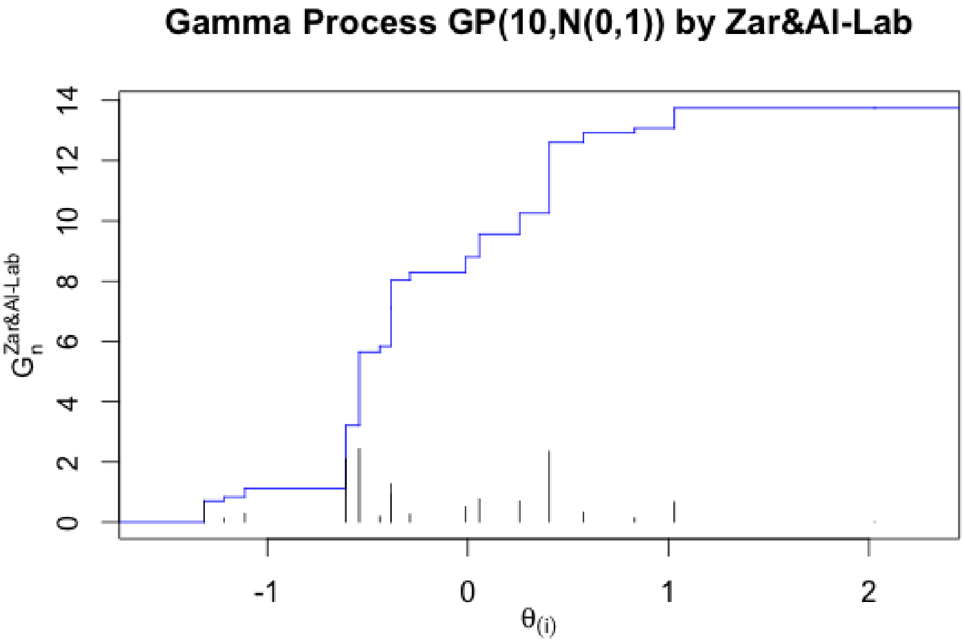


Figure 2.3: One sample path of the Gamma process approximated using the Zarepour and Al Labadi (2012) Gamma process approximation in (2.2.7). We choose $a = 10$, $H = N(0,1)$. For $\epsilon = 0.0001$, we get $n = 21$. Vertical lines represent weights for each $\theta_{(i)}, i = 1, \ldots, n$.

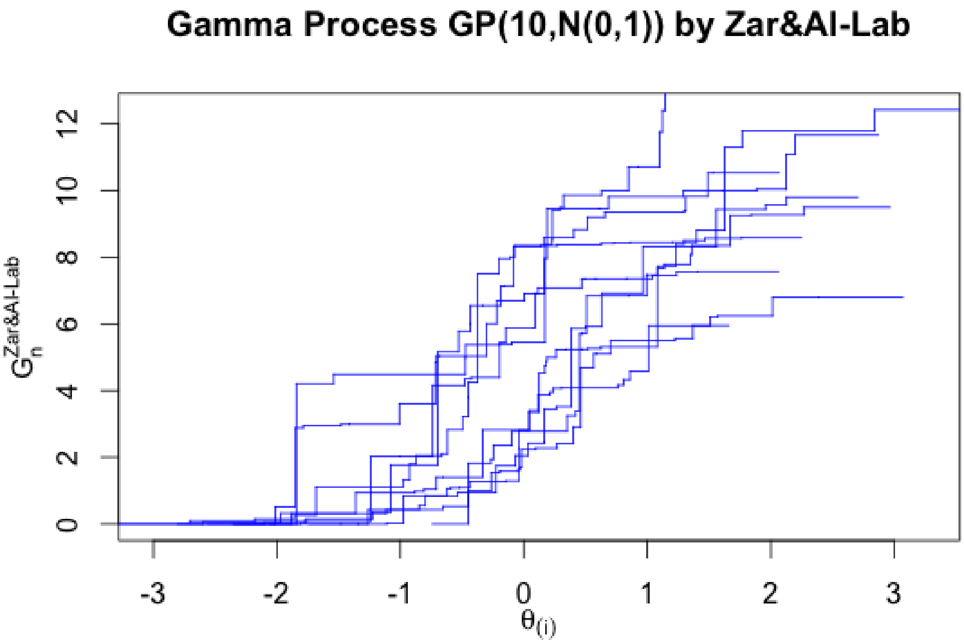


Figure 2.4: Ten sample paths of the Gamma process approximation with $a = 10$, $H = N(0, 1)$ and $\epsilon = 0.0001$. By using the Zarepour and Al Labadi (2012) Gamma process approximation in (2.2.7), the truncation values here are $n = 27$, 25, 16, 19, 19, 19, 18, 21, 25, 21, for the same tolerance value $\epsilon = 0.0001$

## 2.3 Dirichlet Process

### 2.3.1 Definition

In probability theory, the Dirichlet process is a family of stochastic processes whose draws are probability distributions. Just like the Dirichlet distribution which is conjugate for the multinomial distribution, the Dirichlet process prior is also conjugate for estimating a completely unknown distribution from a sample. The Dirichlet process is an extension of the Dirichlet distribution to the infinite-dimensional space. The Dirichlet process was formally introduced by Ferguson (1973) and has been extensively used in Bayesian nonparametric statistics to describe the prior knowledge about the unknown distributions of interest.

**Definition 2.3.1.** *Let $H$ be a probability distribution on a measurable space $(\Theta, \Sigma)$ and $\alpha$ a positive scalar. A random probability measure $G$ defined on $\Theta$ is called a Dirichlet process ($DP$) with parameters $\alpha$ and $H$, denoted by $G \sim DP(\alpha, H)$, if for every finite measurable partition $\{A_1, A_2, \cdots, A_K\}$ on $\Theta$, $(G(A_1), \cdots, G(A_K))$ follows a k-dimensional Dirichlet distribution:*

$$(G(A_1), \cdots, G(A_K)) \sim Dir(\alpha H(A_1), \cdots, \alpha H(A_K)).$$

Note that there is a unique stochastic process satisfying these conditions for any base measure $H$ and concentration parameter $\alpha$.

**Remark:** Notice that the random measure $G$ depends on $\alpha$ and for convenience, we drop $\alpha$, i.e., $G = G_\alpha$.

**Proposition 2.3.2.** *Let $G$ be a Dirichlet process on $(\Theta, \Sigma)$ with parameters $\alpha$ and $H$. For any measurable set $A \in \Sigma$, $G(A) \sim Beta(\alpha H(A), \alpha(1 - H(A)))$. Thus,*

$$E(G(A)) = \frac{\alpha H(A)}{\alpha H(A) + \alpha(1 - H(A))} = H(A)$$

$$Var(G(A)) = \frac{H(A)(1 - H(A))}{\alpha + 1}.$$

The Dirichlet process is specified by a base distribution $H$ and a positive real number $\alpha$ which is called the concentration parameter. The base distribution is the expectation of the process, i.e., the distributions drawn from the Dirichlet process is centred at the base distribution. However, the distributions drawn from the Dirichlet process are almost surely discrete with probability 1 even though the base distribution is continuous. The concentration parameter $\alpha$ determines how many jumps of the realization. Namely, with the decrease of the value of $\alpha$, the distributions drawn from the Dirichlet process become more and more discrete. In the limit of $\alpha \to 0$, the realizations are all concentrated at a single value, and in the limit of $\alpha \to \infty$ the realizations become continuous approaching the base distribution $H$. More explicitly, for any fixed measurable set $A \in \Sigma$ and $\epsilon > 0$, we have $G \xrightarrow{p} H$ as $\alpha \to \infty$ since

$$P\left(|G(A) - H(A)| > \epsilon\right) \leq \frac{H(A)(1 - H(A))}{\epsilon^2(1+\alpha)} \to 0 \qquad (\alpha \to \infty)$$

**Theorem 2.3.3.** *Let $G$ be a Dirichlet process on $(\Theta, \Sigma)$ with parameters $\alpha$ and $H$, and let $X_1, \cdots, X_n$ be a sample of size $n$ from $G$. Then the posterior distribution $G|X_1, \cdots, X_n \sim DP(\alpha^*, H^*)$, where*

$$\begin{aligned} \alpha^* &= \alpha + n \\ H^* &= \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} \frac{\sum_{i=1}^{n} \delta_{X_i}}{n}. \end{aligned}$$

The above theorem confirms that the Dirichlet process is a conjugate prior. From the above we can see that the posterior distribution given a group of samples is approaching the real distribution as the sample size $n \to \infty$. Ferguson (1973) indicates that whatever the real distribution is, the posterior base distribution $H^*$ converges to the actual distribution uniformly almost surely. This fact comes from the Glivenko-Cantelli theorem and the proportion to the Dirichlet prior $\frac{\alpha}{\alpha+n} \to 0$ as $n \to \infty$.

## 2.3.2 Construction of Dirichlet Process

There are several representations for a Dirichlet process. We review three different methods to construct and simulate the Dirichlet process.

### Ferguson (1973) sum representation

Similar to the Dirichlet distribution which can be constructed through a set of gamma random variables divided by the sum, the Dirichlet process can also be defined as a normalized gamma process. A sum representation of Gamma process divided by the total heights of jumps should be distributed as a Dirichlet process.

**Theorem 2.3.4.** *The random measure defined by*

$$\mathcal{P}^{Ferg}(\cdot) = \sum_{i=1}^{\infty} \frac{N^{-1}(\Gamma_i)}{\sum_{j=1}^{\infty} N^{-1}(\Gamma_j)} \delta_{\theta_i}(\cdot) \tag{2.3.1}$$

*is a Dirichlet process on* $(\Theta, \Sigma)$ *with parameters* $\alpha$ *and* $H$*, where* $\{\theta_i, i = 1, 2, \cdots\}$ *is a sequence of i.i.d. random variables with distribution* $H$ *and the distributions of* $\Gamma_i$ *and* $N(x)$ *are defined as before.*

### Bondesson (1982) sum representation

**Theorem 2.3.5.** *Let* $\{\theta_i, i = 1, 2, \cdots\}$ *be a sequence of i.i.d. random variables with distribution* $H$ *and* $\{E_i, i = 1, 2, \cdots\}$ *be a sequence of i.i.d. random variables with a common exponential distribution* $\mathcal{E}xp(1)$*, independent of* $\{\Gamma_i, i = 1, 2, \cdots\}$ *and* $\{\theta_i, i = 1, 2, \cdots\}$*. Then,*

$$\mathcal{P}^{Bond}(\cdot) = \sum_{i=1}^{\infty} \frac{e^{-\frac{\Gamma_i}{a}} E_i}{\sum_{j=1}^{\infty} e^{-\frac{\Gamma_j}{a}} E_j} \delta_{\theta_i}(\cdot) \tag{2.3.2}$$

*is a Dirichlet process with parameter* $\alpha$ *and* $H$*.*

Bondesson (1982) offers a constructive representation in (2.3.2) which we can use to draw samples from a Dirichlet process. We will show a more intuitive and constructive definition of Dirichlet process, called stick-breaking method (Sethuraman, 1994), in the following part.

### Stick-breaking construction by Sethuraman (1994)

It is intuitive that the draws from a Dirichlet process are composed of a weighted sum of point masses. Sethuraman (1994) introduces a stick breaking probability measure, which in fact is a Dirichlet process.

**Theorem 2.3.6.** *Let* $\{\theta_i, i = 1, 2, \cdots\}$ *be a sequence of i.i.d. random variable with distribution* $H$ *and* $\{\beta_i, i = 1, 2, \cdots\}$ *be a sequence of i.i.d random variables with a common*

*Beta distribution* $Beta(1, \alpha)$*, independent of* $\{\theta_i, i = 1, 2, \cdots\}$*. Define*

$$\pi_1 = \beta_1, \qquad \pi_i = \beta_i \prod_{k=1}^{k-1} (1 - \beta_k) \quad i \geq 2.$$

*Then,*

$$\mathcal{P}^{Seth}(\cdot) = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}(\cdot) \tag{2.3.3}$$

*is a Dirichlet process with parameter* $\alpha$ *and* $H$*.*

The construction of $\pi$ in (2.3.3) can be understood metaphorically as breaking a stick. Suppose we have a stick of length 1 and break it at position $\beta_1$, assigning $\pi_1$ to be the length of the broken part. Now recursively break the other portion to obtain $\pi_2$, $\pi_3$ and so forth. The stick-breaking distribution over $\pi$ is sometimes written by $\pi \sim GEM(\alpha)$, where GEM represents Griffiths, Engen and McCloskey.

Ferguson (1973) and Bondesson (1982) define the representation of Dirichlet process by using the normalized Gamma process; however, Sethuraman (1994) uses the stick-breaking method avoiding inverting the Lévy measure, similar to Bondesson (1982) representation.

### 2.3.3 Approximation of Dirichlet Process

Similar to the approximations of a Gamma process, we can also use the truncation method to get the Bondesson (1982) approximation and Sethuraman (1994) approximation.

For Bondesson (1982) approximation, we may approximate the Dirichlet process as follows:

$$\mathcal{P}_n^{Bond}(\cdot) = \sum_{i=1}^{n} \frac{e^{-\frac{\Gamma_i}{a}} E_i}{\sum_{j=1}^{n} e^{-\frac{\Gamma_j}{a}} E_j} \delta_{\theta_i}(\cdot) \tag{2.3.4}$$

where the value of $n$ is derived from the stopping rule:

$$n = \inf \left\{ i : \frac{e^{-\frac{\Gamma_i}{a}} E_i}{\sum_{j=1}^{i} e^{-\frac{\Gamma_j}{a}} E_j} < \epsilon \right\}.$$

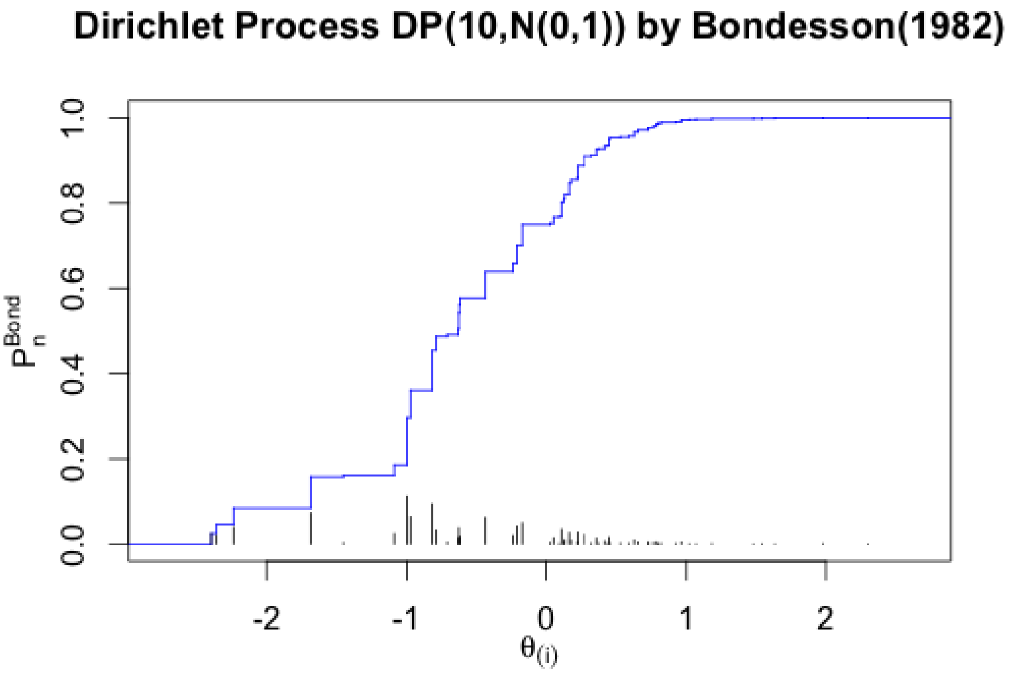


Figure 2.5: One sample path of the Dirichlet process approximated using the Bondesson (1982) Dirichlet process approximation in (2.3.4). We choose $a = 10$, $H = N(0, 1)$. For $\epsilon = 0.0001$, we get $n = 58$. Vertical lines represent weights for each $\theta_{(i)}, i = 1, \ldots, n$.

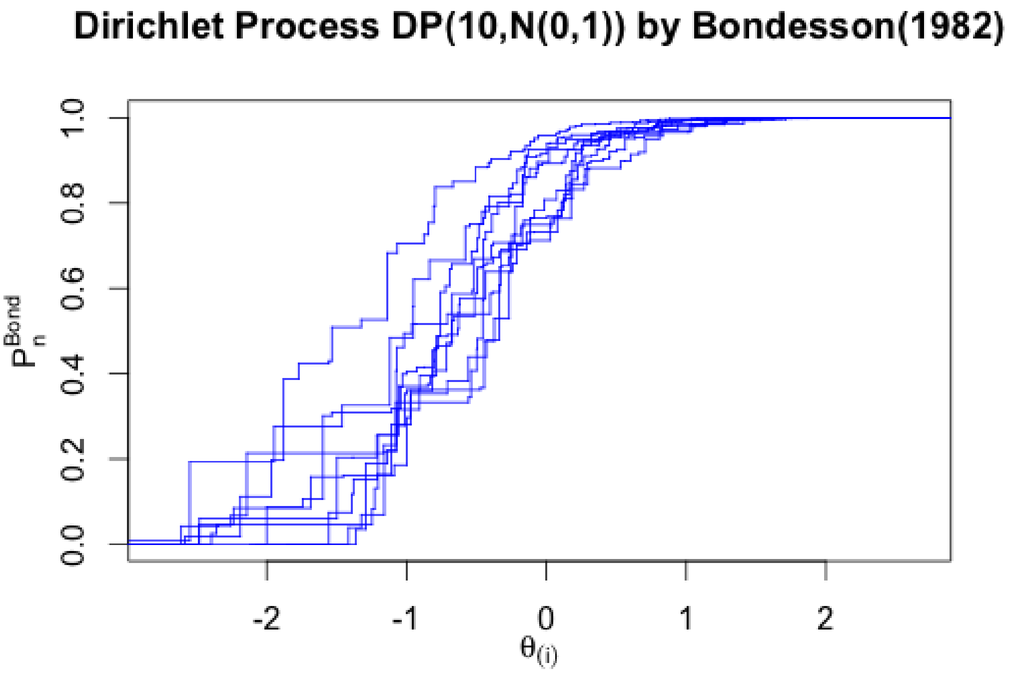


Figure 2.6: Ten sample paths of the Dirichlet process approximated by using the Bondesson (1982) Dirichlet process approximation in (2.3.4). We choose $a = 10$, $H = N(0, 1)$. For $\epsilon = 0.0001$, we get $n = 58, 36, 53, 30, 35, 48, 33, 45, 59, 46$.

For Sethuraman (1994) approximation, we may approximate Dirichlet process as follows:

$$\mathcal{P}_n^{Seth}(\cdot) = \sum_{i=1}^{n} \pi_i \delta_{\theta_i}(\cdot) \tag{2.3.5}$$

where the choice of $n$ is determined by

$$n = \inf \{ i : \pi_i = (1 - \beta_1) \cdots (1 - \beta_{i-1}) \beta_i < \epsilon \} .$$

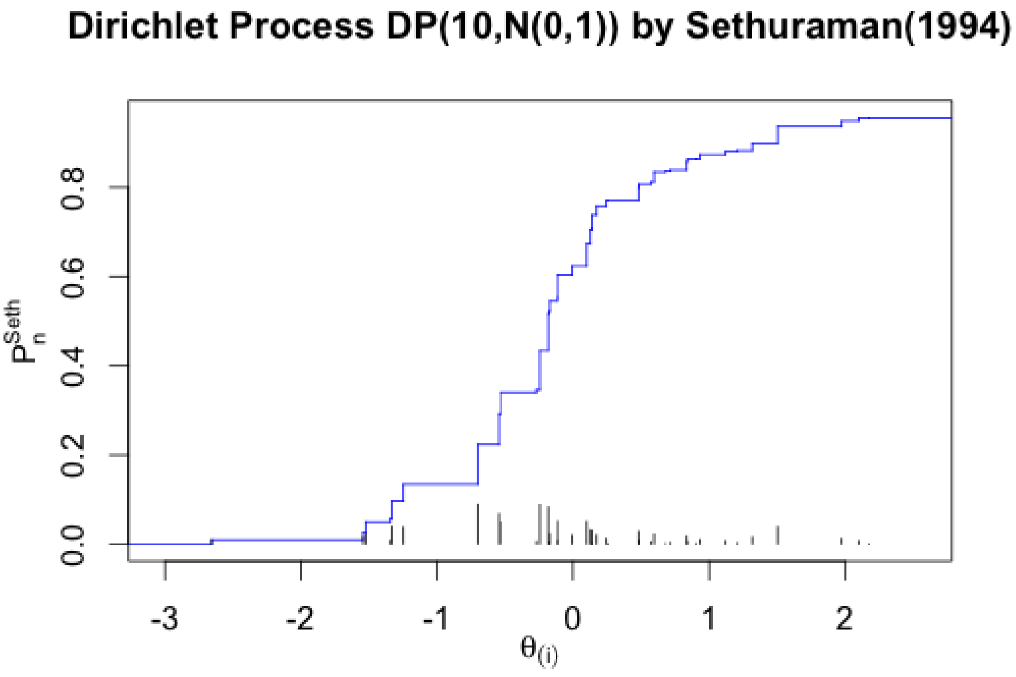


Figure 2.7: One sample path of the Dirichlet process approximated using the Sethuraman(1994) Dirichlet process approximation in (2.3.5). We choose $a = 10$, $H = N(0, 1)$. For $\epsilon = 0.0001$, we get $n = 41$. Vertical lines represent weights for each $\theta_{(i)}, i = 1, \ldots, n$.

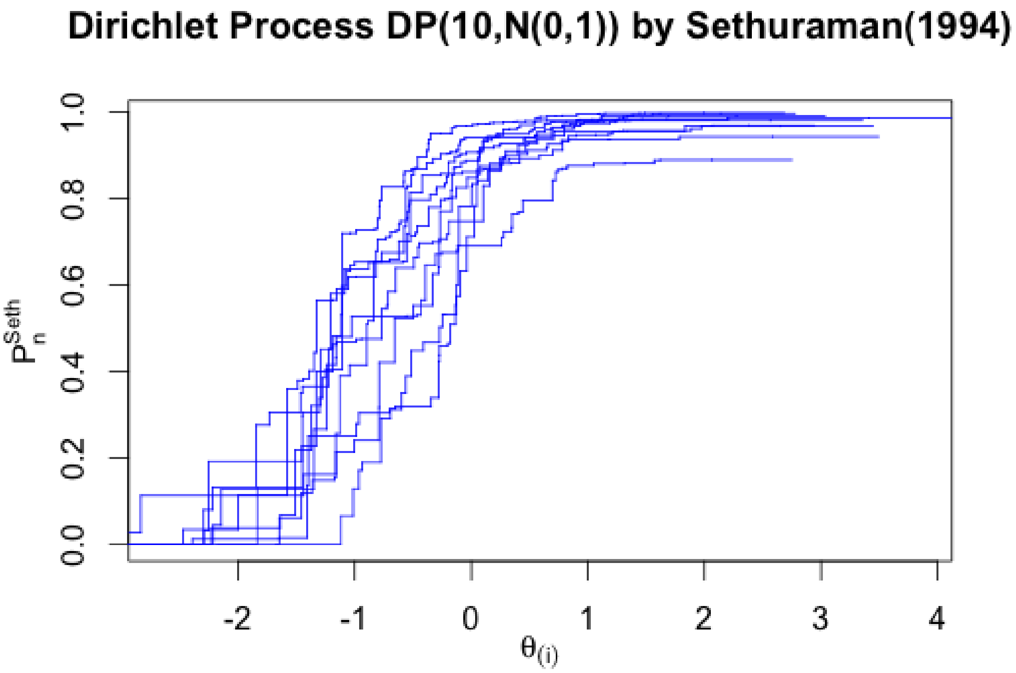


Figure 2.8: Ten sample paths of the Dirichlet process approximation with $a = 10$, $H = N(0,1)$ and $\epsilon = 0.0001$. By using Sethuraman (1994) Dirichlet process approximation in (2.3.5), the truncation values here are $n =$ 32, 38, 32, 28, 40, 53, 40, 69, 34, 52, for the same tolerance value $\epsilon = 0.0001$

**Theorem 2.3.7** (Zarepour & Al Labadi (2012)). *Let $\{\theta_i, i = 1, 2, \cdots\}$ be a sequence of i.i.d. random variables with distribution $H$. Then,*

$$\mathcal{P}_n^{Zar\&Al-Lab} = \sum_{i=1}^{n} \frac{G_n^{-1}(\frac{\Gamma_i}{\Gamma_{n+1}})}{\sum_{j=1}^{n} G_n^{-1}(\frac{\Gamma_j}{\Gamma_{n+1}})} \delta_{\theta_i} \xrightarrow{a.s.} \mathcal{P}^{Ferg}. \tag{2.3.6}$$

*where, $\{\Gamma_i\}$ and $G_n^{-1}(x)$ are defined in (2.2.2) and (2.2.6).*

Even though Zarepour & Al Labadi (2012) approximation is not based on the truncation method, which we have explained in the previous section, we can also define the stopping rule for comparison purpose as follows:

$$n = \inf\left\{ i : \frac{G_i^{-1}(\frac{\Gamma_i}{\Gamma_{i+1}})}{\sum_{j=1}^{i} G_i^{-1}(\frac{\Gamma_j}{\Gamma_{i+1}})} < \epsilon \right\}.$$

For the same tolerance value of $\epsilon$, we can get smaller value of $n$ by using Zarepour & Al Labadi (2012) approximation, compared to Bondesson (1982) approximation and Sethuraman (1994) approximation.

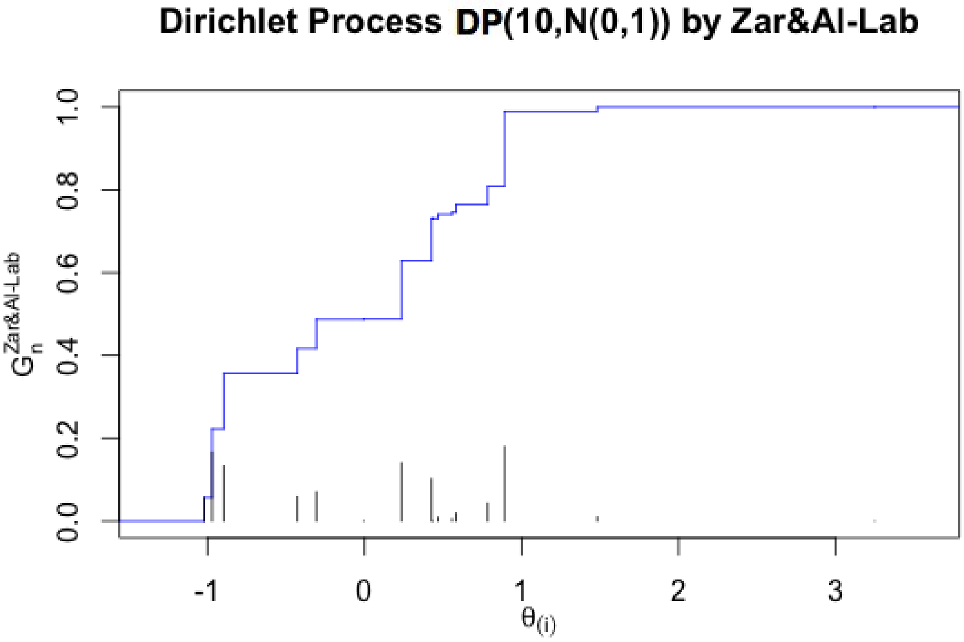


Figure 2.9: One sample path of the Dirichlet process approximated using the Zarepour and Al Labadi (2012) Dirichlet process approximation in (2.3.6). We choose $a = 10$, $H = N(0,1)$. For $\epsilon = 0.0001$, we get $n = 16$. Vertical lines represent weights for each $\theta_{(i)}, i = 1, \ldots, n$.

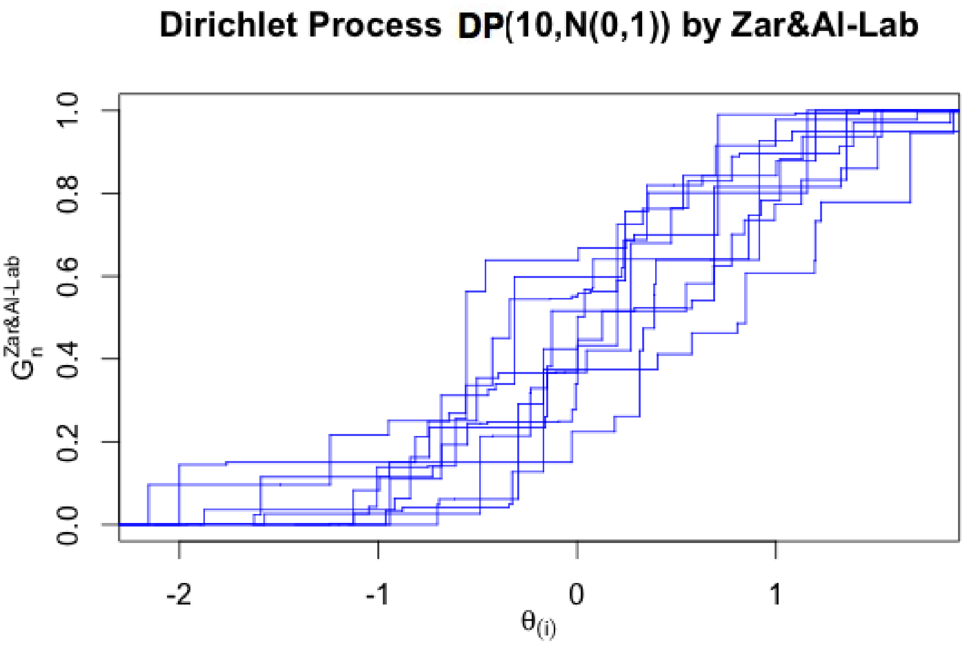
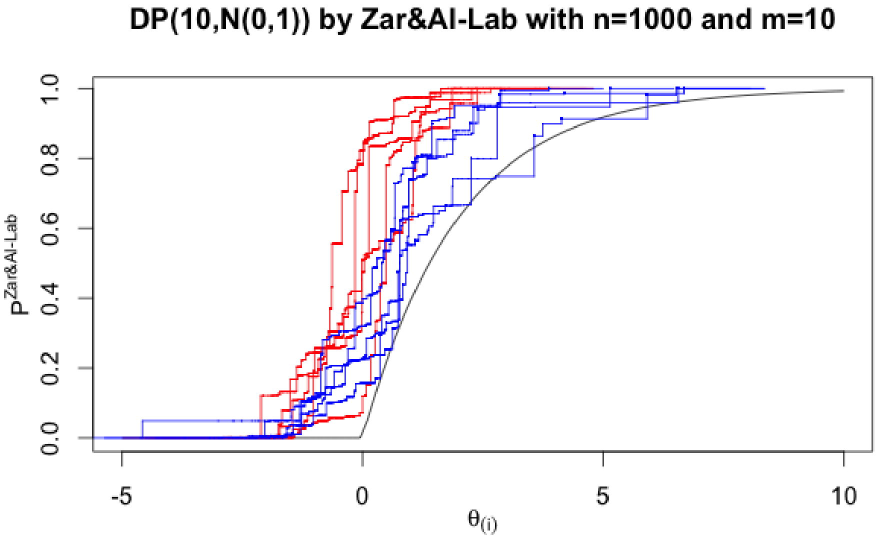
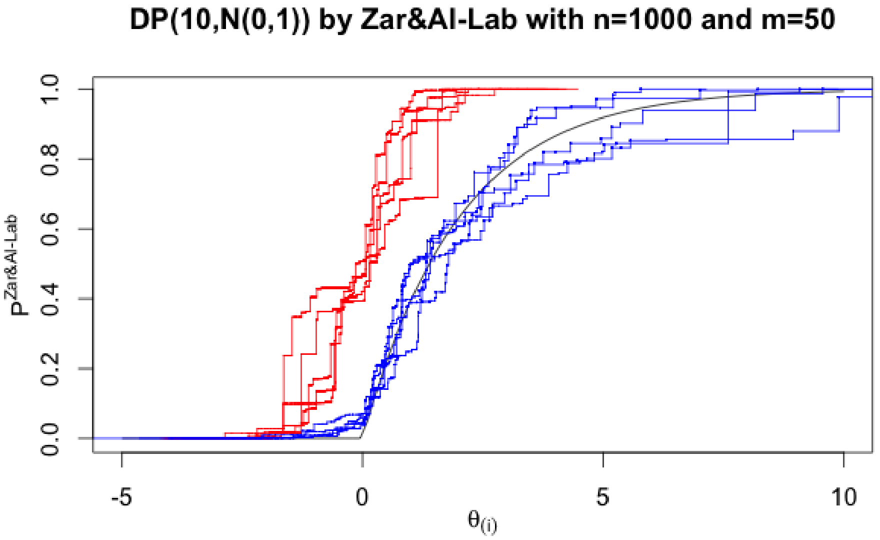
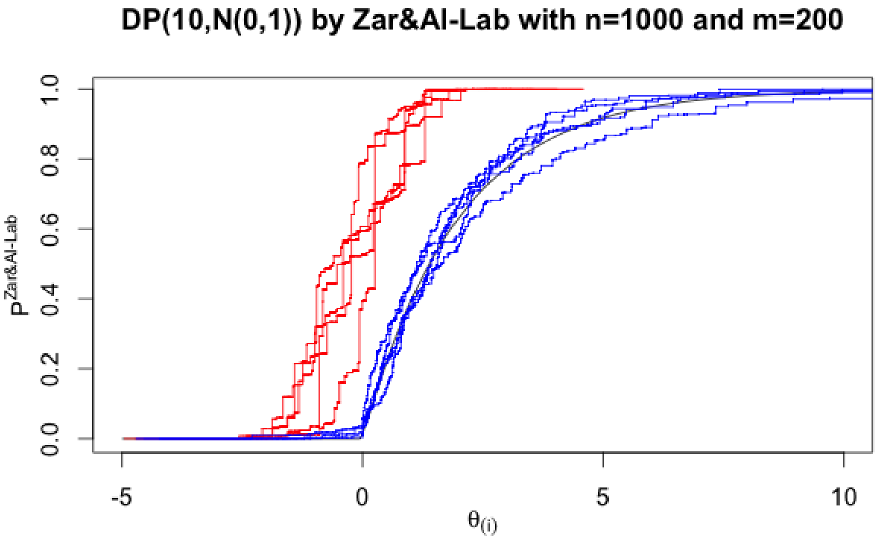


Figure 2.10: Ten sample paths of the Dirichlet process approximation with $a = 10$, $H = N(0, 1)$ and $\epsilon = 0.0001$. By using the Zarepour and Al Labadi (2012) Dirichlet process approximation in (2.3.6), the truncation values here are $n = 13$, 13, 12, 18, 18, 15, 12, 14, 14, 18, for the same tolerance value $\epsilon = 0.0001$

In the following, we give an example to show one of the applications of the Dirichlet process, estimating the distribution of the sample data. For concreteness, considering a sample set $X_1, \ldots, X_m$ from the exponential distribution $F \sim \mathcal{E}xp(0.5)$, we are trying to infer the real distribution of the sample set.


DP(10,N(0,1)) by Zar&Al-Lab with n=1000 and m=10
$P^{Zar\&Al\text{-}Lab}$
0.0
0.2
0.4
0.6
0.8
1.0
-5
0
5
10
$\theta_{(i)}$


DP(10,N(0,1)) by Zar&Al-Lab with n=1000 and m=50
$P^{Zar\&Al\text{-}Lab}$
0.0
0.2
0.4
0.6
0.8
1.0
-5
0
5
10
$\theta_{(i)}$


DP(10,N(0,1)) by Zar&Al-Lab with n=1000 and m=200
$P^{Zar\&Al\text{-}Lab}$
0.0
0.2
0.4
0.6
0.8
1.0
-5
0
5
10
$\theta_{(i)}$

The light-colour (red) paths represent the Dirichlet process prior approximation with $a = 10$, $H = N(0, 1)$ and $n = 1000$ using the Zarepour and Al Labadi (2012)'s approximation. The dark-colour (blue) paths represent the posterior distribution approximation of Dirichlet process given the data set $\{X_i, i = 1, \ldots, m\}$ which in our case follows $P \sim \mathcal{E}xp(0.5)$ with size $m$. The smooth curve (black) is the cumulative distribution function of $\mathcal{E}xp(0.5)$.

Note that the real distribution $F$ of sample data is unknown in advance and what we do is to infer $P$ in practice. The posterior distribution of $F$ given data $\{X_i, i = 1, \ldots, m\}$ and prior distribution $P \sim DP(10, N(0, 1))$ is

$$P|X_1, \ldots X_m \sim DP\left(10 + m, \frac{10}{10 + m} N(0, 1) + \frac{m}{10 + m} \frac{\sum_{i-1}^{m} \delta_{X_i}}{m}\right).$$

Whatever our prior guess is, the posterior distribution will eventually converge to the real distribution as the sample size grows.

# Chapter 3

# Data Depth

The real line has many desirable properties that simplify analysis of the univariate data. Many of those properties are shared by the multivariate space, however, the existence of a natural order does not hold. This is an obstacle when applying univariate analysis techniques to the multivariate case.

Data depth is an important concept of nonparametric approach to multivariate data analysis. It provides a possible way to order the multivariate data. Data depth assigns each point in $\mathbb{R}^d$ a degree of centrality, namely a depth value, with respect to the probability distribution or data set. For the more central points, the depth is greater, and the depth value decreases as moving outward of the distribution. Therefore, the data depth enables us to rank the data points with respect to the centre-outward ordering. In addition to multivariate ordering, we can also obtain some characteristics of a given data set by using data depth.

## 3.1 Depth Function

A depth function measures how central a point is relative to a probability distribution or a data cloud. Given a random vector $X$ on $\mathbb{R}^d$, let $\mathcal{F}$ be the class of distributions on the Borel sets of $\mathbb{R}^d$ and $F_X$ be the distribution function of $X$. A *depth function* is a mapping $D(\cdot;\cdot) : \mathbb{R}^d \times \mathcal{F} \mapsto \mathbb{R}$ which is bounded and nonnegative, and usually satisfies the following four conditions:

1. *Affine invariance*: $D(Ax+b;F_{AX+b}) = D(x;F_X)$ for any random vector $X$, all vector $b \in \mathbb{R}^d$ and each $d \times d$ nonsingular matrix $A$.
2. *Monotonicity with respect to deepest point*: If $\theta = \arg\max_x D(x;F)$ for any $F \in \mathcal{F}$, then $D(x;F) \leq D(\theta + \alpha(x-\theta);F)$ for any $0 \leq \alpha \leq 1$.
3. *Maximality at centre*: $D(\theta;F) = \sup_{x \in \mathbb{R}^d} D(x;F)$ if $F \in \mathcal{F}$ is centred at $\theta$.
4. *Vanishing at infinity*: $\lim_{\|x\| \to \infty} D(x;F_X) = 0$ for each $F \in \mathcal{F}$.

Given a data set $X_1, X_2, \ldots X_n$ in $\mathbb{R}^d$, let $F_n$ denote the empirical distribution on this data set and the *sample version of data depth* $D(x;F_X)$ denoted by $D_n(x)$ is the depth of $x$ with respect to $F_n$, which is defined by replacing $F$ by a suitable empirical measure $F_n$.

For any chosen depth function $D(\cdot;F)$, the *central region* (or *depth-trimmed region*) of level $\alpha$, denoted by $R_F(\alpha)$, is the set of all points whose depth is at least $\alpha$, i.e.

$$R_F(\alpha) = \{x \in \mathbb{R}^d : D(x;F) \geq \alpha\}.$$

Therefore, the *p-th central region* $C_F(p)$ is the central region with probability weight $p$ defined by

$$C_F(p) = \inf_\alpha \{R_F(\alpha) : P(R_F(\alpha)) \geq p\}. \tag{3.1.1}$$

The boundary of $C_F(p)$ is referred to as the *pth level contour*, the multivariate analogue of *quantile* in the univariate case. Also, the sample versions of $p$th central region and $p$th level contour are obtained by replacing $F$ with $F_n$.

The central region satisfies various properties, some of which depend on the notion of data depth function they are built from, but typically they are:

1. *Nested*: $C_F(p_1) \subset C_F(p_2)$ if $p_1 \leq p_2$.
2. *Affine equivariant*: $C_{F_{AX+b}}(p) = C_{F_X}(p)$ for any probability value $p$, all vector $b \in \mathbb{R}^d$ and each $d \times d$ nonsingular matrix $A$.

3. *Compact*: for any probability value $p$, $C_F(p)$ is compact.
4. *Connected*: The central region space $\{C_F(p) : 0 \leq p \leq 1\}$ cannot be divided into two disjoint nonempty open sets.

Several novel depth functions have appeared in the literature, including *Mahalanobis depth (MD)* (Mahalanobis, 1936), *Tukey depth* or *half-space depth (HD)* (Tukey, 1975), *convex hull peeling depth (CHPD)* (Barnett, 1976), *simplicial depth (SD)* (Liu, 1990), *Oja depth (OD)* (Oja, 1985), *Majority depth (MD)* (Singh, 1991) and some other new notions. Here, we mainly introduce the depth functions mentioned above, and use them to present the idea of centre-outward ordering of multivariate data points.

### 3.1.1 Mahalanobis Depth

The *Mahalanobis depth* (Mahalanobis, 1936) is derived from *Mahalanobis distance* which is the most widely-used measure of distance in multivariate analysis. It measures the depth of each point with respect to the mean of a distribution $F$, which is defined as

$$MD(x; F) = [1 + (x - \mu_F)\Sigma_F^{-1}(x - \mu_F)]^{-1} \tag{3.1.2}$$

where $\mu_F$ is the mean vector and $\Sigma_F$ is the covariance matrix of the distribution $F$. The sample version of $MD(x; F)$ is replacing $\mu_F$ and $\Sigma_F$ in (3.1.2) by the sample mean and sample covariance matrix.

However, Mahalanobis depth function is not robust, since it is based on nonrobust statistics: the sample mean and sample covariance matrix. Also, it is not moment-free methodology, which depends on the existence of the second moments.

It is easy to see that the deepest point maximizing the Mahalanobis depth turns out to be the mean of the distribution. For a data set, the sample mean is greatly influenced by outliers, which makes it not robust. One of the suggested estimators dealing with the nonrobustness is the trimmed mean, where most of the outliers are cleaned and only points of the central region are taken into consideration. The trimmed mean is a robust statistics and keeps the affine invariance property.

Generally, the deepest point of most depth functions is the median, as we can see in the following part, which is a robust statistics not affected by outliers. Thus, we usually view the deepest point as the median.

### 3.1.2 Tukey Depth

The *Tukey depth* (Tukey, 1975), which is also called *half-space depth* or *location depth*, of a fixed point $x \in \mathbb{R}^d$ with respect to a probability distribution $F$ is the infimum of the probabilities of all the closed half-spaces containing $x$, which is defined as

$$HD(x; F) = \inf_H \{F(H) : H \text{ is a closed half-space on } \mathbb{R}^d \text{ and } x \in H\}. \tag{3.1.3}$$

Given a data set $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$, a sample version of $HD(x; F)$ is replacing $F$ in (3.1.3) by the empirical distribution $F_n$. In other words, the sample Tukey depth of a point $x \in \mathbb{R}^d$ is the smallest proportion of data points in a closed half-space containing $x$. That is, if we denote the Tukey depth of $x$ with respect to the data set by $HD_n(x)$, then

$$HD_n(x) = \frac{\min\limits_{\|u\|=1} \#\{i : u'x_i \geq u'x\}}{n}$$

where $u$ ranges over all vectors in $\mathbb{R}^d$ with $\|u\| = 1$. In the univariate case, it is easy to see that the Tukey depth of $x$ is given by $\min\{\#\{x_i \leq x\}, \#\{x_i \geq x\}\}$. We show the bivariate case by the following graph.

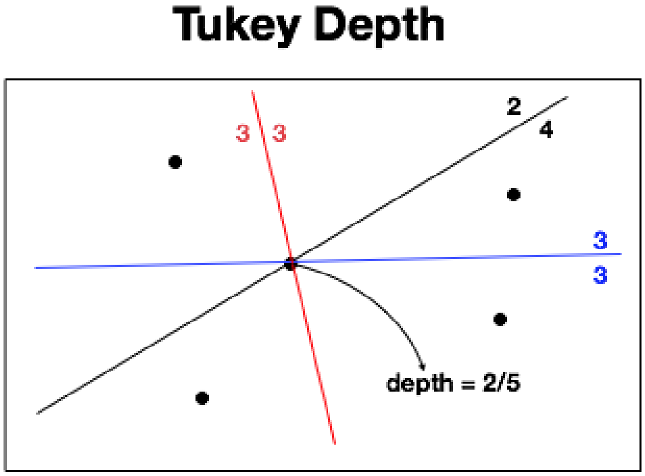


Figure 3.1: The Tukey depth of a point in bivariate case, which in this case is $\frac{2}{5}$.

The point on $\mathbb{R}^d$ with the highest Tukey depth is called the *Tukey median*. And the *sample Tukey median* is defined as the sample point maximizing $HD_n(\cdot)$, or the average if there are more than one such point.

Many interesting multivariate researches are based on Tukey depth and we will employ Tukey depth for further analysis in this paper. Three other popular depth functions which are Oja depth, Simplicial depth and Convex hull peeling depth are introduced in the following part for comparison purposes.

### 3.1.3 Simplicial Depth

The *Simplicial depth* (Liu, 1990) of a data point $x \in \mathbb{R}^d$ with respect to a probability distribution $F$ is defined as

$$SD(x; F) = P_F(x \in S[X_1, \ldots, X_{d+1}]) \tag{3.1.4}$$

where $X_1, \ldots, X_{d+1}$ are $(d+1)$ independent random observations from $F$ and $S[X_1, \ldots, X_{d+1}]$ is the closed simplex with vertices $X_1, \ldots, X_{d+1}$.

Let $S_n = \{X_1, \dots, X_n\}$ be a data set in $\mathbb{R}^d$. The sample version of simplicial depth is defined by replacing $F$ in (3.1.4) with empirical distribution $F_n$, namely $SD(x; F_n)$, or alternatively, by calculating the proportion of closed simplices containing $x$ and having $(d+1)$ vertices in $S_n$. We rewrite $SD(x; F_n)$ as $SD_n(x)$, which is

$$SD_n(x) = \binom{n}{d+1}^{-1} \sum_{1 \leq i_1 < \cdots < i_{d+1} \leq n} I(x \in S[X_{i_1}, \dots, X_{i_{d+1}}])$$

where $I(\cdot)$ is the indicator function and $X_{i_1}, \dots, X_{i_{d+1}} \in S_n$.

Given a bivariate data set of size $n$, the sample simplicial depth of a point $x$ on $\mathbb{R}^2$ is the proportion of all closed triangles containing $x$ inside with vertices in the data set. We may generate $\binom{n}{3}$ closed triangles from this data set and denote the closed triangle containing $x$ as $x \in \triangle(X_i, X_j, X_k)$. Assume that no three points can lie on a straight line with probability zero. For any continuous function $F$, this condition is satisfied. Thus, we have bivariate sample simplicial depth which is $D_n(x) = \binom{n}{3}^{-1} \sum_{1 \leq i < j < k \leq n} I(x \in \triangle(X_i, X_j, X_k))$. For multivariate case, triangles are replaced by simplices with $(d+1)$ vertices in the data set.

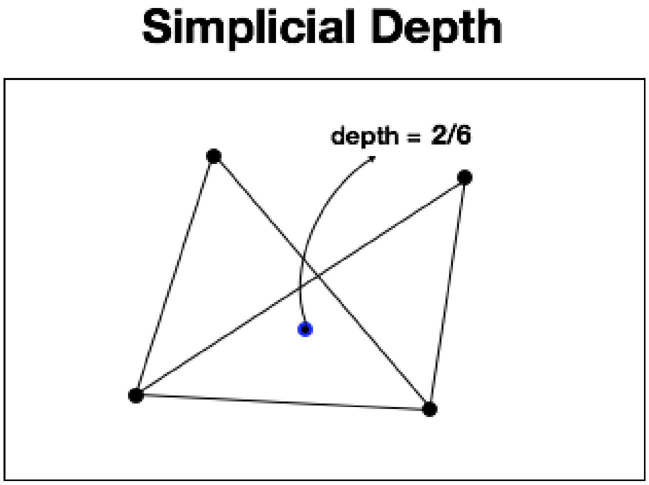


Figure 3.2: The simplicial depth for a point in bivariate data set, where there are two triangles enclosing this point and we can get at most 6 distinct triangles from this black data set of size 4. Thus the depth for this point is $\frac{2}{6}$. To show the graph nicely, we do not include the blue point in the data set in this case.

The *simplicial median* is the point with the highest simplicial depth, that is the point maximizes $SD(\cdot; F)$. The *sample simplicial median* is the sample point attaining the highest sample simplicial depth, or the average of such points if the maximum is achieved at more than one data point.

### 3.1.4 Oja Depth

The *Oja depth* (Oja, 1985) of a point $x \in \mathbb{R}^d$ with respect to a probability distribution $F$ is based on the expected volume of closed simplices having vertices at $x$ and a random sample of size $d$ from $F$, which is defined to be

$$OD(x; F) = [1 + E_F\{\text{Volume}(S[x, X_1, \ldots, X_d])\}]^{-1} \tag{3.1.5}$$

where $X_1, \ldots, X_d$ are $d$ independent random observations from $F$ and $S[x, X_1, \ldots, X_d]$ is the closed simplex with vertices $x$ and $X_1, \ldots, X_d$.

Obviously, given a data set $X_1, X_2, \cdots, X_n$ in $\mathbb{R}^d$, the sample version of Oja depth of $x \in \mathbb{R}^d$ is based on the average volume of all closed simplices with a vertex at $x$ and the others from the data set. From the data set, we can generate $\binom{n}{d}$ such simplices. Therefore, the sample Oja depth is given by

$$OD_n(x) = \binom{n}{d}^{-1} [1 + \sum_{1 \le i_1 < \ldots < i_d \le n} \{\text{Volume}(S[x, X_{i_1}, \ldots, X_{i_d}])\}]^{-1}.$$

In the bivariate case, the Oja depth of a point $x$ relative to a bivariate data set of size $n$ is the average area of all closed triangles with a vertex at $x$ and other two from the data set.

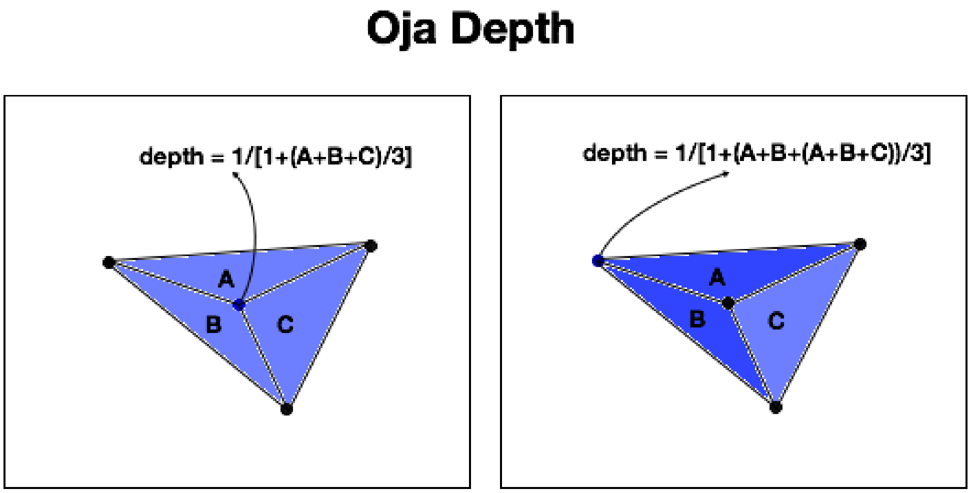


Figure 3.3: The Oja depth for points in a bivariate data set. It is clear to see that the depth of a inner point is greater than the outer point.

Similarly, the point with maximum Oja depth is called the *Oja median*. And the *sample Oja depth* is also defined in a similar way, which is the deepest point or the average of the deepest points if there are more than one.

We have to indicate here that the Oja depth is not affine invariant. That is, the depth value does not remain the same after any affine transformation. In this case the Oja depth will be changed by a factor of the determinant of $A$. However, the ranking induced by the Oja depth and the Oja median remain affine invariant.

### 3.1.5 Convex Hull Peeling Depth

*Convex hull peeling depth* (Barnett, 1976) at a point $x$ with respect to a data set is the *level of convex layer* to which $x$ belongs. Construct the smallest convex hull enclosing all data points. The points on the boundary of the convex hull are assigned as the first convex layer and removed. Construct the convex hull of the remaining points in the same way getting the second convex layer and keep the process repeated. A sequence of convex layers are established. The higher the layer the deeper the points. Therefore, the deepest point of this data set is defined and take the average of the deepest points if there are more than one.

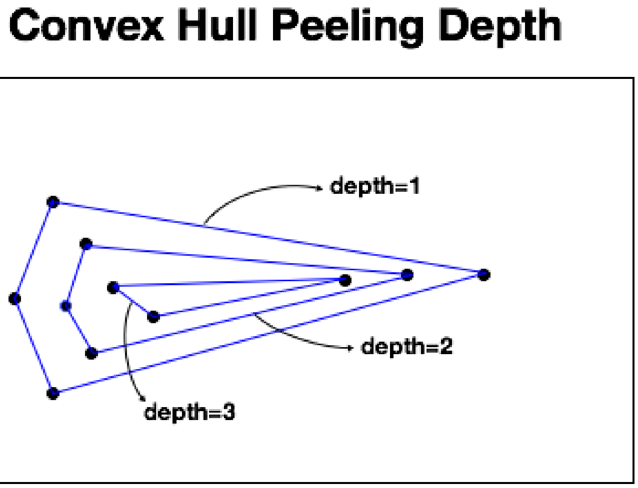


Figure 3.4: Convex hull peeling depth of a bivariate data set.

Intuitively, the three points on the right hand side are supposed to be viewed as outliers and the reasonable centre should be the blue point. However, by convex hull peeling depth, the deepest point is calculated as the average of inner three points, which is greatly affected by outliers. It turns out that convex hull peeling depth keeps some unpleasant nonrobust properties. Also, it does not have a population analogue.

**Remark 3.1.1.** The depth value may vary with the notion of data depth, but for each notion of depth, a larger value of $D(\cdot; F)$ always implies a deeper (or more central) point with respect to $F$. Given a notion of depth, we can compute the depths of all data points $X_1, \ldots, X_n$ and make an ordering of them according to decreasing depth values, i.e. from the centre outward, denoted as $X_{[1]}, \ldots, X_{[n]}$. For symmetric distributions, the orderings derived from different notions of depth function vary little from one to the other. However, they may be quite distinct if the distribution is asymmetric, especially for the deepest points. This will be shown visually in the next section.

## 3.2 Graphical Representation

### 3.2.1 Contours

**Symmetric Distributions**

Given a sample of 500 points drawn from bivariate normal distribution with mean $\mu = (2, 3)'$ and covariance matrix $\Sigma = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}$, we use Mahalanobis depth, Tukey depth, Oja depth and simplicial depth to order data points and draw some contour graphs applying the different depth orderings. The following graphs show the sample $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$. As the $p$-value increases, the contour expands from centre outward. The deepest point in the sample is marked by a cross. There is little difference between Figure 3.5 to 3.8. The shapes of contours and $p$th central regions in the figures are pretty similar and the deepest points of these four depth functions are all the same, which is (1.901198, 3.126400). This is because the sample data is normal and symmetric.

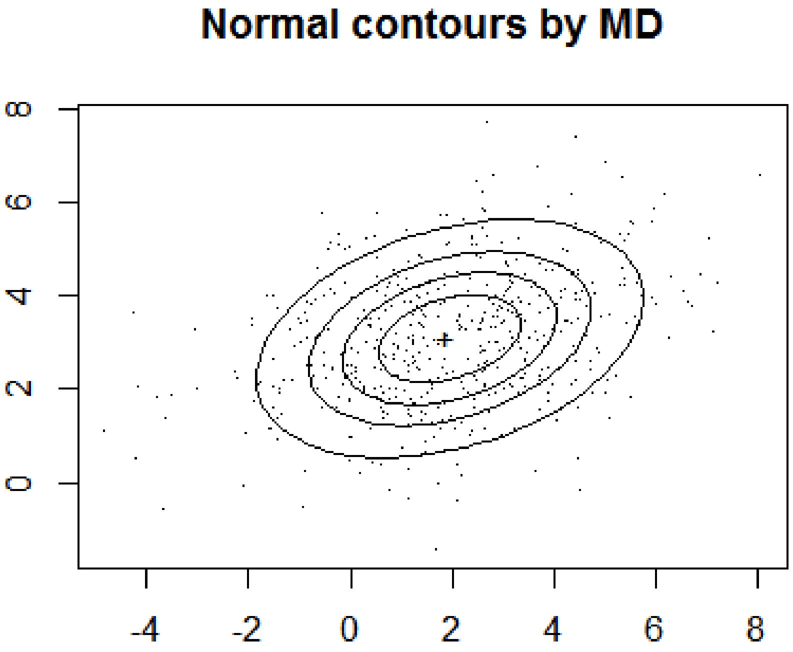


Figure 3.5: Normal $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Mahalanobis depth.

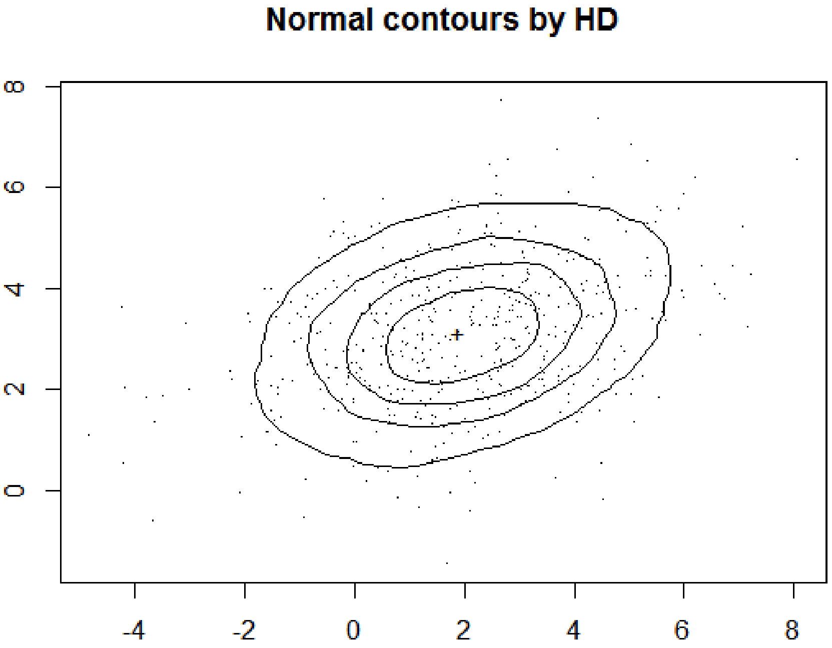


Figure 3.6: Normal $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Tukey depth.

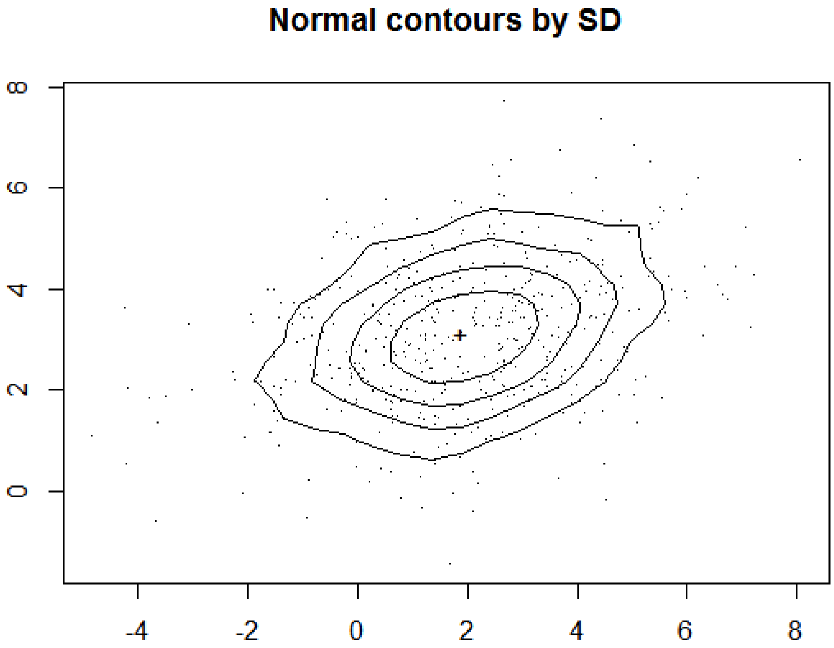


Figure 3.7: Normal $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by simplicial depth.

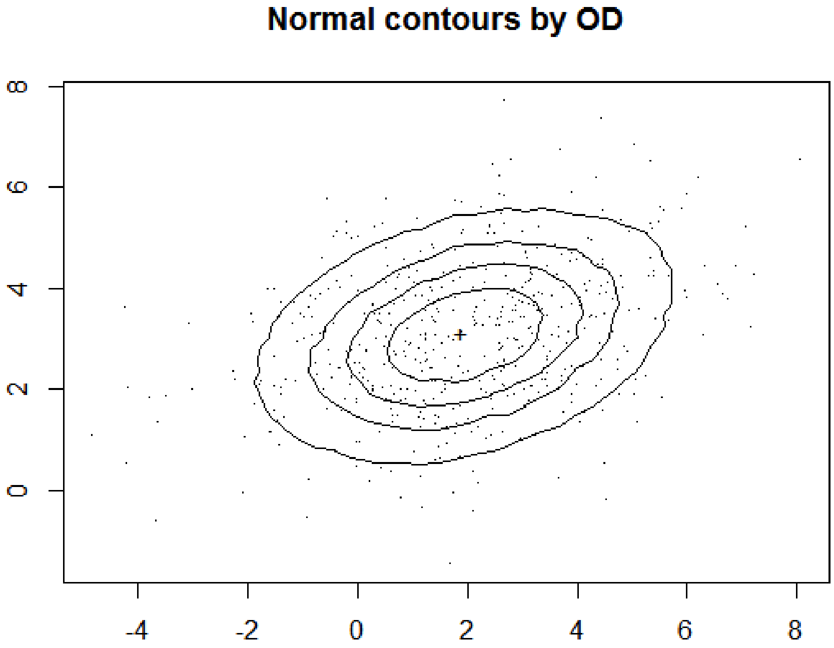


Figure 3.8: Normal $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Oja depth.

Figure 3.9 is a 3-dimensional graph, where the vertical height represents the depth value. It is obvious to see from the graphs that the highest depth is obtained at the centre and the depths of outlying points are approaching to zero. Also, the depth values vary with the notions of depth functions but larger depth values always imply more central location.

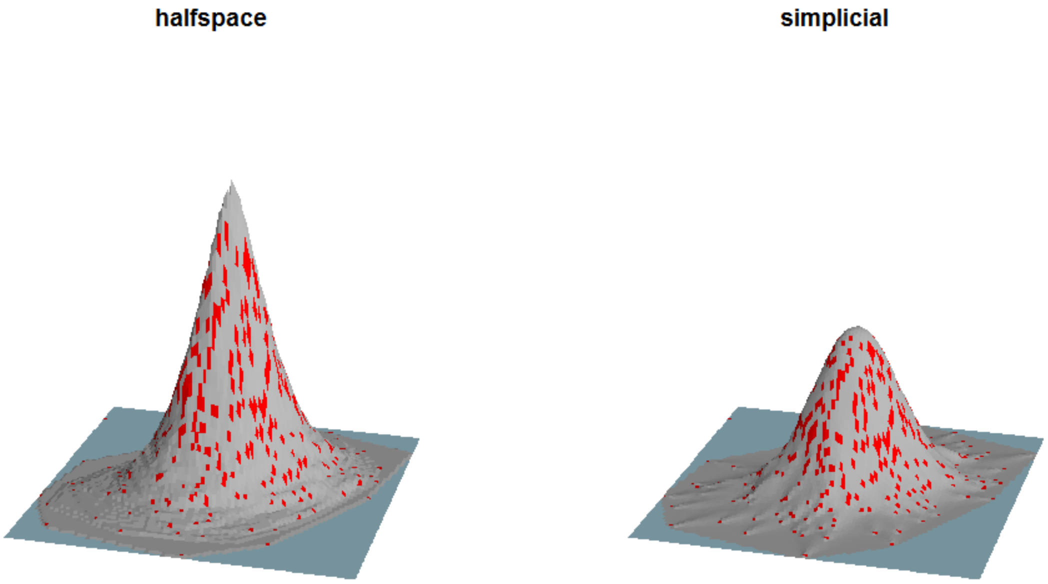


Figure 3.9: Put Tukey depth graph and simplicial depth graph in the same coordinate system and we can easily see the difference between them. Different depth functions may assign sample points different depth values even though we get the similar rankings in this sample.

### Asymmetric Distributions

For asymmetric distributions, the difference between different notions of depth function is noticeable. Draw a sample from the standard bivariate exponential distribution and common rate and do the same as above getting the following graphs. It is worth noticing that there exists remarkable difference between individual depth functions. The following table shows us the location of the deepest point of each depth function, which are almost different.

| Depth | x | y |
|---|---|---|
| Mahalanobis | 0.9972922 | 0.9849399 |
| Tukey | 0.7595659 | 0.7435865 |
| Simplicial | 0.7595659 | 0.7435865 |
| Oja | 0.7914264 | 0.6955089 |

Table 3.1: The deepest points by different depth functions.

Obviously, the shapes of contours between each graph vary greatly, which means we get distinct $p$th central regions by using different notions of depth function, whereas they are pretty similar for symmetric distributions. The following graphs show the sample $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$.

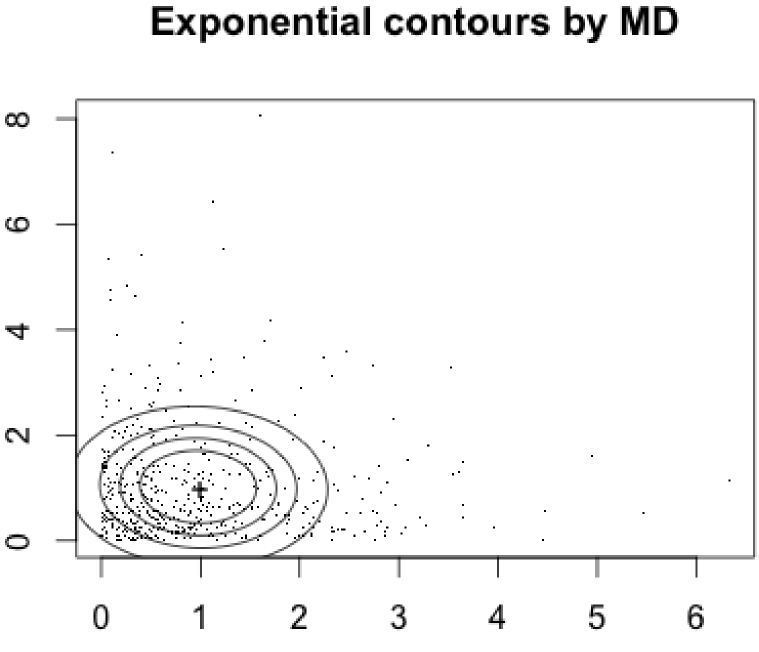


Figure 3.10: Exponential $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Mahalanobis depth, where the deepest point is (0.9972922, 0.9849399).

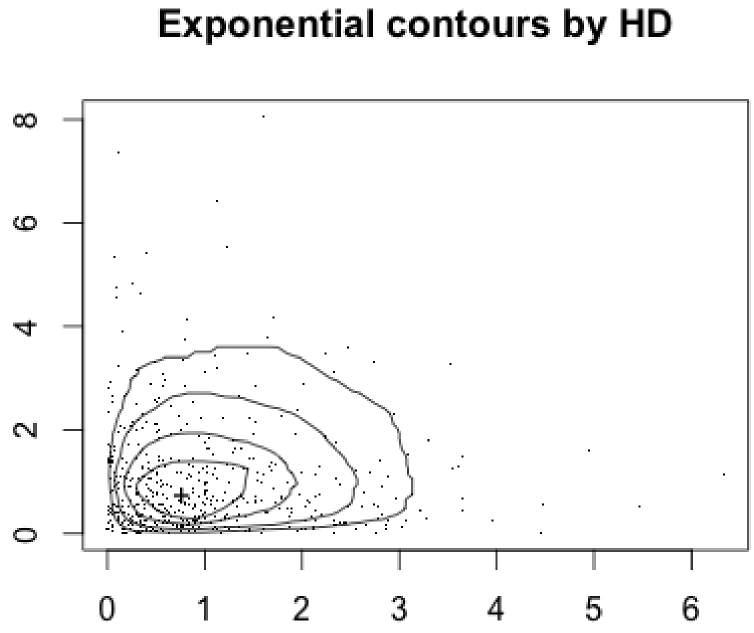


Figure 3.11: Exponential $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Tukey depth, where the median is (0.7595659, 0.7435865).

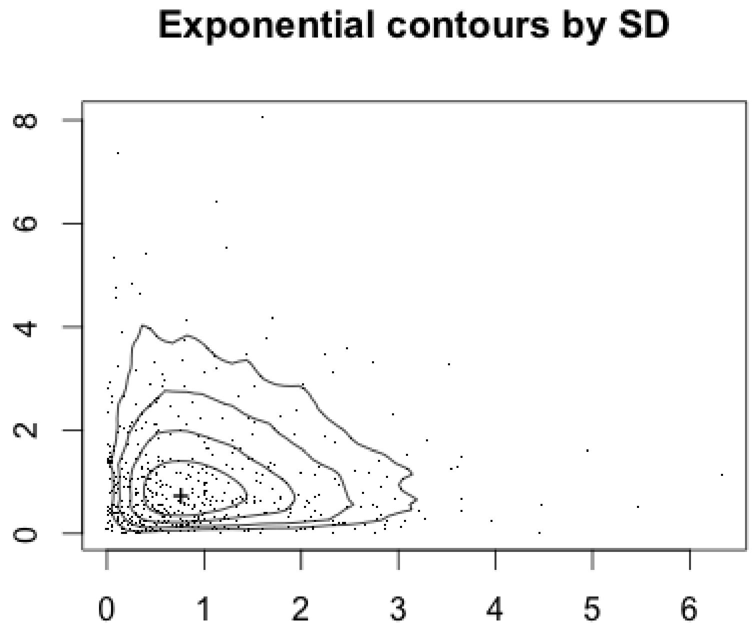


Figure 3.12: Exponential $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Simplicial depth, where the median is (0.7595659, 0.7435865).

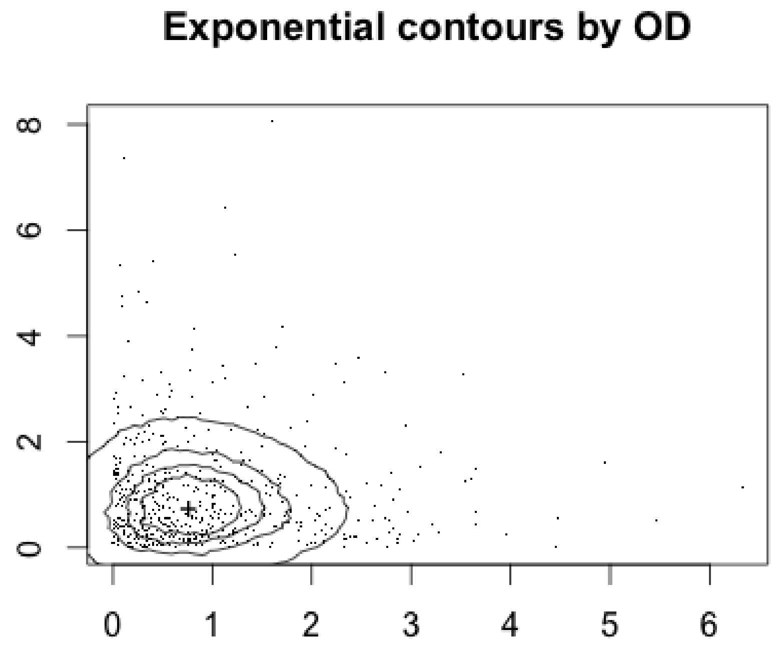


Figure 3.13: Exponential $p$th contours for $p = 0.2, 0.4, 0.6, 0.8$ by Oja depth, where the median is (0.7914264, 0.6955089).

### 3.2.2 Bagplot

Another way to present bivariate data is the *bagplot*. The bagplot was proposed by Rousseeuw (1999) considered as a bivariate generalization of the univariate boxplot. This data depth representation is constructed by using Tukey depth, however, other notions of data depth might be used as well. The main component of the bagplot is the *bag* which is the central region containing 50% of all points. This plot is also composed by a *fence*, which separates the outliers from other points, and a *loop* which is an area between the bag and the fence.

Like the univariate boxplot, the bagplot also provides us several characteristics of the data, such as the location (median), dispersion (the size of the bag), correlation (the orientation of the bag) and skewness (the shape of the bag and the loop).

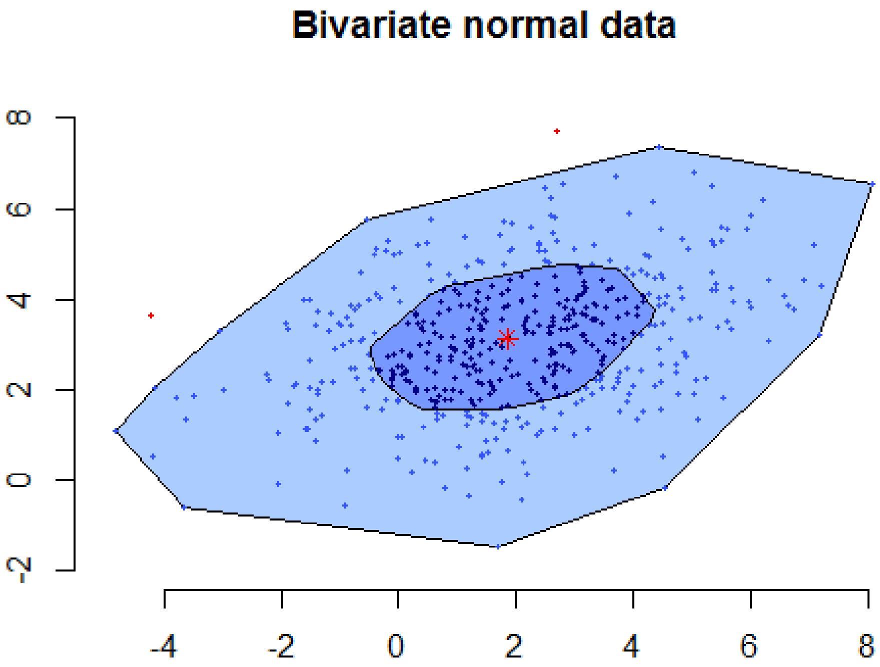


Figure 3.14: The bagplot of bivariate normal data set mentioned above. The anticlockwise rotation represents the positive correlation between two variables. The shape of the bagplot is an ellipse because of the variance of first variable is relatively larger.

The bagplot is an intuitive and informative graphic representation to visualize data, as well as the contour plot. We apply the bagplot to analyze a real data set, which is *Daily air quality measurements in New York*. There are two variables in this data set, temperature and wind speed.

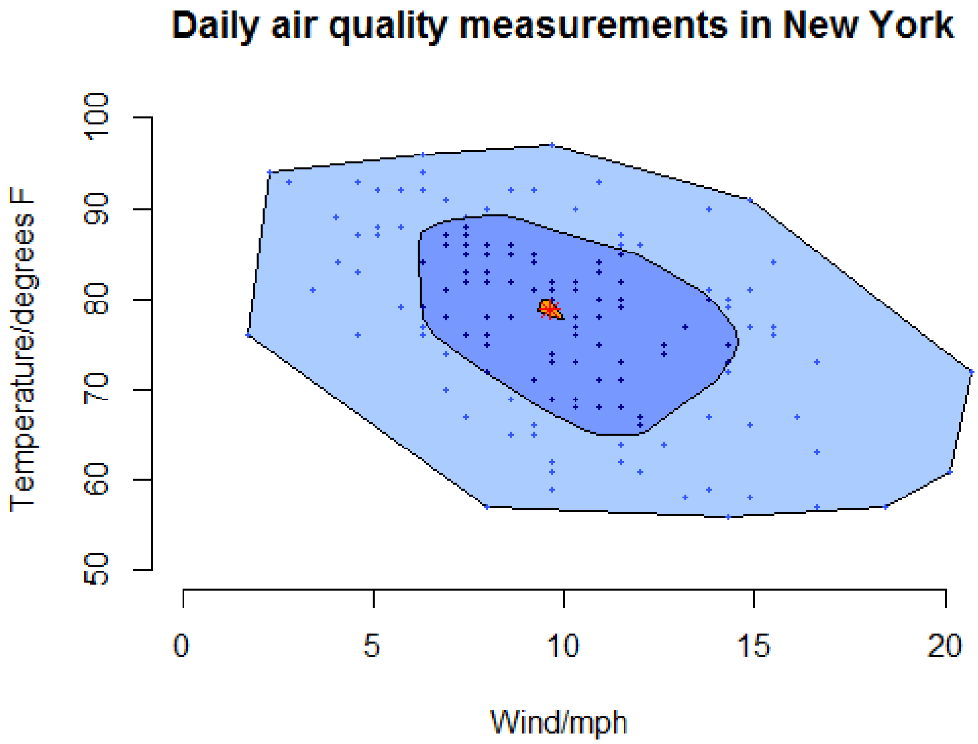


Figure 3.15: The bagplot of air quality data in New York, May to September 1973. The wind speed has a negative correlation with the temperature which virtually makes sense. From the graph, we can roughly see the distribution is also close to bivariate normal distribution.

# Chapter 4

# Simulation of the Bootstrap Methods

The bootstrap is a well-used resampling technique and simple to operate. In this chapter, we review two main bootstrap methods: Efron's regular bootstrap and Bayesian bootstrap with a nonparametric prior, as well as their relevant properties and their extensions. For simplicity, we simulate data from the bivariate normal distribution and the bivariate Cauchy distribution. Then apply these bootstrap methods to the sample data respectively. The statistics of interest is the Tukey median and we compare these bootstrap methods on the Tukey median.

## 4.1 Efron's Regular Bootstrap

### 4.1.1 Introduction

The regular bootstrap method was first introduced by Efron (1979). This is a very general resampling technique for estimating the statistics of the unknown distribution and drawing inferences from the sampling distribution of the statistics based on the observed sample data. The basic idea of the regular bootstrap is to draw samples from the empirical distribution $F_n$ and the population distribution $F$ is completely unknown in this case. A sample of size $n$ from $F_n$ is whereby a sample of size $n$ randomly drawn with replacement from the sample data set $X_1, \ldots, X_n$, which is also called a *bootstrap sample* or a *resample*.

Considering a sample set $X_1, \ldots, X_n$ from the population distribution $F$, let $\theta$ be the unknown parameter and $\hat{\theta}_n$ be an estimate for $\theta$. The basic steps of the regular bootstrap can be written as follows.

**Step 1.** Place a probability of $1/n$ to each sample point $X_1, X_2, \ldots, X_n$ to construct the empirical probability distribution $F_n$ of the sample points. The empirical distribution $F_n$ is the nonparametric maximum likelihood estimate of the population distribution $F$ when $F$ is completely unknown.

**Step 2.** Draw a random sample of size $n$ with replacement from the empirical distribution function $F_n$ to form a new sample $X_1^*, \ldots, X_n^*$. This is a bootstrap sample with the bootstrap empirical distribution $F_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i^*}$.

**Step 3.** Calculate the statistic of interest, $\hat{\theta}_n$, for this bootstrap sample, denoted by $\hat{\theta}_n^*$.

**Step 4.** Repeat steps 2 and 3 a large number of times, say B times, to generate $B$ resamples and get a collection of values for the statistics of interest, denoted by $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Typically, $B$ is at least equal to 1000 in order to obtain an estimate of the confidence interval for $\theta$.

**Step 5.** Construct the empirical probability distribution of $\hat{\theta}_n^*$ by placing a probability of $1/B$ at each point of $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$. Then we can make inferences about $\theta$ by using this empirical distribution. This sampling distribution is usually presented by the histogram in the univariate case.

Since the replacement is allowed in the bootstrap procedure, it is virtually impossible that the bootstrap samples are identical to the original sample data when n is large enough. Intuitively, some data points may be duplicated whereas some other points from the original sample may be missed in a resample.

To show this concretely, let $X_1, \ldots, X_n$ be the sample set and $X_1^*, \ldots, X_n^*$ be a bootstrap sample.

Define the following random variables associated with the bootstrap sample:

$M^*_{1,n}$ = number of times that $X_1$ is chosen,

$M^*_{2,n}$ = number of times that $X_2$ is chosen,

...

$M^*_{n,n}$ = number of times that $X_n$ is chosen.

Namely, $M^*_{i,n} = \sum_{j=1}^{n} I(X^*_j = X_i)$, $i = 1, 2, \ldots, n$. It is obvious to see that $M^*_{1,n} + M^*_{2,n} + \cdots + M^*_{n,n} = n$ and $(M^*_{1,n}, M^*_{2,n}, \ldots, M^*_{n,n})$ has a multinomial distribution with parameters $n$ and $\pi = (1/n, \ldots, 1/n)$, $(M^*_{1,n}, \ldots, M^*_{n,n}) \sim Multi(n, 1/n, \ldots, 1/n)$. If the bootstrap sample is identical to the original sample set, $M^*_{i,n}$ must be 1 for $i = 1, 2, \ldots, n$. That is,

$$P(M^*_{1,n} = 1, \ldots, M^*_{n,n} = 1) = \frac{n!}{n^n} = \frac{n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}}{n^n} = \frac{\sqrt{2\pi n}}{e^n}$$

As we can see, the probability that the bootstrap sample is identical to the original sample is $\frac{\sqrt{2\pi n}}{e^n} \to 0$ when $n \to \infty$.

On the other hand, the number of all distinct bootstrap samples is $\binom{2n-1}{n}$. This can be demonstrated by counting all the *combinations with repetition*, that is

$$\binom{r+n-1}{r} = \frac{(r+n-1)!}{r!(n-1)!}$$

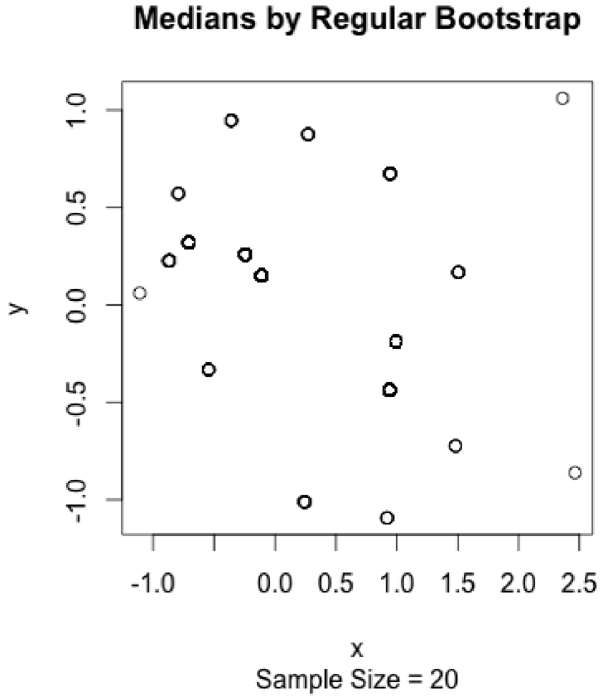
where $n$ the number of things to choose from and we choose $r$ of them with possible repetition. In our case, we have $r = n$ and whereby the number of all distinct bootstrap samples is $\binom{2n-1}{n}$.

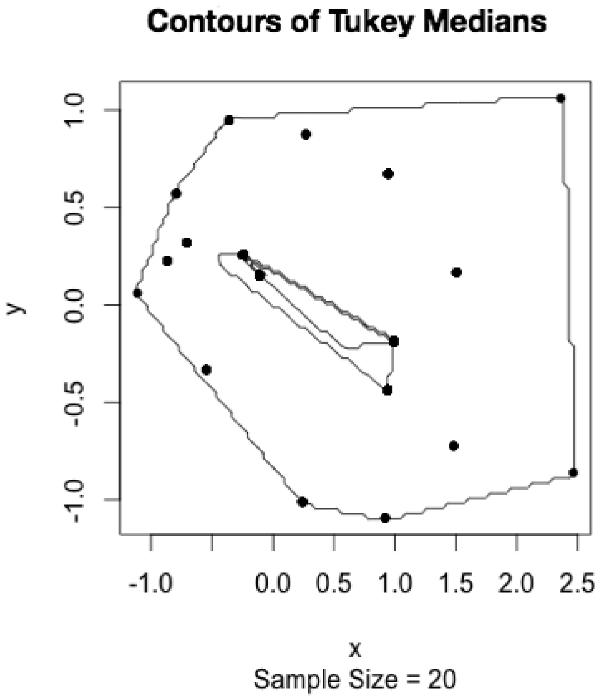
In the case of the $m$ out of $n$ bootstrap (discussed later), the number of all possible distinct bootstrap samples is $\binom{m+n-1}{n}$, where $r = m$.

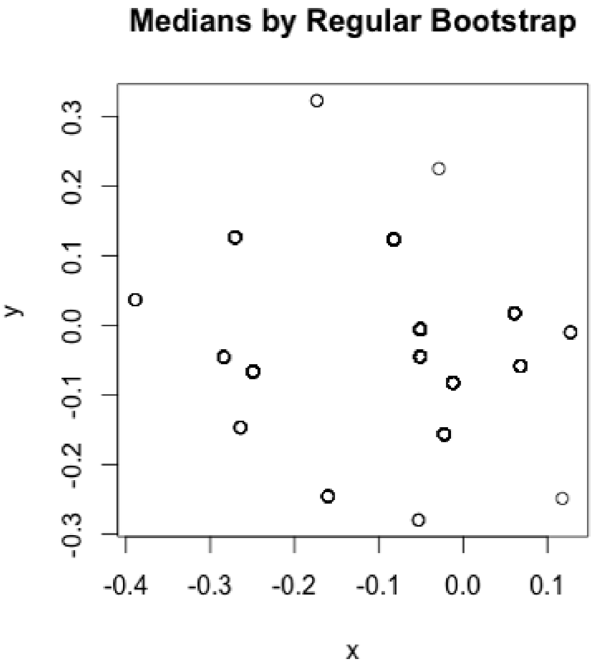
### 4.1.2 A Simulation Study of Regular Bootstrap for Tukey Median

**The Bivariate Normal Distribution**

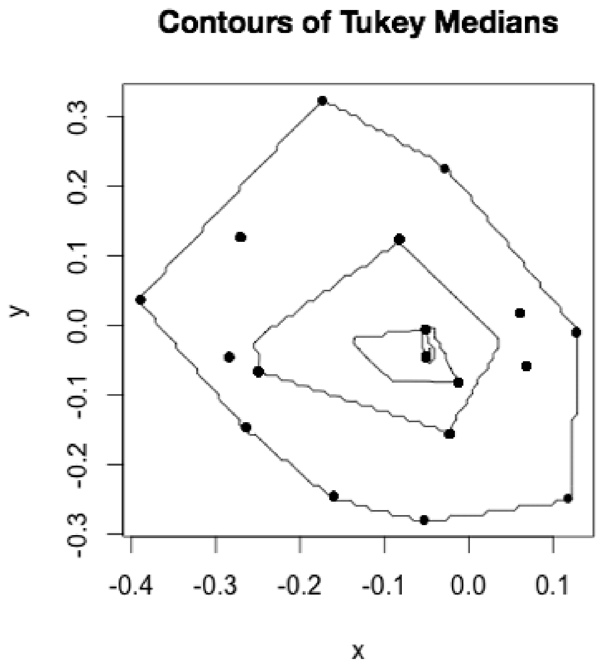
Draw three sample sets from the bivariate normal distribution $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right)$ of size 20, 200 and 500, denoted by $\mathcal{N}_{20}$, $\mathcal{N}_{200}$ and $\mathcal{N}_{500}$. We use the regular bootstrap for 1000 times ($B = 1000$) to generate bootstrap samples and calculate the Tukey median of each bootstrap sample. A collection of medians of size 1000 is obtained and shown in the following graphs.



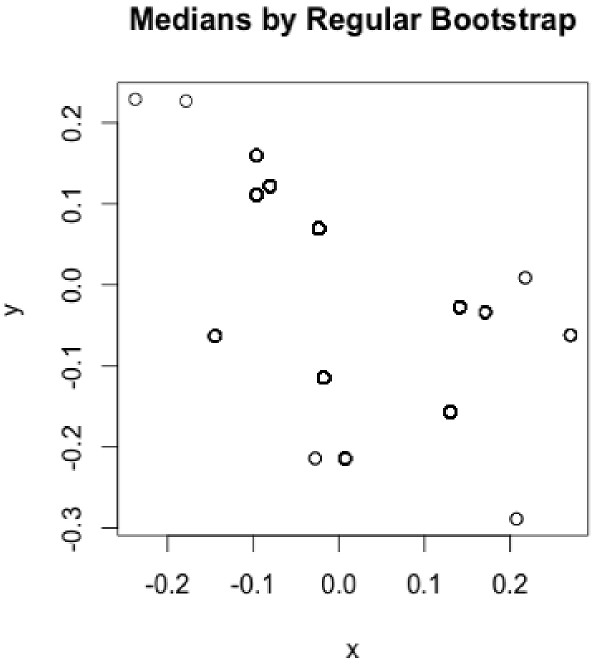




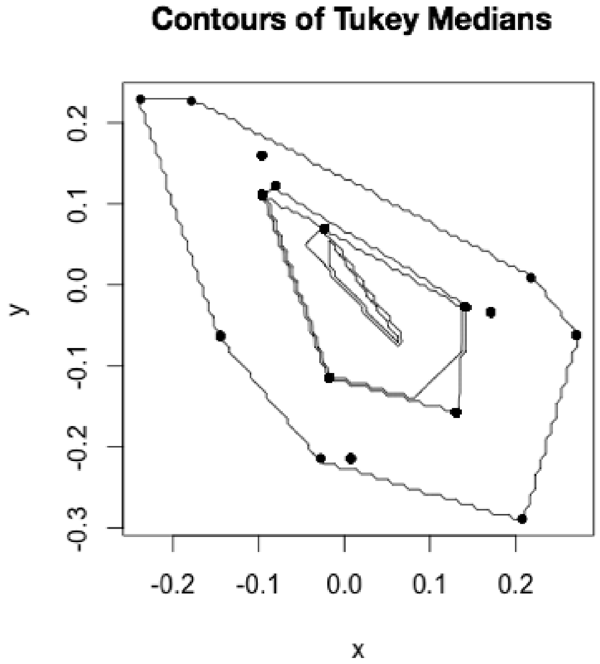
Sample Size = 200



Sample Size = 200



Sample Size = 500



Sample Size = 500

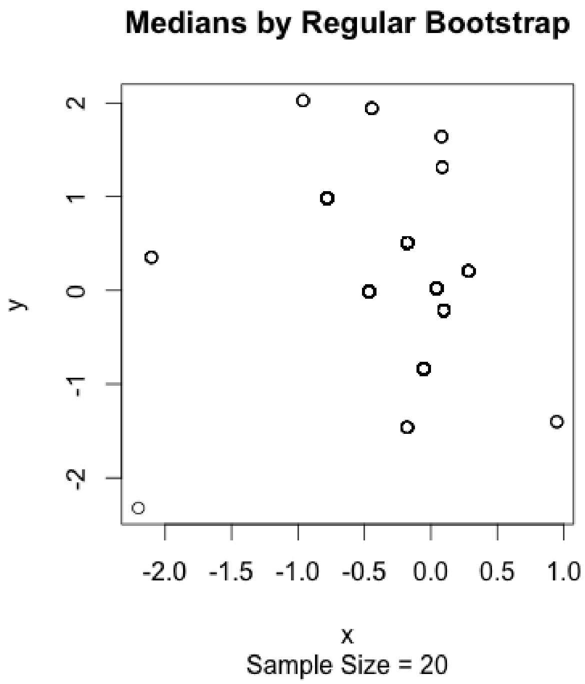
From the above, we can clearly see the Tukey medians are randomly distributed around the actual Tukey median (the original point) and the darker the point is, the greater the number of this point is. In the left-hand graphs, there are no more than 20 distinguished points of each collection of Tukey medians in our bootstrap samples no matter what the size of the original sample is. The resampling shows that the Tukey median of the bootstrap samples is not very sensitive from one sample to another sample. This is because the estimate for the Tukey median is a robust estimator and not sensitive to outliers. Applying the Tukey depth to the collection of Tukey medians and drawing quantile contour graphs as we do in Chapter 3, we get three graphs shown on the right with the quantile contours of 25%, 50%, 75% and 100% from the inside out. The 25% central region of each graph roughly falls around the actual Tukey median (the original point) and the larger sample size results in the 25% central region closer to the actual Tueky median (the original point). This is obvious as the sample is approaching to the population when the sample size is large enough.

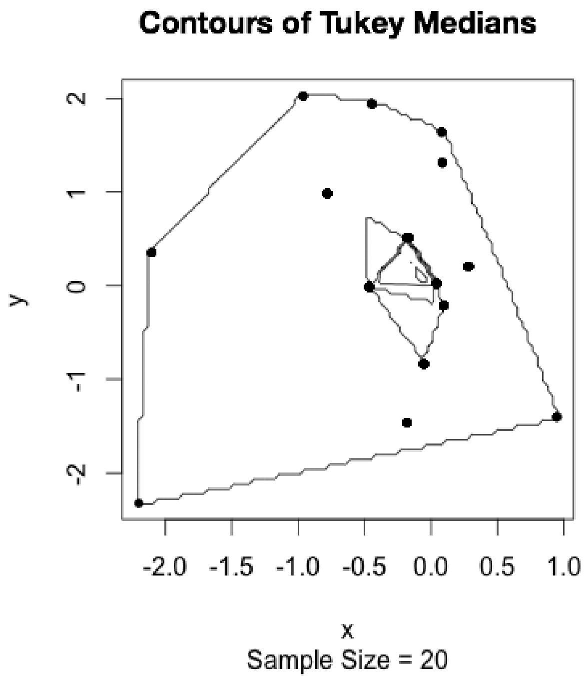
**The Bivariate Cauchy Distribution**

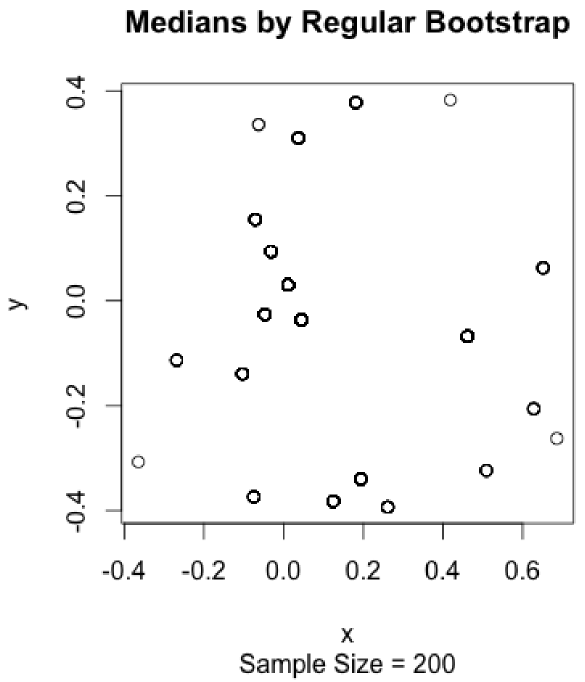
The following simulation is based on the sample sets drawn from the standard bivariate Cauchy distribution of size 20, 200 and 500, denoted by $\mathcal{C}_{20}$, $\mathcal{C}_{200}$ and $\mathcal{C}_{500}$. The distribution density of the standard bivariate Cauchy distribution is

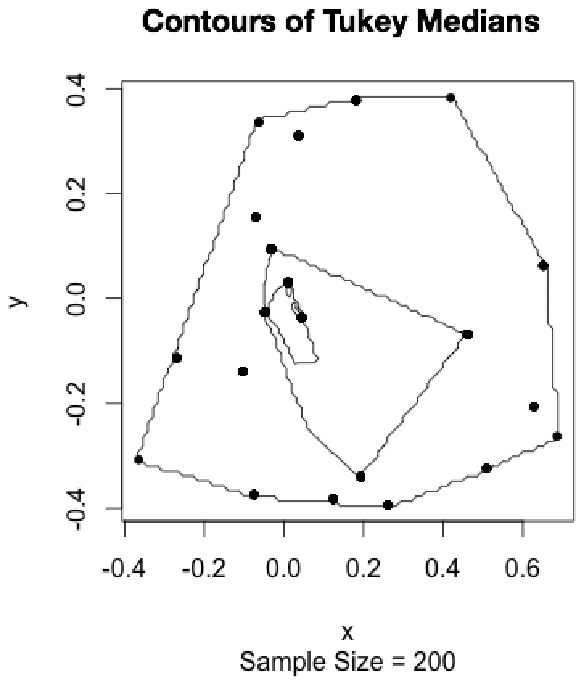
$$f(x, y) = \frac{1}{2\pi(1 + x^2 + y^2)^{3/2}} \qquad x, y \in \mathbb{R}.$$

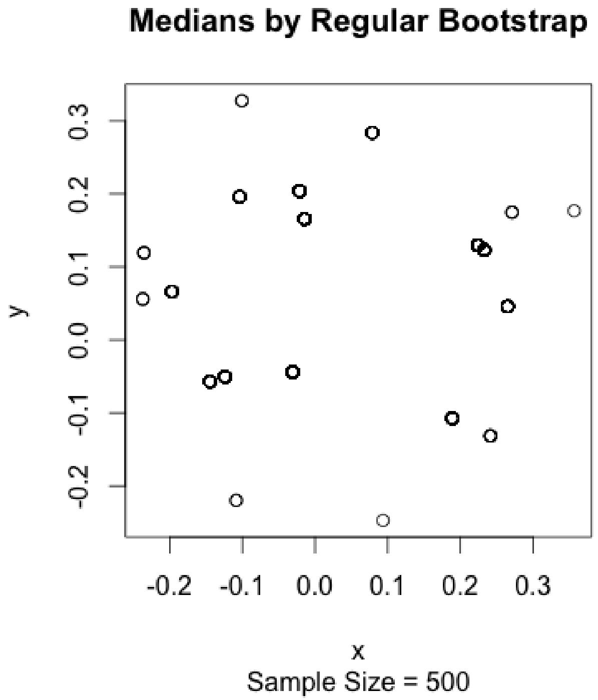
We continue using the regular bootstrap 1000 times and analyzing the collections of medians by applying the Tukey depth. This simulation is aimed to verify our results above. As we expect, the number of the distinct points in each collection of medians is still modest and does not change much with the increase of the sample size. The 25% central region still falls around the actual Tukey median (the original point). This shows us intuitively that the bootstrap Tukey median is consistent for estimating the Tukey median of the population. In the next section, we will discuss the general results of the consistency of the regular bootstrap.

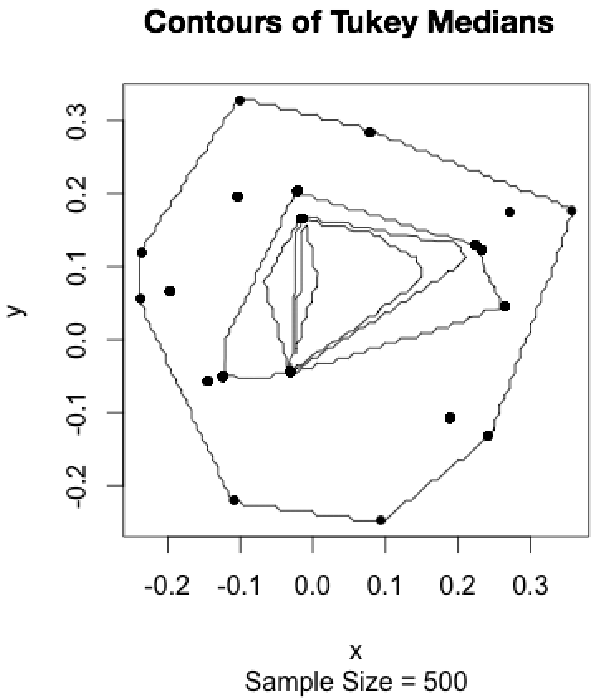








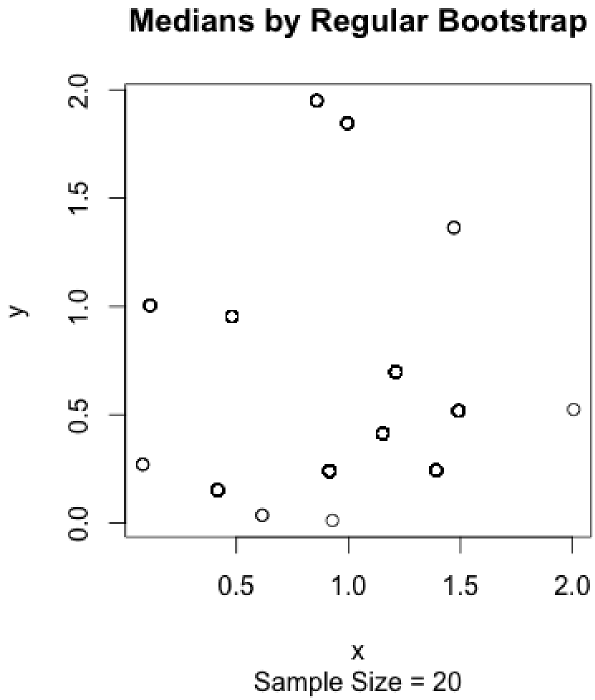


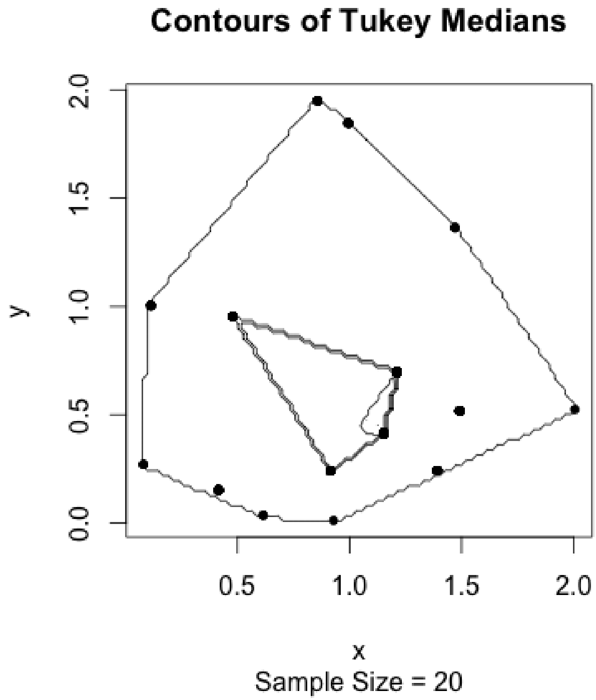


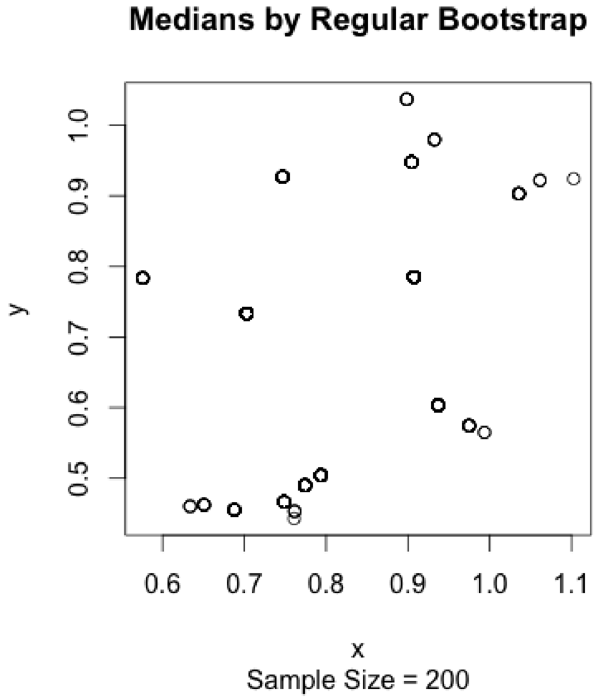
**The Bivariate Exponential Distribution**

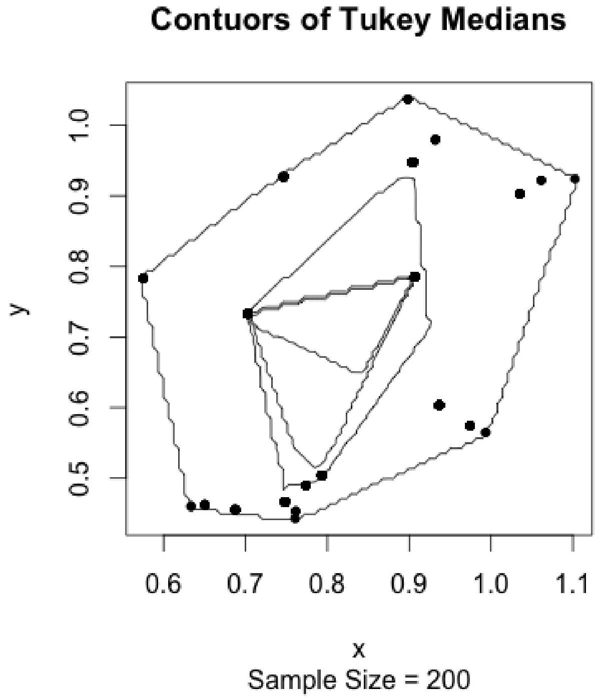
The following simulation is based on the sample sets drawn from the standard bivariate exponential distribution (i.e. the marginal distributions are standard exponential distribution and they are independent) of size 20, 200 and 500, denoted by $\mathcal{E}_{20}$, $\mathcal{E}_{200}$ and $\mathcal{E}_{500}$. We use the regular bootstrap for 1000 times and analyzing the collections of medians by applying the Tukey depth. The quantile contours represent 25%, 50%, 75% and 100% respectively, from the inside out.

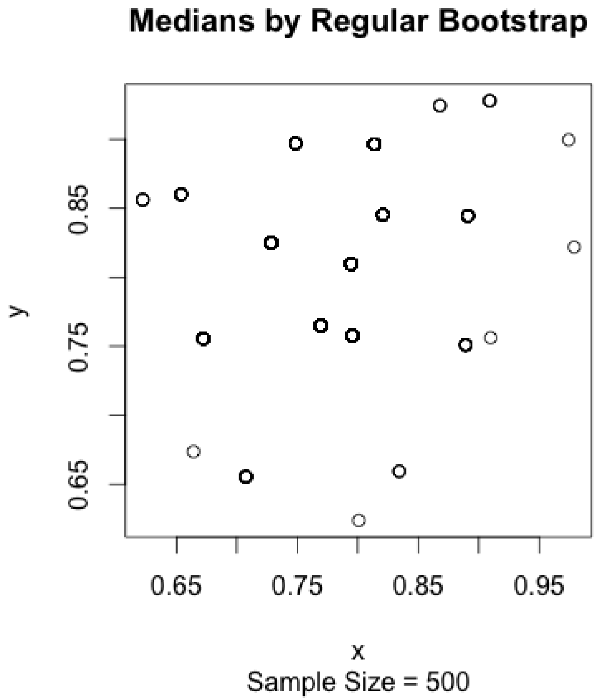
The actual Tukey median of the standard bivariate exponential distribution is very complicated to calculate. However, by using a sample of size 100,000 in our simulation, we can get the actual Tukey median which is approximately around (0.75, 0.75). From the contour graphs below, we can observe that the points is distributed around the actual Tukey median, which shows the bootstrap sample Tukey median is consistent.

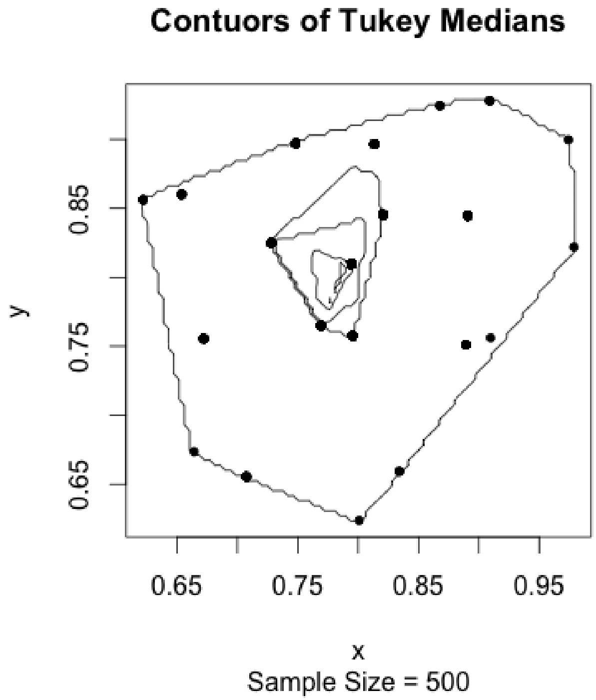

Medians by Regular Bootstrap
y
0.0
0.5
1.0
1.5
2.0
0.5
1.0
1.5
2.0
x
Sample Size = 20


Contours of Tukey Medians
y
0.0
0.5
1.0
1.5
2.0
0.5
1.0
1.5
2.0
x
Sample Size = 20


Medians by Regular Bootstrap
y
0.5
0.6
0.7
0.8
0.9
1.0
0.6
0.7
0.8
0.9
1.0
1.1
x
Sample Size = 200


Contuors of Tukey Medians
y
0.5
0.6
0.7
0.8
0.9
1.0
0.6
0.7
0.8
0.9
1.0
1.1
x
Sample Size = 200





### 4.1.3 The Validity of Regular Bootstrap

In this section, we review some existing results for the validity of the regular bootstrap. The multivariate case can be easily extended by the Cramér-Wold theorem.

Let $X_1, \dots, X_n$ be a sequence of i.i.d. observations coming from an unknown distribution $F$. Let $F_n$ be the empirical distribution of this sample and $X_1^*, \dots, X_n^*$ be the bootstrap sample drawn from $F_n$ independently. The random vector $(M_{1,n}^*, \dots, M_{n,n}^*)$ is defined previously and follows the multinomial distribution $Multi(n, 1/n, \dots, 1/n)$. We first study the properties of the bootstrap sample mean $\bar{X^*} = \frac{1}{n}\sum_{i=1}^n X_i^* = \frac{1}{n}\sum_{i=1}^n M_{i,n}^* X_i$ to find its relationship with the sample mean $\bar{X}$. Given the sample set $X_1, \dots, X_n$,

$$E(\bar{X^*}) = E\left(\frac{1}{n}\sum_{i=1}^n M_{i,n}^* X_i\right) = \frac{1}{n}\sum_{i=1}^n E(M_{i,n}^*) X_i = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$$

$$
\begin{aligned}
Var(\bar{X^*}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} M_{i,n}^* X_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(M_{i,n}^*)X_i^2 + 2\sum_{i<j} Cor(M_{i,n}^*, M_{j,n}^*)X_iX_j\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} n\left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)X_i^2 + 2\sum_{i<j}\left(-\frac{1}{n}\right)X_iX_j\right) \\
&= \frac{1}{n^2}\left(\frac{n-1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i<j} X_iX_j\right) = \frac{S_n^2}{n}
\end{aligned}
$$

The result here is extremely interesting as the calculation shows $E(\bar{X^*})$ and $Var(\bar{X^*})$ are exactly what we need in practice to construct a proper C.I. for the mean of an unknown distribution $F$. Historically, many authors have tried to establish asymptotic behaviour of the bootstrap estimator for its consistency to imitate the actual asymptotic distribution. Therefore, we would like to know $a_n(\hat{\theta}_n - \theta)$ converges to the same limiting distribution as $a_n(\hat{\theta}_n^* - \hat{\theta}_n)$, in order that we can use the bootstrap estimator to draw inferences about the real parameter $\theta$.

More formally, we rewrite $\hat{\theta}_n^* = \theta(F_n^*)$, $\hat{\theta}_n = \theta(F_n)$ and $\theta = \theta(F)$. In here $\theta(F)$ can be any parameter of the population, such as the mean, the variance and the median. The corresponding (bootstrap) estimators are the (bootstrap) sample mean, the (bootstrap) sample variance and the (bootstrap) sample median. The first two articles talking about the theoretical support of the bootstrap methods are from Singh (1981) and Bickel and Freedman (1981). They provide the asymptotic consistency of the bootstrap for a variety of statistics, such as the mean, the median, the empirical process and the quantile process.

### The Mean

For concreteness, we first consider the sample mean of $F$. In this simple case, $a_n = \sqrt{n}$. Let $X \sim F$ and $X_1, \ldots, X_n$ be an i.i.d. sample from $F$. Then we have $\theta(F) = \mu$, $\theta(F_n) = \bar{X}$ and $\theta(F_n^*) = \bar{X^*}$. By the central limit theorem, it is well-known that

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, Var(X)). \tag{4.1.1}$$

If the bootstrap works, our goal is to get the same asymptotic distribution by replacing the parameter $\mu$ by $\bar{X}$ and the estimator $\bar{X}$ by $\bar{X}^*$. That is

$$\sqrt{n}(\bar{X}^* - \bar{X}) \xrightarrow{d} N(0, Var(X)).$$

The following theorem states the consistency result of bootstrap sample mean. The proof can be found in a variety of places, such as Bickel and Freedman (1981) and Singh (1981).

**Theorem 4.1.1.** *Given a sample set $X_1, \ldots, X_n$ from $F$ with finite positive variance, let $\bar{X}$ be the sample mean and $\bar{X}^*$ be the bootstrap sample mean. Then as $n \to \infty$*

$$\sqrt{n}\left(\bar{X}^* - \bar{X}\right) \xrightarrow{d} N(0, Var(X)). \tag{4.1.2}$$

Singh (1981) provide more refined statements and proves the bootstrap method offers a better approximation than the central limit theorem in the case of standardized sample mean. Refer to Singh (1981) for more details.

**The Median**

Suppose $F$ is a continuously differentiable function which is differentiable at its median $\widetilde{M} = F^{-1}(1/2)$ and $f(\widetilde{M}) = F'(\widetilde{M}) > 0$. Let $X_1, \ldots, X_n$ be the i.i.d. random sample from $F$ and $\widetilde{M}_n = F_n^{-1}(1/2)$. Then we have

$$\sqrt{n}(\widetilde{M}_n - \widetilde{M}) \xrightarrow{d} N\left(0, \frac{1}{4f^2(\widetilde{M})}\right). \tag{4.1.3}$$

This result first appeared in Bickel and Freedman (1981), where they show that under the condition of the continuous differentiability of $F$ the bootstrap works almost surely.

**Theorem 4.1.2.** *Let $\widetilde{M}$, $\widetilde{M}_n$ and $\widetilde{M}_n^*$ be the medians of $F$, $F_n$ and $F_n^*$, respectively. Suppose that $F$ is continuously differentiable in a neighbourhood of $\widetilde{M}$ and the derivative $f(\widetilde{M}) > 0$. Then for almost all the sample paths $X_1, X_2, \ldots,$ given the sample set*

$X_1, \ldots, X_n$,

$$\sqrt{n}(\widetilde{M}_n^* - \widetilde{M}_n) \xrightarrow{d} N\left(0, \frac{1}{4f^2(\widetilde{M})}\right). \tag{4.1.4}$$

From (4.1.3) and (4.1.4), we can see the bootstrap estimator and the original estimator converge to the same asymptotic distribution. Therefore, $\sqrt{n}(\widetilde{M}_n - \widetilde{M}_n^*)$ can be used to the construct confidence intervals for $\widetilde{M}$.

More generally, let $Q_n(t) = \sqrt{n}(F_n^{-1}(t) - F^{-1}(t))$ be the quantile process and $Q_n^*(t) = \sqrt{n}(F_n^{*-1}(t) - F_n^{-1}(t))$ be the bootstrap quantile process.

It is well known that given $0 < \alpha \leq \beta < 1$, if $F$ is continuously differentiable with the positive derivative $f$ on $\mathbb{R}$, then,

$$Q_n(t) = \sqrt{n}(F_n^{-1}(t) - F^{-1}(t)) \xrightarrow{d} \frac{B(t)}{f(F^{-1}(t))} \tag{4.1.5}$$

where $B(t)$ is the Brownian bridge on [0,1] and $t \in [\alpha, \beta]$.

Bickel and Freedman (1981) justify the consistency of the bootstrap quantile process in the following theorem. Following this theorem, the asymptotic distribution of the median is one of the applications.

**Theorem 4.1.3.** *Let $F$ be a continuously differentiable function with the positive derivative $f$ on $\mathbb{R}$ and $X_1, \ldots, X_n$ be an i.i.d. random sample set from $F$. Define the bootstrap quantile process as $Q_n^*(t) = \sqrt{n}(F_n^{*-1}(t) - F_n^{-1}(t))$. Given $0 < \alpha \leq \beta < 1$, then for almost all the sample paths,*

$$Q_n^*(t) = \sqrt{n}(F_n^{*-1}(t) - F_n^{-1}(t)) \xrightarrow{d} \frac{B(t)}{f(F^{-1}(t))} \tag{4.1.6}$$

*where $B(t)$ is the Brownian bridge on [0,1] and $t \in [\alpha, \beta]$.*

### 4.1.4 The m out of n Regular Bootstrap

The bootstrap is consistent in most situations. However, the inconsistency sometimes occurs. The following simple example gives us an application where the bootstrap fails to work.

Let $X_1, \dots, X_n$ be an i.i.d. random sample drawn from the uniform distribution $F \sim \mathit{Unif}(0, \theta)$ and $\hat{\theta}_n = \max(X_1, \dots, X_n)$. Let $F_n$ be the empirical distribution of this sample and $X_1^* \dots, X_n^*$ be the bootstrap sample drawn from $F_n$ independently. Define $\hat{\theta}_n^* = \max(X_1^*, \dots, X_n^*)$. Then we have

$$\frac{n(\theta - \hat{\theta}_n)}{\theta} \xrightarrow{d} \mathcal{E}xp(1)$$

where $\mathcal{E}xp(1)$ represents the standard exponential distribution.

It is natural to expect $n(\hat{\theta}_n - \hat{\theta}_n^*)/\hat{\theta}_n \xrightarrow{d} \mathcal{E}xp(1)$ if the bootstrap estimator $\hat{\theta}_n^*$ is asymptotic consistent. However, noticing that

$$\begin{aligned} P(n(\hat{\theta}_n - \hat{\theta}_n^*) = 0 | X_1, \dots, X_n) &= P(\hat{\theta}_n^* = \hat{\theta}_n | X_1, \dots, X_n) \\ &= 1 - P(\hat{\theta}_n^* < \hat{\theta}_n | X_1, \dots, X_n) \\ &= 1 - P(X_1^*, \dots, X_n^* < \hat{\theta}_n | X_1, \dots, X_n) \\ &= 1 - \left(\frac{n-1}{n}\right)^n \\ &= 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \qquad (n \to \infty) \end{aligned}$$

Whereas $P(n(\theta - \hat{\theta}_n) = 0) \to 0$ which results in the inconsistency of the bootstrap estimator.

More generally, let $X_{(1)}, \dots, X_{(n)}$ be the order statistics. Then

$$P(\hat{\theta}_n^* < X_{(n-k+1)} | X_1, \dots, X_n) = \left(1 - \frac{k}{n}\right)^n \to e^{-k} \qquad (n \to \infty)$$

The limiting distribution of the bootstrap estimator is actually a random distribution and $n(\hat{\theta}_n - \hat{\theta}_n^*)/\hat{\theta}_n$ does not have a weak limit.

We will introduce an alternative resampling method whose conditions for the consistency are weaker than those for the bootstrap. The idea is to draw a bootstrap sample of size $m$

instead of size $n$, where $m \to \infty$ and $m/n \to 0$. This new resampling scheme is called the *$m$ out of $n$ bootstrap*. It usually resolves the inconsistency problem of the regular bootstrap. Bickel, Götze and van Zwet (1997) have reviewed the successes and failure of the regular bootstrap and give a variety of results on how $m$ out of $n$ bootstrap works. They also indicate that in the successful cases of the bootstrap, the $m$ out of $n$ bootstrap is not necessary and generally leads to less efficiency.

The following example is an application of the $m$ out of $n$ bootstrap from Swanepoel (1986), which remedies the inconsistency of the regular bootstrap for estimating the maximum of an uniform sample.

**Theorem 4.1.4** (Swanepoel (1986))**.** *Let $X_1, \ldots, X_n$ be an i.i.d. random sample from the uniform distribution $F \sim \mathrm{Unif}(0, \theta)$ and $X_{(1)}, \ldots, X_{(n)}$ be the order statistics. Let $F_n$ be the empirical distribution of this sample and $X_1^*, \ldots, X_m^*$ be the bootstrap sample drawn from $F_n$ independently. Define $\hat{\theta}_n = X_{(n)}$ and $\hat{\theta}_m^* = \max(X_1^*, \ldots, X_m^*)$. Let $An^{-\epsilon} \le m \le o(n^{\frac{\epsilon+1}{2}}(\mathrm{ln}n)^{-\frac{1}{2}})$, where $A$ is a positive constant and $\epsilon \in (0,1)$. Then for almost all the sample paths, $m(\hat{\theta}_n - \hat{\theta}_m^*)/\hat{\theta}_n \xrightarrow{d} \mathcal{E}xp(1)$, i.e. the bootstrap estimator $\hat{\theta}_m^*$ is consistent.*

**Proof:** Define a new sample set $U_1 = \frac{X_1}{\theta}, U_2 = \frac{X_2}{\theta}, \ldots, U_n = \frac{X_n}{\theta}$ with corresponding order statistics $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$. Then $U_i \sim \mathit{Unif}(0,1)$ for $i = 1, \ldots, n$ and let $\hat{F}_n$ be the empirical distribution.

Denote the conditional probability as $P^*(\cdot) = P(\cdot|F_n)$, then

$$\begin{aligned} P^*\left(\frac{m(\hat{\theta}_n - \hat{\theta}_m^*)}{\hat{\theta}_n} \le x\right) &= P^*\left(\hat{\theta}_m^* \ge \hat{\theta}_n\left(1 - \frac{x}{m}\right)\right) \\ &= 1 - P^*\left(\hat{\theta}_m^* \le \hat{\theta}_n\left(1 - \frac{x}{m}\right)\right) + P^*\left(\hat{\theta}_m^* = \hat{\theta}_n\left(1 - \frac{x}{m}\right)\right) \end{aligned} \tag{4.1.7}$$

Noticing that,

$$\begin{aligned}&P^*\left(\hat{\theta}^*_m \leq \hat{\theta}_n\left(1-\frac{x}{m}\right)\right) = P^*\left(X^*_1,\ldots,X^*_m \leq \hat{\theta}_n\left(1-\frac{x}{m}\right)\right)\\ &= \left(P^*\left(X_1 \leq \hat{\theta}_n\left(1-\frac{x}{m}\right)\right)\right)^m = \left(F_n\left(X_{(n)}\left(1-\frac{x}{m}\right)\right)\right)^m\\ &= \left(\hat{F_n}\left(U_{(n)}\left(1-\frac{x}{m}\right)\right)\right)^m\end{aligned} \tag{4.1.8}$$

Now, we define the uniform empirical process as

$$u_n(t) = \sqrt{n}(\hat{F_n}(t) - t) \qquad 0 \leq t \leq 1 \tag{4.1.9}$$

It is well-known that $X_{(n)} = \theta + O((\mathrm{lnln}n)/n)$ as $n \to \infty$, i.e. $U_{(n)} = 1 + O((\mathrm{lnln}n)/n)$. Then, along with the condition $An^{-\epsilon} \leq m \leq o(n^{\frac{\epsilon+1}{2}}(\mathrm{ln}n)^{-\frac{1}{2}})$ and $n$ large enough, we have

$$\left|U_{(n)}\left(1-\frac{x}{m}\right) - 1\right| \leq \frac{x}{m} + O(\frac{\mathrm{lnln}n}{n}) \leq Bn^{-\epsilon} \tag{4.1.10}$$

where $B$ is a constant and $\epsilon \in (0,1)$. Denote $a_n = Bn^{-\epsilon}$ in (4.1.10) and refer to Theorem 0.2 of Stute (1982). For $n$ large enough, we get

$$\begin{aligned}&\left|u_n\left(U_{(n)}\left(1-\frac{x}{m}\right)\right) - u_n(1)\right| = \left|\sqrt{n}\hat{F_n}\left(U_{(n)}\left(1-\frac{x}{m}\right)\right) - \sqrt{n}U_{(n)}\left(1-\frac{x}{m}\right)\right|\\ &\leq \sup_{|t-s|\leq a_n} |u_n(t) - u_n(s)| = O\left(n^{-\frac{\epsilon}{2}}(\mathrm{ln}n)^{\frac{1}{2}}\right)\end{aligned} \tag{4.1.11}$$

Thus, along with (4.1.9) and $U_{(n)} = 1 + O((\mathrm{lnln}n)/n)$,

$$\begin{aligned}\hat{F_n}\left(U_{(n)}\left(1-\frac{x}{m}\right)\right) &= U_{(n)}\left(1-\frac{x}{m}\right) + O(n^{-\frac{\epsilon+1}{2}}(\mathrm{ln}n)^{\frac{1}{2}})\\ &= \left(1-\frac{x}{m}\right) + O(n^{-\frac{\epsilon+1}{2}}(\mathrm{ln}n)^{\frac{1}{2}})\end{aligned} \tag{4.1.12}$$

Then, from (4.1.8) and (4.1.12), when $n \to \infty$,

$$\begin{aligned}&P^*\left(\hat{\theta}^*_m \leq \hat{\theta}_n\left(1-\frac{x}{m}\right)\right) = \left(\hat{F_n}\left(U_{(n)}\left(1-\frac{x}{m}\right)\right)\right)^m\\ &= \left(\left(1-\frac{x}{m}\right) + O\left(n^{-\frac{\epsilon+1}{2}}(\mathrm{ln}n)^{\frac{1}{2}}\right)\right)^m \to e^{-x}\end{aligned} \tag{4.1.13}$$

Swanepoel indicates that in exactly the same way as above, we can also get

$$P^*\left(\hat{\theta}_m^* \leq X_{(n-1)}\left(1-\frac{x}{m}\right)\right) \to e^{-x} \qquad as \quad n \to \infty$$

Therefore, as $n \to \infty$,

$$P^*\left(\hat{\theta}_m^* = \hat{\theta}_n\left(1-\frac{x}{m}\right)\right) = P^*\left(\hat{\theta}_m^* \leq \hat{\theta}_n\left(1-\frac{x}{m}\right)\right) - P^*\left(\hat{\theta}_m^* \leq X_{(n-1)}\left(1-\frac{x}{m}\right)\right) \to 0 \tag{4.1.14}$$

From (4.1.7), (4.1.13) and (4.1.14), it follows that

$$P^*\left(\frac{m(\hat{\theta}_n - \hat{\theta}_m^*)}{\hat{\theta}_n} \leq x\right) \to 1 - e^{-x}$$

i.e.

$$\frac{m(\hat{\theta}_n - \hat{\theta}_m^*)}{\hat{\theta}_n} \xrightarrow{d} \mathcal{E}xp(1)$$

∎

Zarepour and Knight (1999) derive a more general result for bootstrapping extremes. It is known that if there exists $a_n > 0$ and $b_n$ such that $a_n(X_{(n)} - b_n) \xrightarrow{d} Y$ as $n \to \infty$, then $Y$ is a random variable with one of the three types of distributions and in each case, the regular bootstrap fails asymptotically but the $m$ out of $n$ bootstrap works by setting $m = o(n) \to \infty$. (See Zarepour and Knight (1999)).

For more specific applications of the $m$ out of $n$ bootstrap, refer to Bickel and Yahav (1988), Hall (1990), Bickel and Ren (1996) and Andrews (2000).

Besides, a crucial question remains subjective as the choice of $m$ is not clear, since $m = \sqrt{n}$, $m = n^{2/3}$ and $m = \ln n$ all are of type $o(n)$. Many articles have discussed this question, for example, see Datta and McCormick (1995), Hall, Horowitz and Jing (1995), Politis, Romano and Wolf (1999), Sakov (1998), Götze and Račkauskas (2001) and Bickel and Sakov (2008).

## 4.2 Bayesian Bootstrap

### 4.2.1 Rubin's Bayesian Bootstrap

The idea of the Bayesian bootstrap was first introduced by Rubin (1981). It can be interpreted as an analogue of the regular bootstrap by placing a random probability to each sample point instead of a fixed probability of $1/n$. This random probability is called a posterior probability. In Rubin's Bayesian bootstrap, the posterior probability is constructed by generating $n-1$ uniform random numbers on $[0,1]$ and assigning the gaps of them to the sample points. Besides, Rubin also proved the Bayesian bootstrap inference about the statistics of interest will be similar to the regular bootstrap inference and therefore the criticisms of the Bayesian bootstrap are the same as that of the regular bootstrap.

Given an i.i.d. sample $X_1, \ldots, X_n$ randomly drawn from the population distribution $F$ where $F$ is completely unknown, let $\theta$ be the unknown parameter of $F$ and $\hat{\theta}_n$ be an estimate for $\theta$. The basic steps of the Rubin's Bayesian bootstrap can be presented as follows.

**Step 1.** Simulate $n-1$ random numbers on $[0,1]$ from uniform distribution *Unif*$(0,1)$ (independent of the sample points $X_1, X_2, \ldots, X_n$), denoted as $U_1, U_2, \ldots, U_{n-1}$. Define the ordered statistics by $0 = U_{(0)} < U_{(1)} < \cdots < U_{(n-1)} < U_{(n)} = 1$ and assign each sample point $X_i$ a weight $w_i = U_{(i)} - U_{(i-1)}$. A discrete distribution function $G_n$ of the sample points is constructed, called the posterior distribution.

**Step 2.** Calculate the statistics of interest, $\hat{\theta}_n$, for this posterior distribution $G_n$, denoted by $\hat{\theta}_n^*$. This is the Bayesian bootstrap estimator for $\theta$.

**Step 3.** Repeat steps 2 and 3 a large number of times, say $B$ times, to generate B posterior distributions and whereby get a collection of the statistics of interest, denoted by $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

**Step 4.** Put a probability of $1/B$ on each point $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ and construct an empirical distribution for $\hat{\theta}_n^*$. Then we can analyze $\theta$ by this empirical distribution.

**Remark 4.2.1.** If $t_{(1)}, \ldots, t_{(n)}$ are the ordered statistics of a sample from a continuous distribution function $F_0(t)$, and define $u_1 = F_0(t_{(1)})$, $u_2 = F_0(t_{(2)}) - F(t_{(1)})$, $\ldots$, $u_n = F_0(t_{(n)}) - F_0(t_{(n-1)})$, then $u_1, \ldots, u_n$ are random variables following the $n$-variate flat Dirichlet distribution $Dir(1, \ldots, 1)$. (Wilks, 1962). Therefore, the corresponding prior distribution for the parameter $\theta$ of the population distribution $F$ in the Rubin's Bayesian bootstrap is actually an $n-1$ variate flat Dirichlet distribution.

Moreover, Rubin also demonstrated by the following theorem that his Bayesian bootstrap is actually simulating the posterior distribution of the parameter $\theta$ under a specific model. Operationally and inferentially, the Rubin's Bayesian bootstrap and the regular bootstrap have much similarity.

**Theorem 4.2.2** (Rubin (1981))**.** *Let $X$ be a random variable and $d = (d_1, \ldots, d_K)$ be the vector of all possible distinct values of $X$. Let $\theta = (\theta_1, \ldots, \theta_K)$ be the vector of the corresponding probabilities for $d$, i.e. $P(X = d_k|\theta) = \theta_k$ for $k = 1, 2, \ldots, K$ and $\sum_{k=1}^K \theta_k = 1$. Let $X_1, \ldots, X_n$ be $n$ i.i.d. random observations of $X$ and $n_k = \sum_{i=1}^n I(X_i = d_k)$ for $k = 1, 2, \ldots, K$. If $\theta$ has a Dirichlet prior distribution $Dir(l_1 + 1, \ldots, l_K + 1)$, then the posterior distribution of $\theta$ is the $K-1$ variate Dirichlet distribution $Dir(n_1 + l_1 + 1, \ldots, n_K + l_K + 1)$.*

In practice, let $m = n + K + \sum_{k=1}^K l_k$ and $U_1, \ldots, U_{m-1}$ be i.i.d. random numbers from *Unif*$(0, 1)$ with corresponding order statistics $U_{(1)}, \ldots, U_{(m-1)}$. Define $g_1 = U_{(1)}$, $g_2 = U_{(2)} - U_{(1)}$, $\ldots$, $g_m = 1 - U_{(m-1)}$. Then divide the $g_1, \ldots, g_m$ into $K$ groups, so that there are $n_k + l_k + 1$ items in the $k$-th group. Therefore, the weight $w_k$ for each possible value $d_k$ of $X$ is the sum of all the $g_i$ in the $k$-th group. Furthermore, by Wilks (1962), we have that $(w_1, \ldots, w_K)$ follows the Dirichlet distribution $Dir(n_1 + l_1 + 1, \ldots, n_K + l_K + 1)$. Lo (1987) extends the Rubin's Bayesian bootstrap to a more general form by using a Dirichlet process prior. We will introduce the Lo's Bayesian bootstrap in the next section.

Moreover, Lo also verifies the validity of the Rubin's Bayesian bootstrap in the following theorems.

**Theorem 4.2.3** (Lo (1987))**.** *Given an i.i.d. random sample* $X_1, \ldots, X_n$ *from* $F$*, where* $F$ *is an unknown continuous distribution function, let* $F_n$ *be the empirical distribution and* $G_n$ *be a posterior distribution. Define a process* $d_n = \sqrt{n}(G_n - F_n)$*. Then there exists a Brownian bridge* $\{B(s) : 0 \leq s \leq 1\}$ *independent of* $X_1, \ldots, X_n$*, such that*

$$\sup_x |d_n(x) - B(F(x))| \to 0 \qquad a.s. \quad as \quad n \to \infty \tag{4.2.1}$$

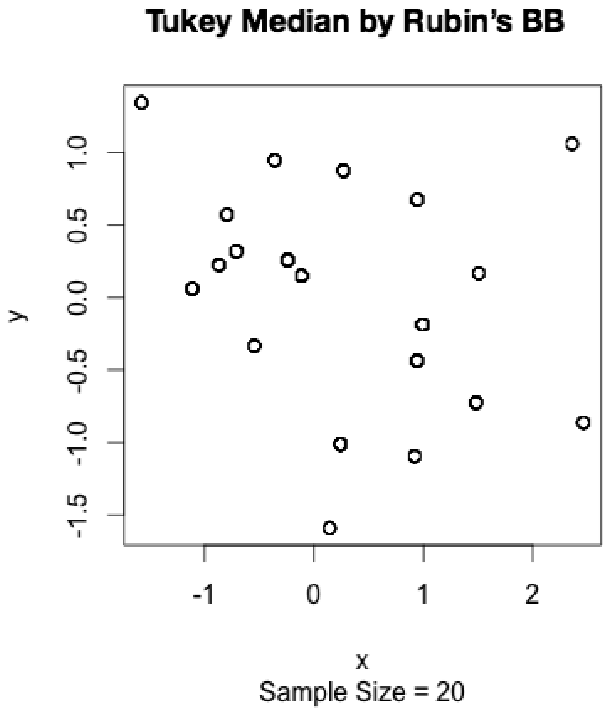
**Theorem 4.2.4** (Lo (1987))**.** *Let* $X_1, \ldots, X_n$ *be an i.i.d. random sample from* $F$ *and* $F_n$ *is the empirical distribution. Define* $\mu(F)$*,* $\mu(F_n)$ *and* $\mu(G_n)$ *as the population mean, the sample mean and the Rubin's Bayesian bootstrap sample mean, respectively. If* $\int x^2 F(dx) < \infty$*, then as* $n \to \infty$*,*

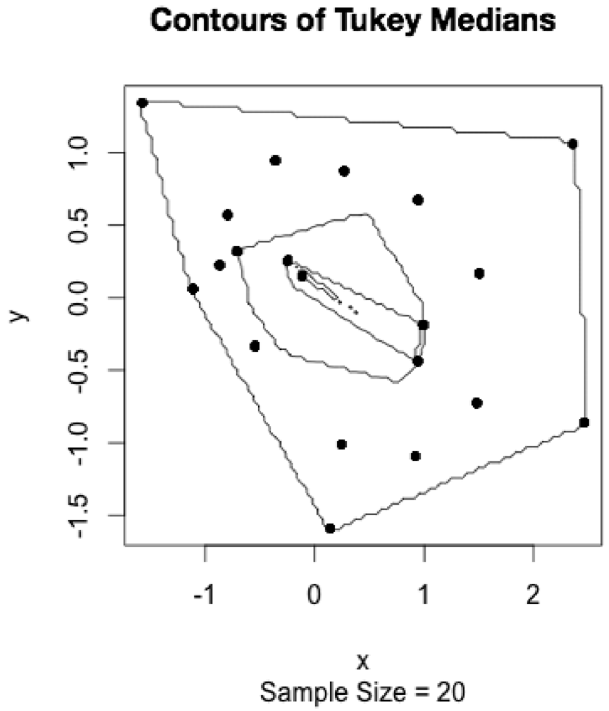
$$\sqrt{n}\left(\mu(G_{\alpha n}) - \mu(F_n)\right) \xrightarrow{d} N\left(0, Var(X)\right). \tag{4.2.2}$$

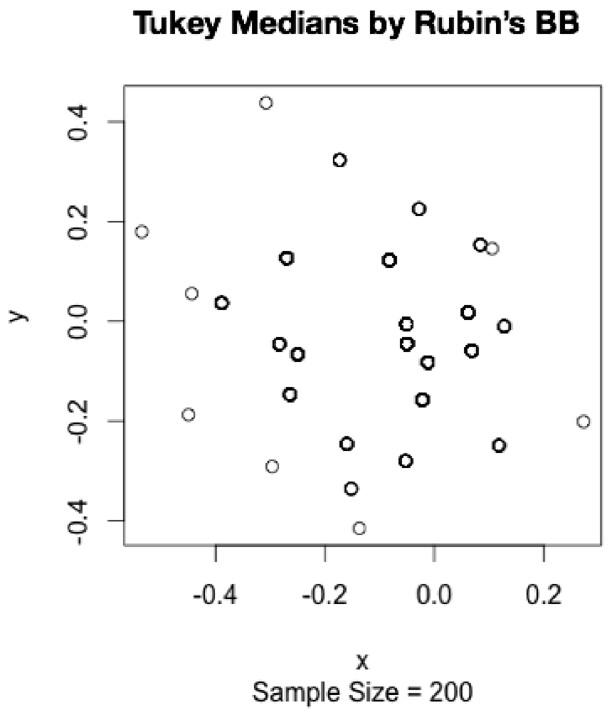
### 4.2.2 A Simulation study of Rubin's Bayesian Bootstrap for Tukey Median

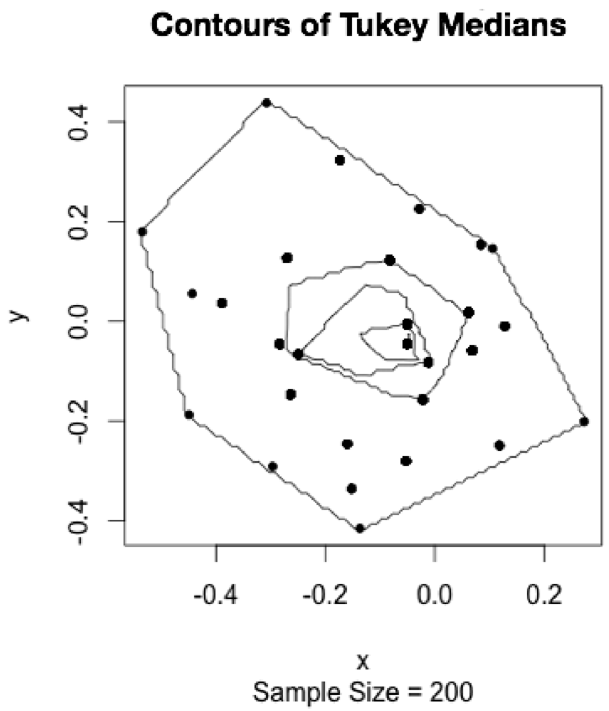
#### The Bivariate Normal Distribution

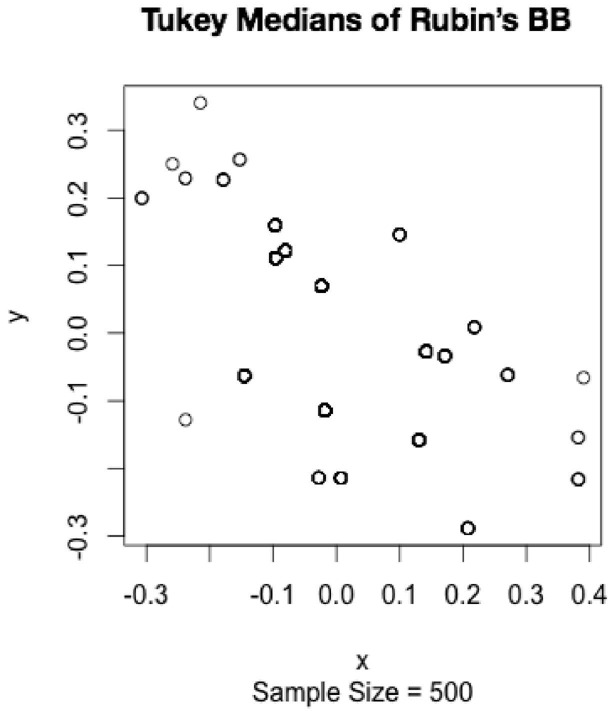
The simulation in this section is based on the sample sets $\mathcal{N}_{20}$, $\mathcal{N}_{200}$ and $\mathcal{N}_{500}$. To each sample set, we use the Rubin's Bayesian bootstrap for 1000 times, namley $B = 1000$, and get 1000 Rubin's Bayesian bootstrap samples. By applying the Tukey depth, we can get a Tukey median for each bootstrap sample. Then, a collection of Tukey medians of size 1000 is obtained for each sample set and shown in the following graphs.

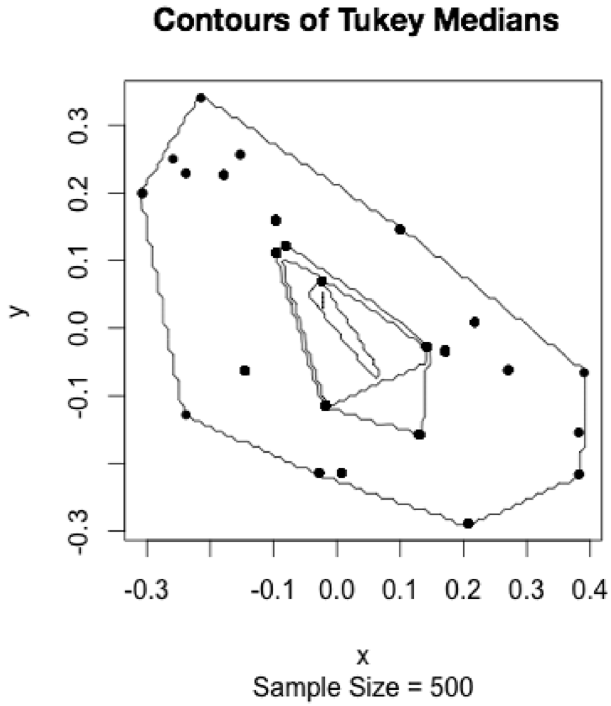

Tukey Median by Rubin's BB
y
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1
0
1
2
x
Sample Size = 20


Contours of Tukey Medians
y
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1
0
1
2
x
Sample Size = 20


Tukey Medians by Rubin's BB
y
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
x
Sample Size = 200


Contours of Tukey Medians
y
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
x
Sample Size = 200

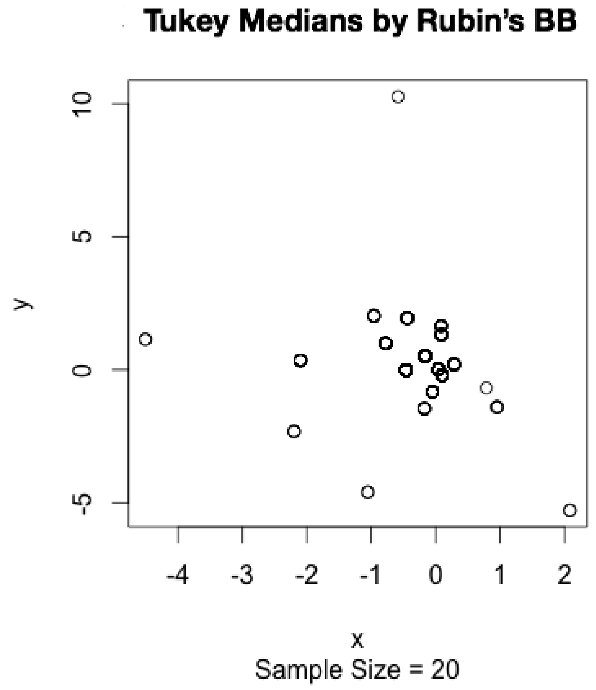


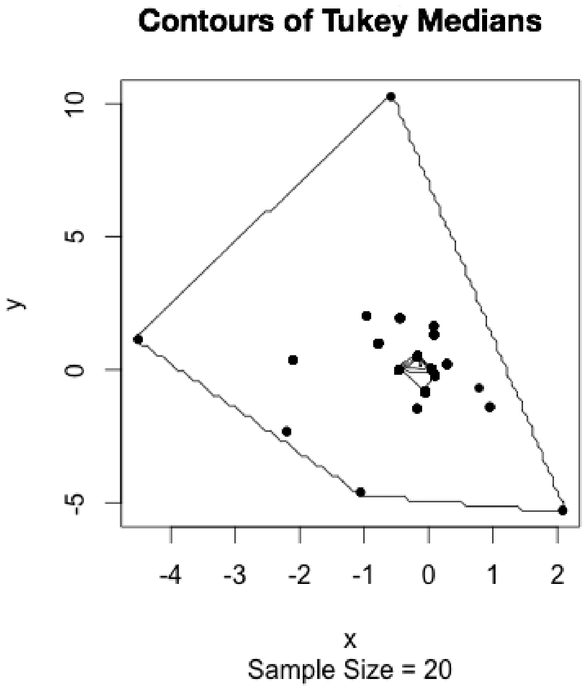


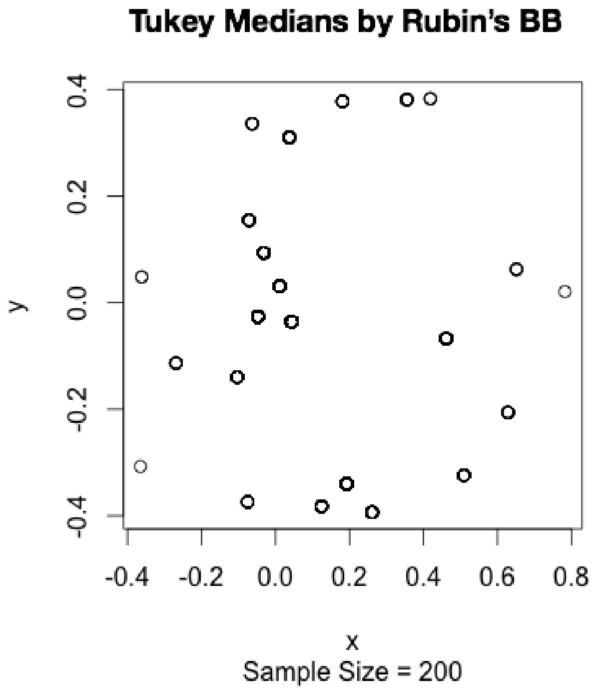
From the graphs above, it is clear to observe that the number of the distinct points in the left-hand graphs increases a little bit, compared to the graphs we presented before for the regular bootstrap. This shows that the diversity of the Rubin's Bayesian bootstrap samples increases to some extent. This is because the weight of each sample point is not the same anymore, so that some points are replicated more often whereas other points have a larger chance to be omitted. Therefore, we get more distinct Tukey median points. However, the number of the distinct Tukey medians in each graph is still modest. This comes from the fact that the Tukey median is a robust measurement. The darker points represent the greater number of observed points in the resampling simulation. Applying the Tukey depth to the collection of Tukey medians and drawing quantile contour graphs, we get three contour graphs shown on the right with the quantile contours of 25%, 50%, 75% and 100% from the inside out. The 25% central region of each contour graph is still around the actual Tukey median (the original point) which is the same as the regular bootstrap. Actually, the regular bootstrap and Rubin's Bayesian bootstrap behave in a similar way when the sample size is large enough.

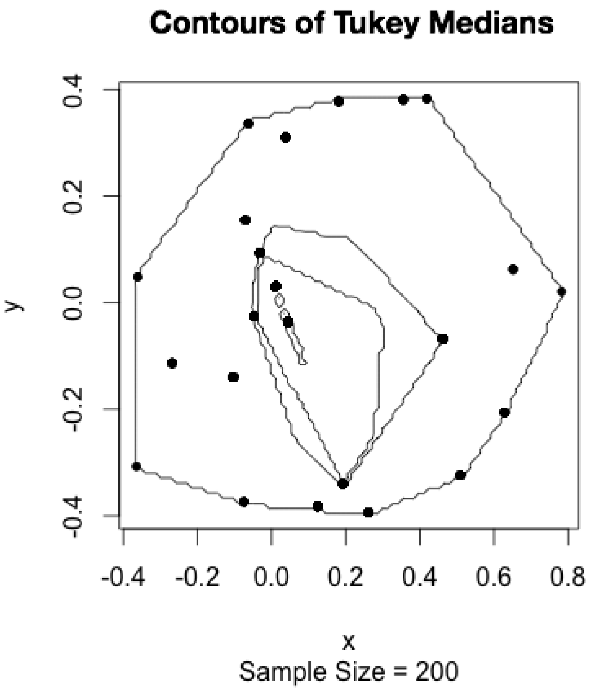
### The Bivariate Cauchy Distribution

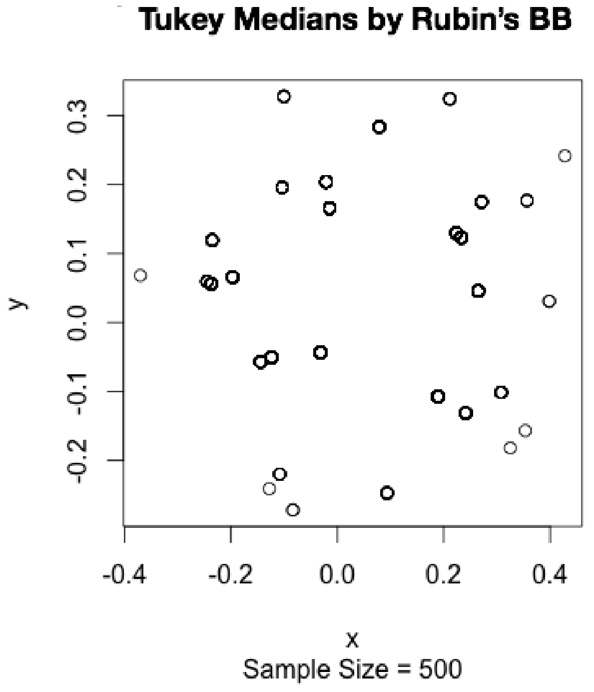
Applying the Rubin's Bayesian bootstrap and the Tukey depth for 1000 times, to the sample sets $\mathcal{C}_{20}$, $\mathcal{C}_{200}$ and $\mathcal{C}_{500}$ respectively, we get a collection of Tukey medians for each sample set. The scattergrams and the contour graphs of the collection of Tukey medians are shown below. The quantile contours represent 25%, 50%, 75% and 100% respectively, from the inside out. As we can see, the points are distributed around the actual Tukey median (the original point). Here, the analysis is very similar to the previous cases.

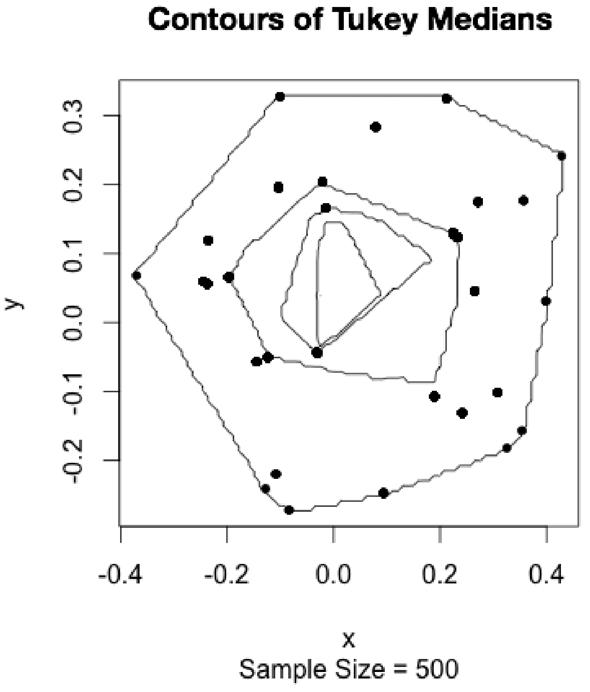





Tukey Medians by Rubin's BB
y
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
x
Sample Size = 200

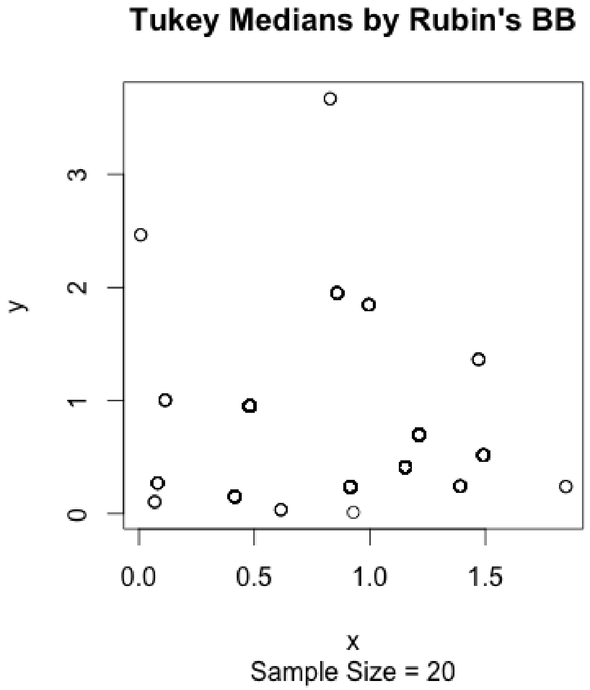

Contours of Tukey Medians
y
0.4
0.2
0.0
-0.2
-0.4
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
x
Sample Size = 200

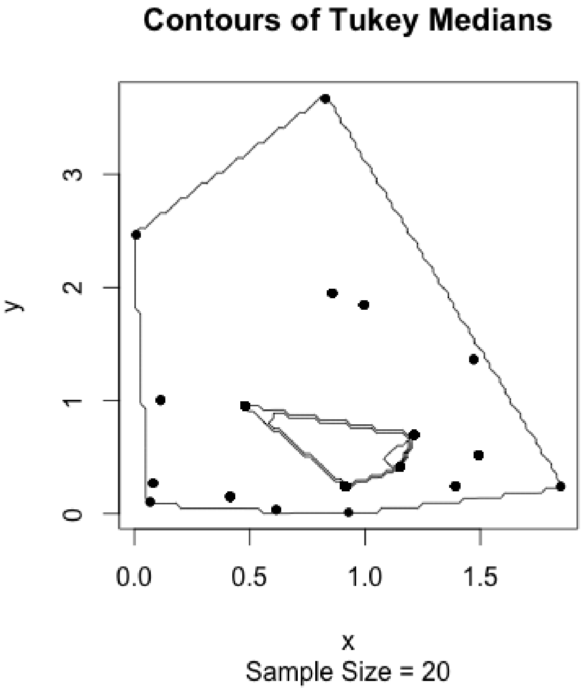

Tukey Medians by Rubin's BB
y
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.2
0.0
0.2
0.4
x
Sample Size = 500

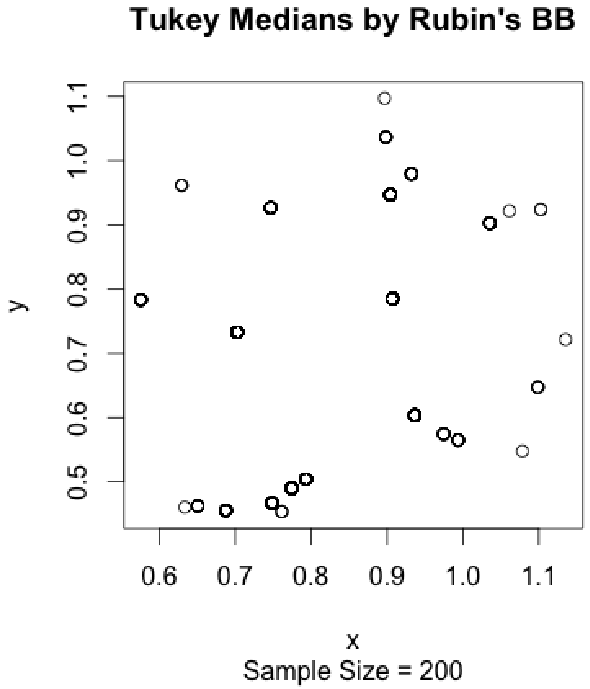

Contours of Tukey Medians
y
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.2
0.0
0.2
0.4
x
Sample Size = 500

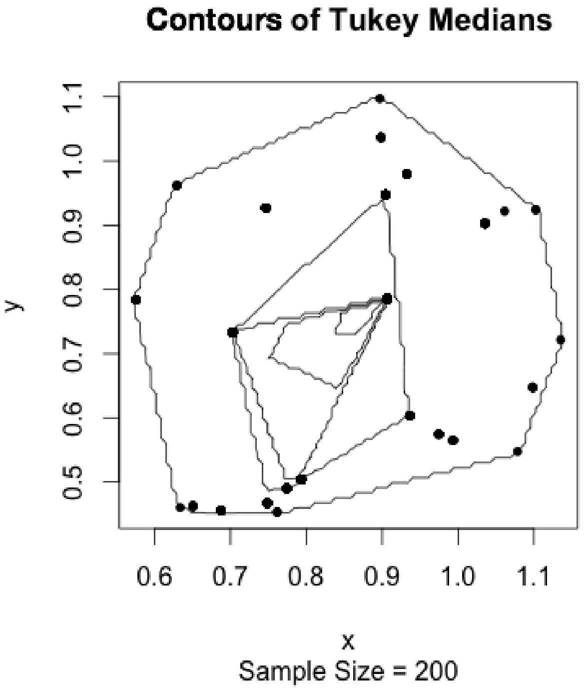
### The Bivariate Exponential Distribution

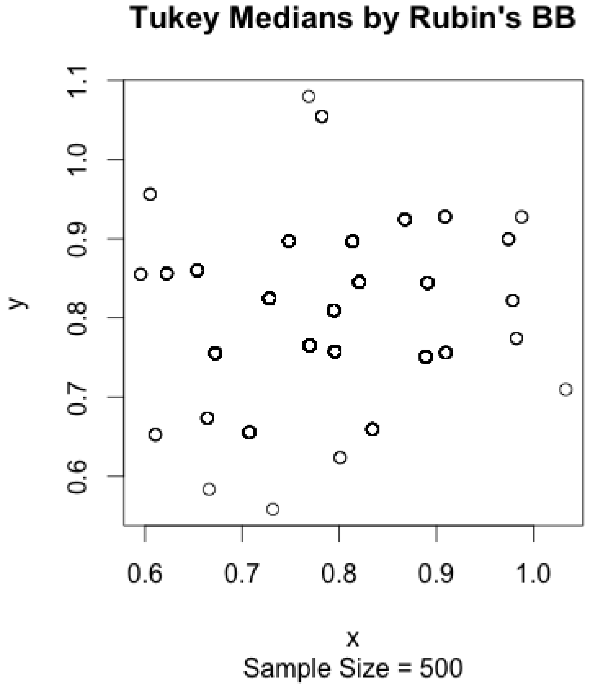
Applying the Rubin's Bayesian bootstrap and the Tukey depth for 1000 times, to the sample sets $\mathcal{E}_{20}$, $\mathcal{E}_{200}$ and $\mathcal{E}_{500}$ respectively, we get a collection of Tukey medians for each sample set. The scattergrams and the contour graphs of the collection of Tukey medians are shown below. The quantile contours represent 25%, 50%, 75% and 100% respectively, from the inside out. From the contour graphs below, we can observe that the points is distributed around the actual Tukey median (0.75, 0.75), which shows the bootstrap sample Tukey median is consistent.

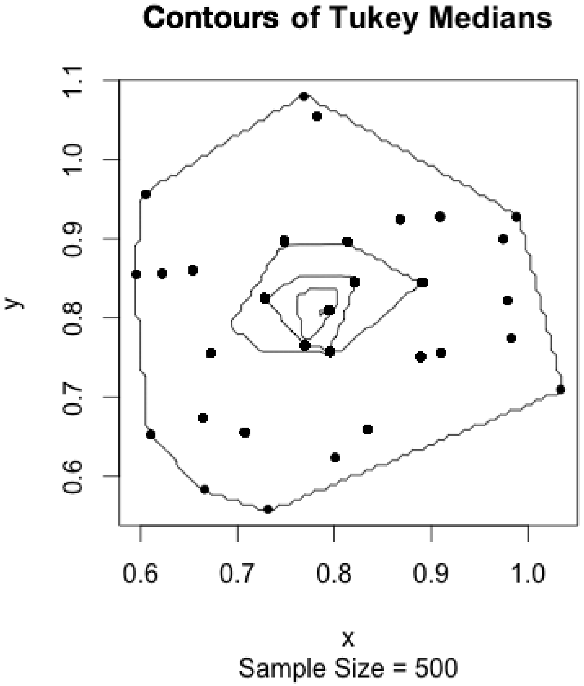





Tukey Medians by Rubin's BB
y
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0.6
0.7
0.8
0.9
1.0
1.1
x
Sample Size = 200


Contours of Tukey Medians
y
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0.6
0.7
0.8
0.9
1.0
1.1
x
Sample Size = 200


Tukey Medians by Rubin's BB
y
0.6
0.7
0.8
0.9
1.0
1.1
0.6
0.7
0.8
0.9
1.0
x
Sample Size = 500


Contours of Tukey Medians
y
0.6
0.7
0.8
0.9
1.0
1.1
0.6
0.7
0.8
0.9
1.0
x
Sample Size = 500

### 4.2.3 Lo's Bayesian Bootstrap

Lo (1987) extends the Rubin's Bayesian bootstrap to a more general technique by placing a Dirichlet process prior to the sample set and draw Bayesian bootstrap samples from the posterior Dirichlet process. To distinguish the Lo's Bayesian bootstrap from the Rubin's Bayesian bootstrap, we call this bootstrap method as Lo's Bayesian Bootstrap. Lo demonstrates that the Bayesian bootstrap estimator for the unknown parameter of interest is valid if the prior probability is a Dirichlet process. However, this result is not surprising as the prior distribution in the Rubin's Bayesian bootstrap follows the flat Dirichlet distribution.

Given an i.i.d. random sample $X_1, \ldots, X_n$ from the completely unknown distribution $F$ and denote the empirical distribution of this sample by $F_n$. If we have a Dirichlet process prior $DP(\alpha, H)$ on $F$, where $\alpha$ is our faith in the prior, the posterior probability is whereby $DP(\alpha^*, H^*)$, where $\alpha^* = \alpha + n$ and $H^* = \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}F_n$. Then we can draw a posterior distribution of $F$ from the posterior Dirichlet process $DP(\alpha^*, H^*)$, denoted by $G_{\alpha n}$, and estimate the unknown parameter, say $\theta$, by this posterior distribution $G_{\alpha n}$. Let $\hat{\theta}_n$ be the estimator for $\theta$ given $F_n$. The detailed steps of the Lo's Bayesian bootstrap are described as follows.

**Step 1.** Simulate a posterior Dirichlet process path from $DP(\alpha^*, H^*)$, where $\alpha^* = \alpha + n$ and $H^* = \frac{\alpha}{\alpha+n}H + \frac{n}{\alpha+n}F_n$, as we do in Chapter 2. This path is a random discrete posterior distribution, denoted as $G_{\alpha n}$.

**Step 2.** Calculate the statistics of interest, $\hat{\theta}_n$, for the posterior distribution $G_{\alpha n}$, denoted by $\hat{\theta}_n^*$.

**Step 3.** Repeat steps 2 and 3 a large number of times, say $B$ times, to generate B posterior distributions and get a collection of the statistics of interest, denoted by $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

**Step 4.** Put a probability of $1/B$ on each point $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ and construct an empirical distribution for $\hat{\theta}_n^*$. Then we can analyze $\theta$ by this empirical distribution.

Lo indicates that the regular bootstrap and the Bayesian bootstrap are asymptotically equivalent, namely, given a random sample, both of them have the same asymptotic distributions for almost all the sample paths when $n$ is large enough. Furthermore, there is no definite answer for which bootstrap method is better.
Next, we define the centred and rescaled posterior Dirichlet process $d_{\alpha n} = \sqrt{n}(G_{\alpha n} - F_n)$. The asymptotic approximation of $d_{\alpha n}$ is given by Lo (1987) in (4.2.3).

**Theorem 4.2.5** (Lo (1987))**.** *Given an i.i.d random sample* $X_1, \ldots, X_n$ *from an unknown continuous distribution function* $F$*, there exists a Brownian bridge* $\{B(s) : 0 \leq s \leq 1\}$ *independent of the sample* $X_1, \ldots, X_n$*, such that*

$$\sup_x |d_{\alpha n}(x) - B(F(x))| = O(n^{-1/4}(\text{lnln}n)^{1/4}(\text{ln}n)^{1/2}) \qquad a.s. \tag{4.2.3}$$

One of the applications of the above theorem is to construct a confidence band of $F$ for a large enough sample size $n$. Refer to Lo (1987) for an example of this application. Furthermore, Lo (1987) also shows the consistency of the Lo's Bayesian bootstrap sample mean. The limiting distribution of Lo's bootstrap sample mean in (4.2.4) is exactly the same as the regular bootstrap sample mean.

**Theorem 4.2.6** (Lo (1987))**.** *Let* $X_1, \ldots, X_n$ *be an i.i.d. random sample from* $F$ *and* $F_n$ *is the empirical distribution. Define* $\mu(F)$*,* $\mu(F_n)$ *and* $\mu(G_{\alpha n})$ *as the population mean, the sample mean and the Lo's Bayesian bootstrap sample mean, respectively. If* $\int x^2 H^*(dx) < \infty$*, then as* $n \to \infty$*,*

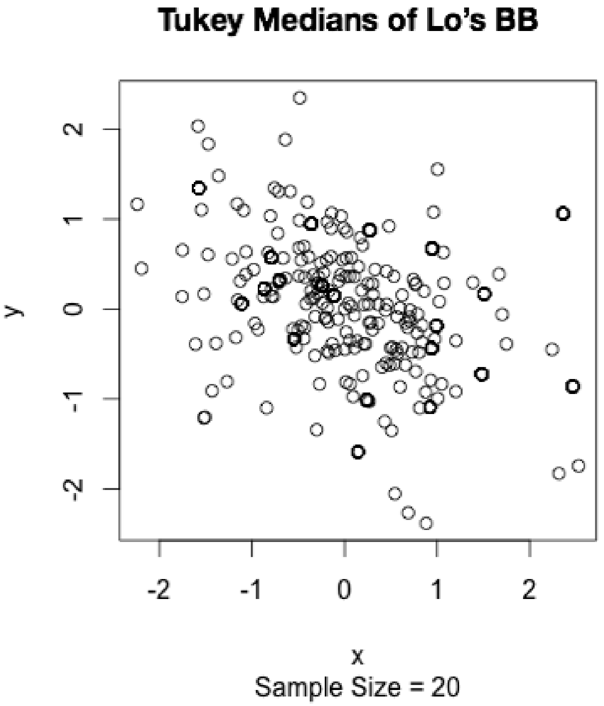
$$\sqrt{n}\left(\mu(G_{\alpha n}) - \mu(F_n)\right) \xrightarrow{d} N\left(0, Var(X)\right). \tag{4.2.4}$$

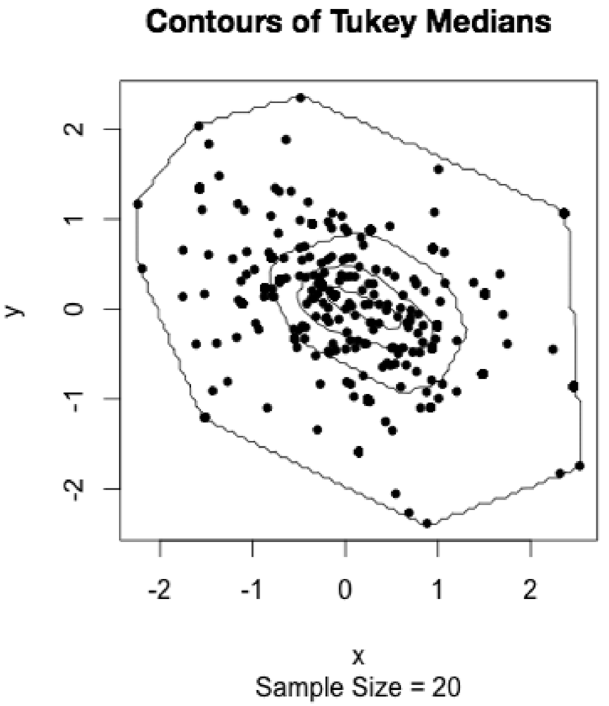
### 4.2.4 A Simulation study of Lo's Bayesian Bootstrap for Tukey Median

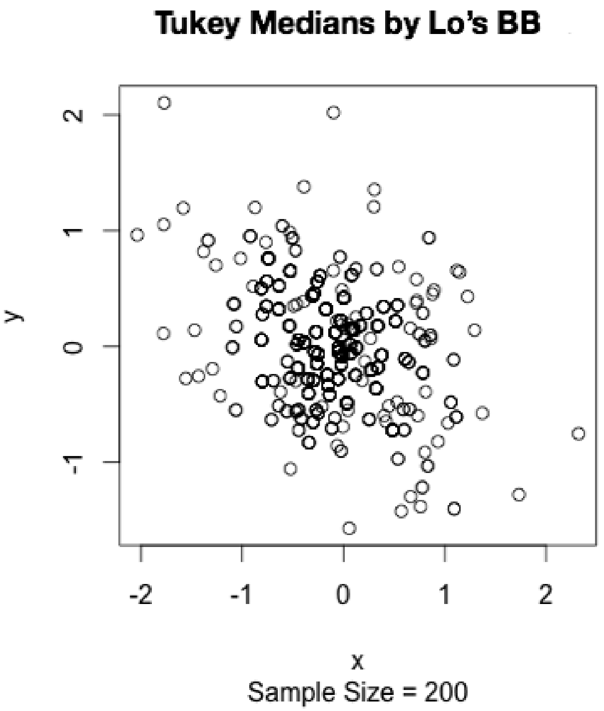
#### The Bivariate Normal Distribution

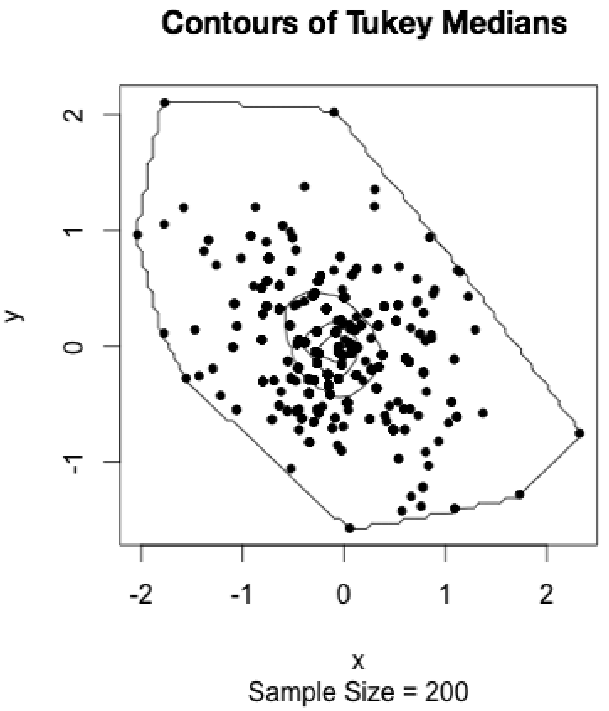
Given a sample set from an unknown distribution function $F$, we put a Dirichlet process prior on $F$ based on some known information and resample from the posterior Dirichlet process. This is the Lo's Bayesian bootstrap. We continue to use the bivariate normal sample sets $\mathcal{N}_{20}$, $\mathcal{N}_{200}$ and $\mathcal{N}_{500}$. Then we give the unknown distribution $F$ a prior $DP(5, H)$ for the sample set $\mathcal{N}_{20}$ and $DP(10, H)$ for the sample sets $\mathcal{N}_{200}$ and $\mathcal{N}_{500}$, where $H \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right)$. In real life, we never know the real underlying distribution of the samples. However, for the simulation purpose, we employ the real distribution as the base measure in the Dirichlet process prior. Then we simulate a posterior distribution from the posterior Dirichlet process for 1000 times and calculate the Tukey median for each posterior distribution. Thus, we get a collection of Tukey medians of size 1000 for each sample set. Draw the scatter plots and contour graphs of the collection of medians and then show them below.

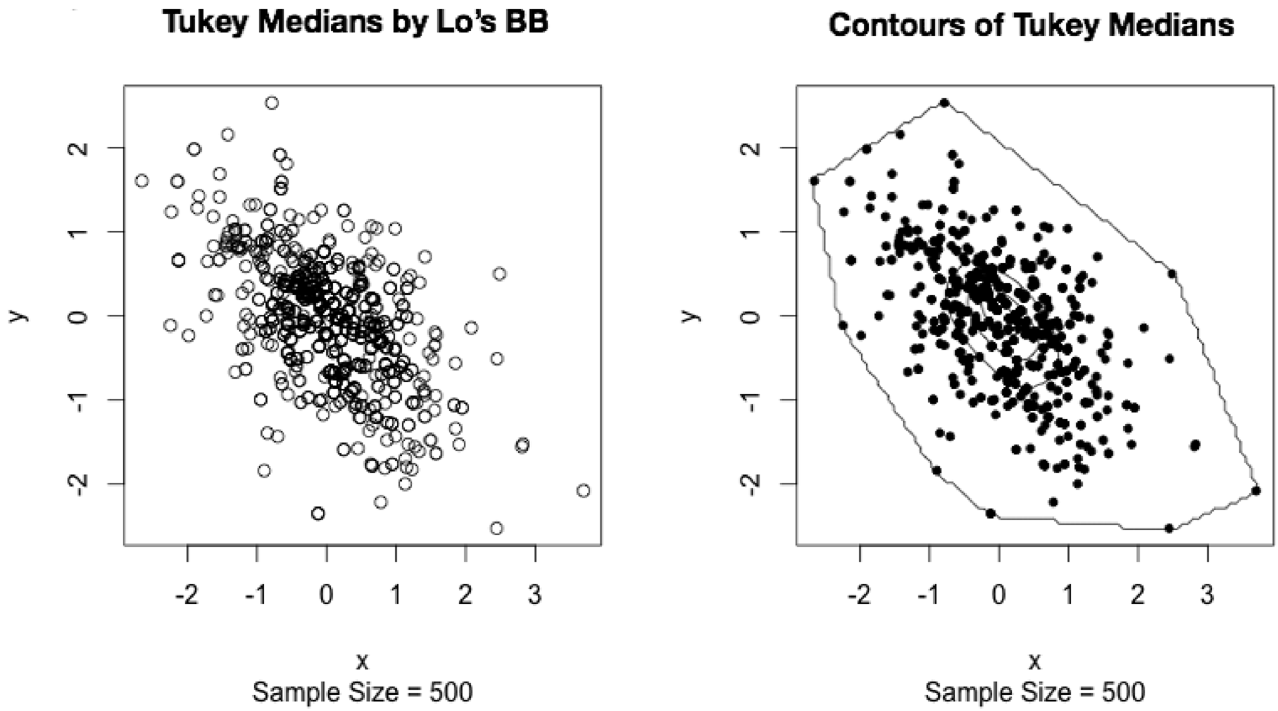
The number of the distinct points on the left-hand graphs increases significantly. This is because a portion of the Lo's bootstrap samples are from the prior distribution rather than the original sample. These new sample points do not exist in the original sample set and whereby increase the diversity of the Lo's bootstrap samples. As a result we get more distinct Tukey medians in the simulations. In each contour graphs, the quantile contours represent 25%, 50%, 75% and 100% from the inside out. The 25% central region falls around the actual Tukey median (the original point) and shrinks with the increase of the sample size. This method presents a much better representation for the Tukey median compared to other bootstrap cases.


Tukey Medians of Lo's BB
y
2
1
0
-1
-2
-2
-1
0
1
2
x
Sample Size = 20


Contours of Tukey Medians
y
2
1
0
-1
-2
-2
-1
0
1
2
x
Sample Size = 20

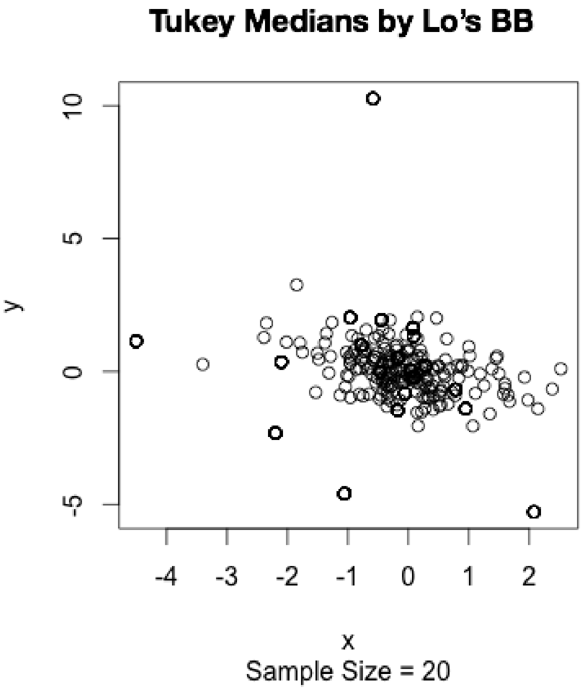

Tukey Medians by Lo's BB
y
2
1
0
-1
-2
-1
0
1
2
x
Sample Size = 200

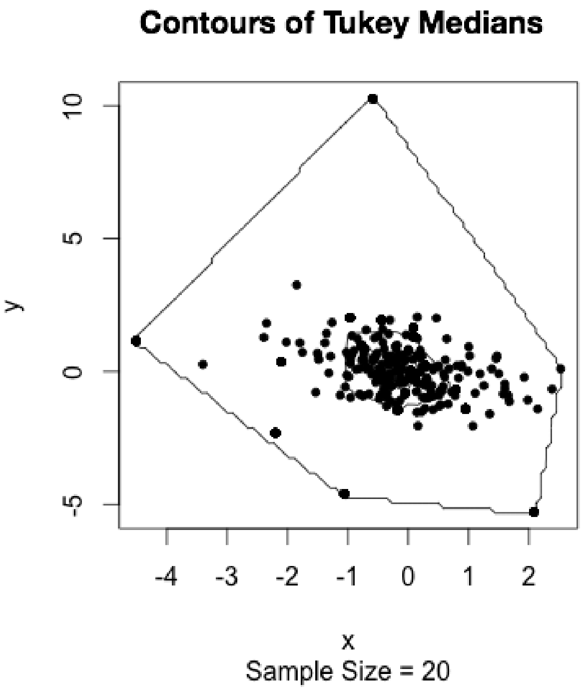

Contours of Tukey Medians
y
2
1
0
-1
-2
-1
0
1
2
x
Sample Size = 200

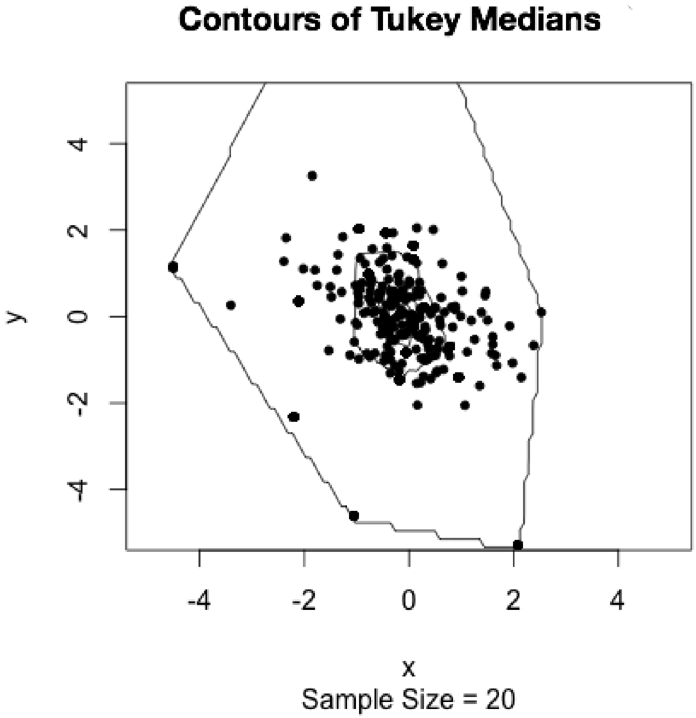


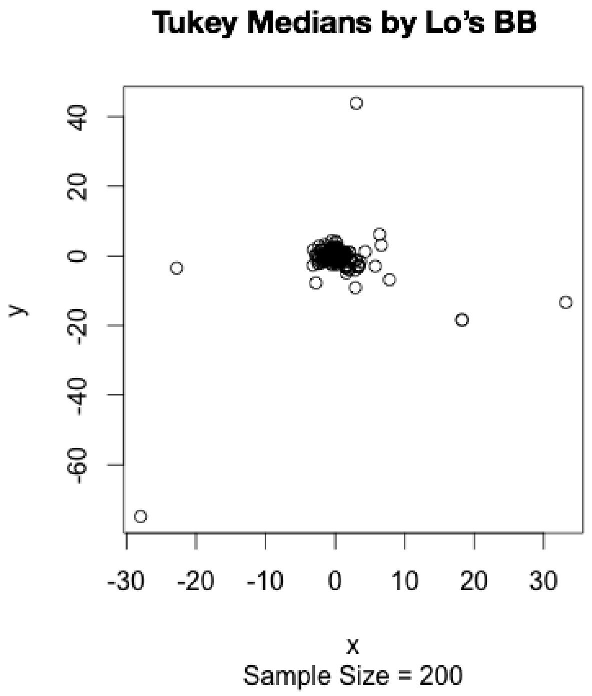
### The Bivariate Cauchy Distribution

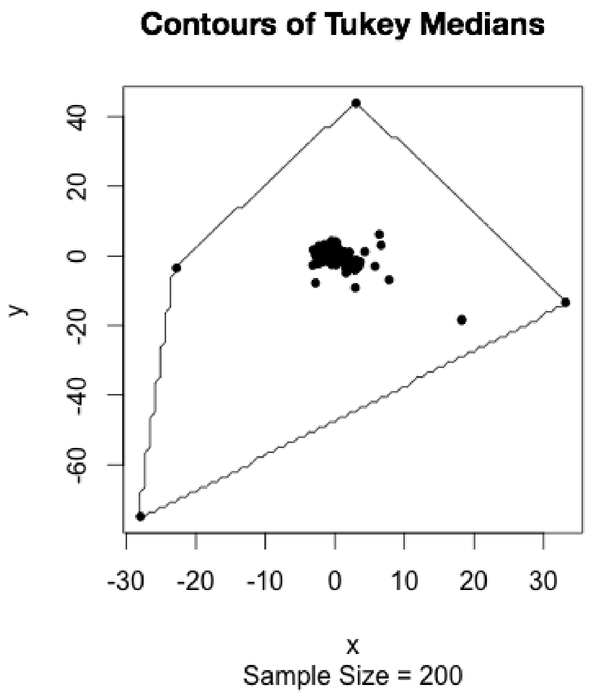
Applying the Lo's Bayesian bootstrap and the Tukey depth for 1000 times, to the sample sets $\mathcal{C}_{20}$, $\mathcal{C}_{200}$ and $\mathcal{C}_{500}$ respectively, we get a collection of Tukey medians for each sample set. Here, the base distribution $H$ of the Dirichlet process prior is the standard bivariate Cauchy distribution. The scatter plots and contour graphs of the collection of Tukey medians are shown below. The quantile contours represent 25%, 50%, 75% and 100% respectively, from the inside out. As we can observe, the range of the set of Tukey medians is pretty large compared to the bivariate normal case. The outermost points can be seen as outliers and we draw new contour graphs by omitting theses points. The points are distributed around the actual Tueky median (the original point) and the analysis is similar to the previous cases.

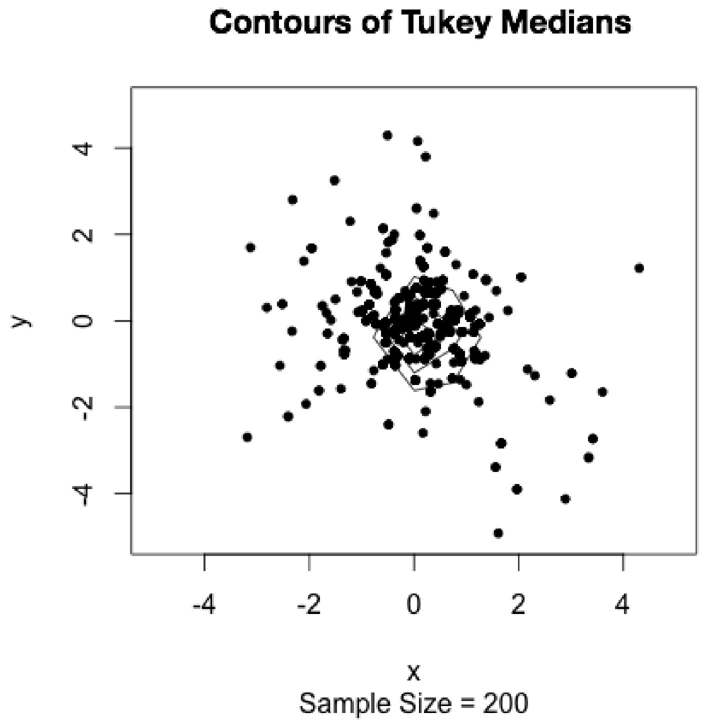

Tukey Medians by Lo's BB

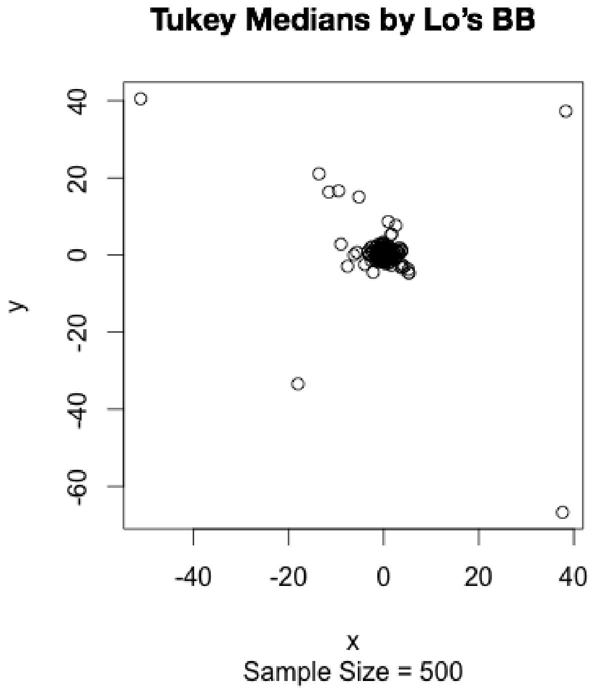

Contours of Tukey Medians


Contours of Tukey Medians


Tukey Medians by Lo's BB

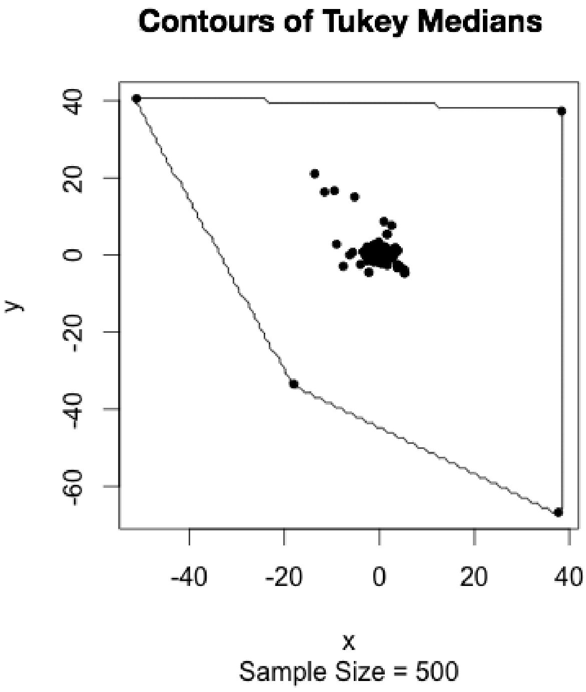

Contours of Tukey Medians


Contours of Tukey Medians



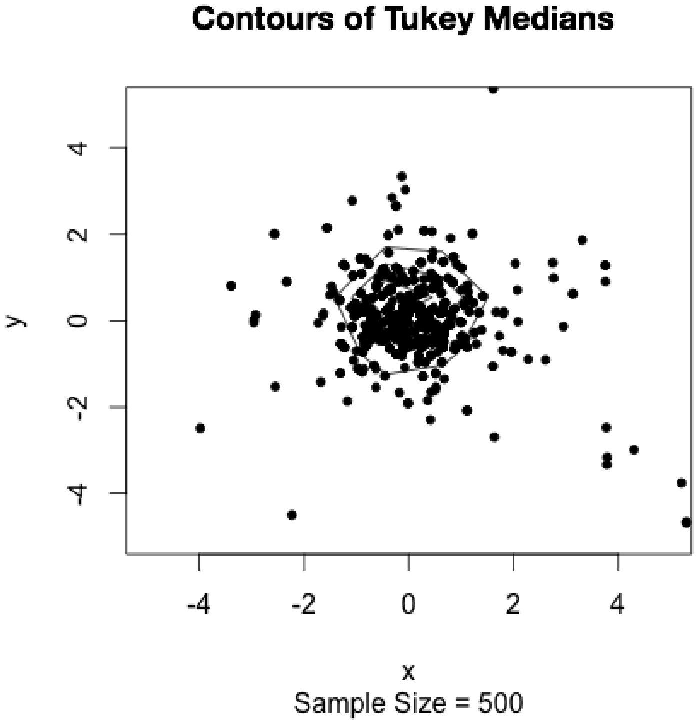
Tukey Medians by Lo's BB

Sample Size = 500



Contours of Tukey Medians

Sample Size = 500

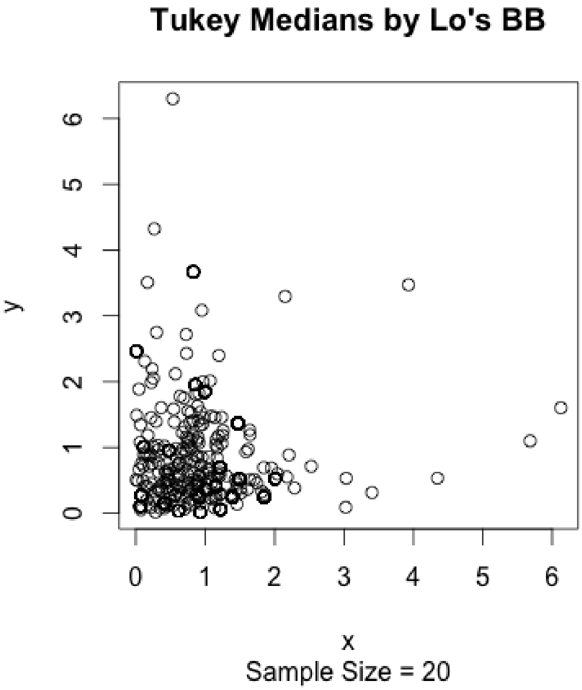


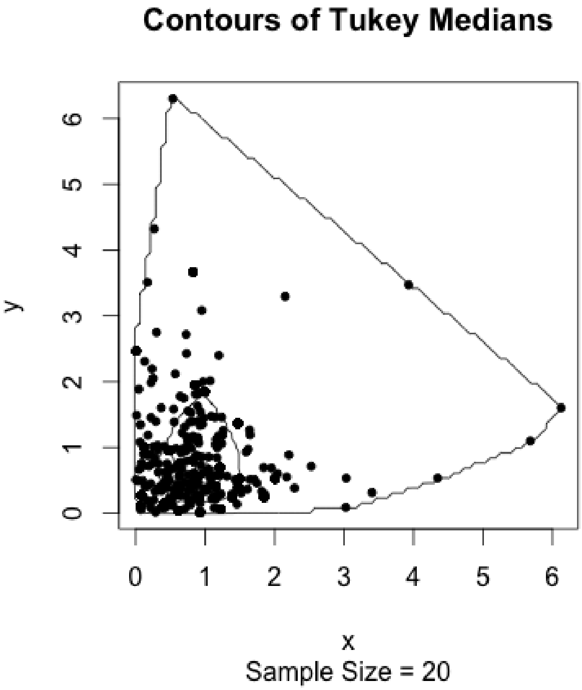
Contours of Tukey Medians

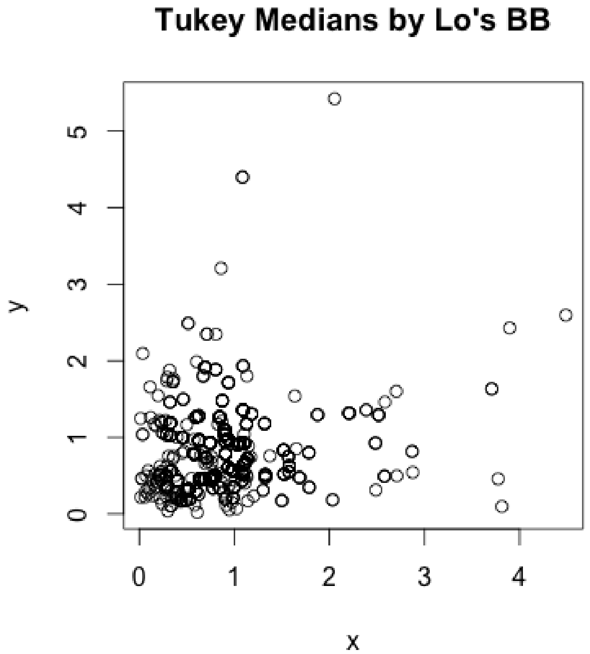
Sample Size = 500

### The Bivariate Exponential Distribution

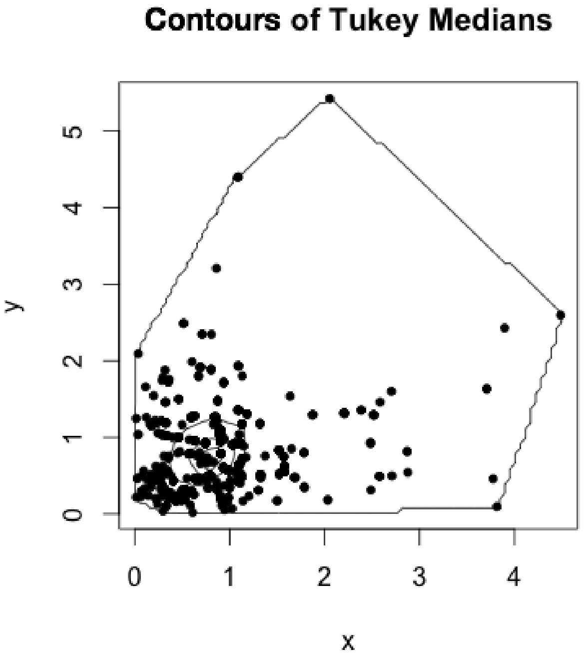
Applying the Lo's Bayesian bootstrap and the Tukey depth for 1000 times, to the sample sets $\mathcal{E}_{20}$, $\mathcal{E}_{200}$ and $\mathcal{E}_{500}$ respectively, we get a collection of Tukey medians for each sample set. The base distribution $H$ of the Dirichlet process prior is the standard bivariate exponential distribution. The scatter plots and contour graphs of the collection of Tukey medians are shown below. The quantile contours represent 25%, 50%, 75% and 100% respectively, from the inside out. From the contour graphs below, we can see that the points is distributed around the actual Tukey median (0.75, 0.75), which shows the bootstrap sample Tukey median is consistent.



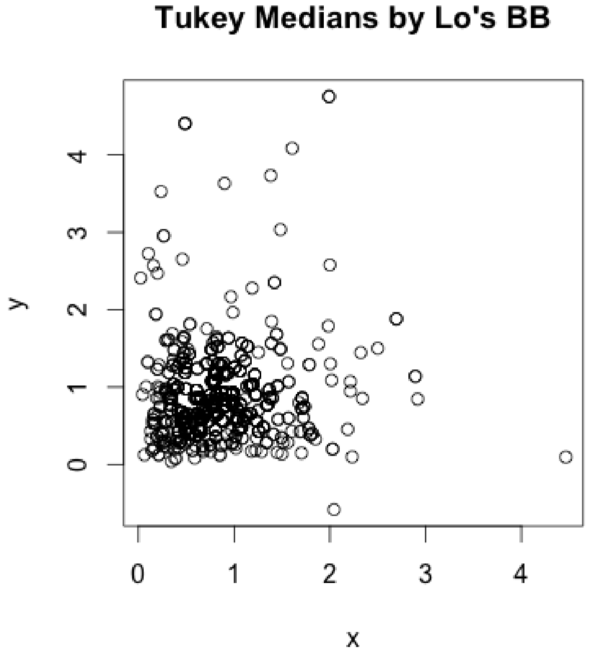



Tukey Medians by Lo's BB

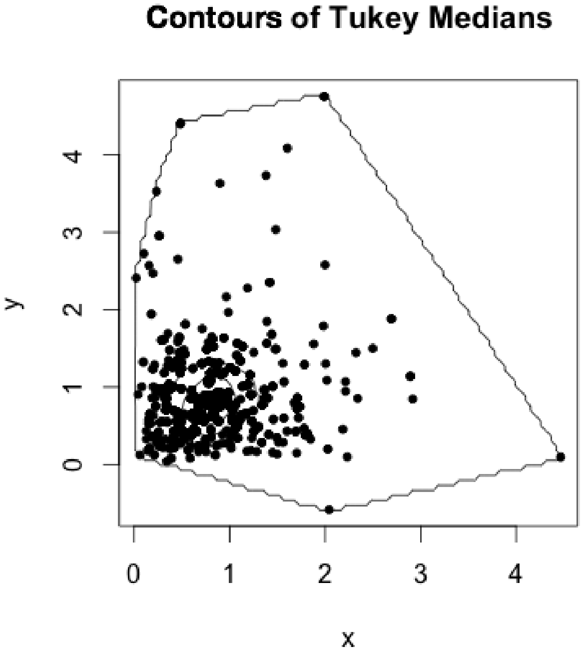
Sample Size = 200


Contours of Tukey Medians

Sample Size = 200


Tukey Medians by Lo's BB

Sample Size = 500


Contours of Tukey Medians

Sample Size = 500

**Remark 4.2.7.** In the bivariate Cauchy case, the range of the set of Tukey medians is pretty large compared to the bivariate normal case. This phenomenon is due to the fact that Cauchy distribution has heavy tails. Therefore, there exists a few extremely large or small points in the bivariate Cauchy sample set, however seldom in the bivariate normal case. In the bootstrap procedure, the larger weights are given to these unrepresentative points so that we get the Tukey medians far away from the actual Tukey median, which is the original point. A small sample of size 10 from both distributions are tabulated in Table 4.1. As can be seen, the point number 1, 2 and 7 are unusual in the Cauchy case.

| | Bivariate Normal | Sample Set | Bivariate Cauchy | Sample Set |
|---|---|---|---|---|
| | x | y | x | y |
| 1 | -0.81824738 | 0.5901294 | -23.61440227 | -8.8578689 |
| 2 | -0.54511944 | 0.4329529 | 687.14010642 | -513.3356351 |
| 3 | 0.01845658 | -0.1606049 | 1.82094584 | 3.4121933 |
| 4 | -0.40482528 | 0.3961275 | -0.12012679 | -0.6836541 |
| 5 | 0.11445618 | -0.2920520 | 0.31569745 | -0.3332673 |
| 6 | -0.29569293 | 0.9147552 | 0.97497865 | -0.3190016 |
| 7 | 0.76295831 | -0.7436176 | -40.59375977 | -10.2633795 |
| 8 | -0.43963404 | 0.2251028 | 0.07295006 | -0.2180382 |
| 9 | -1.17312054 | 1.3973587 | 9.67134273 | 4.8360161 |
| 10 | -0.84282873 | 0.7147185 | 0.51697718 | 0.2155644 |

Table 4.1

**Remark 4.2.8.** Obviously, in the Lo's Bayesian bootstrap, there is only a small proportion in the resamples that produced near identical medians. On the other hand, the Efron's regular bootstrap and the Rubin's Bayesian bootstrap are mainly applied on the original sample data. If the original sample is not a good representative of the real population, the result will be inaccurate. With the credible prior information, the Lo's Bayesian bootstrap method will supplement such weakness to a certain extent. Moreover we should note that the median is robust to the outliers. However we believe that asymptotically all three methods are valid.

## 4.3 Muliere and Secchi's Bootstrap

Muliere and Secchi (1996) give a generalization of the Bayesian bootstrap associated with a Dirichlet process prior $DP(\alpha, H)$, including the Efron's regular bootstrap and Rubin's Bayesian bootstrap. They indicate that the Efron's regular bootstrap and Rubin's bootstrap can be viewed as the special cases of the general Bayesian bootstrap by setting $\alpha = 0$. The proof is given by the Theorem 2.1 of Muliere and Secchi (1996). Then, a new Bayesian bootstrap technique is suggested, which is similar to the Lo's Bayesian bootstrap but simpler to operate.

The next lemma shows a trivial case that how to simulate a Dirichlet process when the base distribution $H$ is a discrete distribution function with finite support.

**Lemma 4.3.1** (Muliere and Secchi (1996)). *Let $H$ be a discrete distribution function with finite support $h_1, \ldots, h_K$ on $\mathbb{R}$ and $p_i$ be the probability of each point $h_i$ for $i = 1, \ldots, K$. Let $V_1, \ldots, V_K$ be i.i.d. random variables from the Gamma distribution $V_i \sim Gamma(\alpha p_i, 1)$ for $i = 1, \ldots, K$, where $\alpha > 0$. Define a distribution function $G(x)$:*

$$G(x) = \frac{1}{\sum_{i=1}^{K} V_i} \sum_{i=1}^{K} V_i I_{[h_i, \infty)}(x) \tag{4.3.1}$$

*Then, $G(x)$ is a Dirichlet process $DP(\alpha, H)$.*

Let $X_1, \ldots, X_n$ be an i.i.d. random sample from an unknown distribution $F$ and a Dirichlet process prior $DP(\alpha, H)$ is given on $F$. Therefore, the posterior probability is derived, which is $DP(\alpha^*, H^*)$ where $\alpha^* = \alpha + n$ and $H^* = \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} F_n$. Given a random sample set $X_1, \ldots, X_n$, the detailed procedures of the new bootstrap technique are presented as followed, based on (4.3.1).

**Step 1.** Draw an i.i.d. random sample of size m $X_1^*, \ldots, X_m^*$ from $H^* = \frac{\alpha}{\alpha+n} H + \frac{n}{\alpha+n} F_n$.

**Step 2.** Simulate $m$ random numbers from the Gamma distribution $Gamma(\frac{n+\alpha}{m}, 1)$, denoted by $V_1, \ldots, V_m$. Then assign each sample point $X_i^*$ a weight $w_i = V_i / \sum_{i=1}^m V_i$. A random discrete posterior distribution $G_{\alpha m}$ is constructed.

**Step 3.** Calculate the statistics of interest, $\hat{\theta}_n$, for the posterior distribution $G_{\alpha m}$, denoted by $\hat{\theta}_n^*$.

**Step 4.** Repeat steps 2 and 3 a large number of times, say $B$ times, to generate B posterior distributions and get a collection of the statistics of interest, denoted by $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.

**Step 5.** Put a probability of $1/B$ on each point $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ and construct an empirical distribution for $\hat{\theta}_n^*$. Then we can analyze $\theta$ by this empirical distribution.

Furthermore, Muliere and Secchi (1996) demonstrate the bootstrap quantile process and the bootstrap sample mean derived by this new bootstrap method are also consistent in $\mathbb{R}$. Refer to the Theorem 3.1 and Theorem 3.2 of Muliere and Secchi (1996) for more details.

# Chapter 5

# Asymptotics for the Tukey Median

The asymptotic distribution of the Tukey median has been fully developed in both univariate and multivariate cases under the symmetry or asymmetry hypothesis. Of course, in the univariate case, the quantile process has been studied thoroughly, and the median is only a special case. Nolan (1999) obtained the asymptotic properties of the Tukey median for the bivariate case under the angularly symmetry assumption. Bai and He (1999) developed the asymptotic results for the multivariate case but also under hypothesis of symmetry. Massé (2001) extended the above results into a more general situation without the symmetry assumption. In this chapter, we review the asymptotic properties of the Tukey Median mainly based on the results from Nolan (1999) and Massé (2001) and conjecture that the bootstrap Tukey median converges to the same limiting distribution. As we noted in the simulation section (Chapter 4) that the Bayesian bootstrap had a better performance due to the ties in the regular bootstrap.

## 5.1 Introduction

To be formal, define all the directions on $\mathbb{R}^d$ by the space $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : |u| = 1\}$ and all the closed half-spaces containing $x$ by $H[x, u] = \{t \in \mathbb{R}^d : u't \geq u'x, x \in \mathbb{R}^d, u \in \mathbb{S}^{d-1}\}$. Let $F$ be a distribution function on $\mathbb{R}^d$ and $D(x)$ be the *Tukey depth function* with respect to

$F$. Then,

$$D(x) = \inf_{u \in \mathbb{S}^{d-1}} F(H[x,u]).$$

Given a sample set $X_1, \ldots, X_n$ from $F$, let $F_n$ be the empirical distribution of this sample. Then, the *sample Tukey depth function* $D_n(x)$ of a point $x \in \mathbb{R}^d$ is defined as the Tukey depth function by replacing $F$ by $F_n$. Namely,

$$D_n(x) = \inf_{u \in \mathbb{S}^{d-1}} F_n(H[x,u]).$$

Therefore, the sample Tukey depth of a point $x \in \mathbb{R}^d$ is the smallest proportion of the sample set in a closed half-space containing $x$. Refer to Figure 3.1 (Chapter 3) for a graphic representation of the Tukey depth.

In the following, we assume $F$ has a unique Tukey median. This is followed that $F$ has a connected support and is absolutely continuous. Let $F_u$ be the one dimensional orthogonal projection of $F$ for each $u \in \mathbb{R}^d$, i.e. the one dimensional marginal distribution of $u'X$ where $X \sim F$. Thus $F_u$ is absolutely continuous for all $u \in \mathbb{R}^d$. Since the Tukey depth is affine invariant, as we mentioned previously, the Tukey median can be defined to be equal to 0, without loss of generality. The sample Tukey median denoted by $T_n$ is defined by

$$D_n(T_n) = \max_{x \in \mathbb{R}^{d-1}} D_n(x) \quad \textit{or} \quad T_n = \arg\max_{x \in \mathbb{R}^d} \inf_{u \in \mathbb{S}^{d-1}} F_n(H[u,x]).$$

Under a strong angularly symmetry assumption, Nolan (1999) obtains the asymptotic distribution of the sample Tukey median in the univariate and bivariate cases. She conjectures the result holds in higher dimensions.

**Definition 5.1.1** (Liu (1990))**.** *A random vector $X$ on $\mathbb{R}^d$ or its distribution $F$ is* angularly symmetric *about the point $\theta \in \mathbb{R}^d$ if and only if*

$$\frac{X-\theta}{||X-\theta||} \overset{\mathrm{d}}{=} -\frac{X-\theta}{||X-\theta||}$$

*where $||\cdot||$ is the Euclidean norm.*

**Theorem 5.1.2** (Nolan (1999)). *For $d = 1, 2$, suppose that:*

*a. $F_u$ has a unique median at 0 for all $u \in \mathbb{R}^d$;*

*b. $F_u$ has a bounded positive density $f_u$ at 0 and $f_u$ is continuous at 0 for any $u \in \mathbb{R}^d$.*

*Then,*

$$\sqrt{n}T_n \xrightarrow{d} \arg\max_{t} \inf_{u \in \mathbb{R}^d} \{Z(u) - u'tf_u(0)\}$$

*where $Z$ is a Gaussian process on the unit sphere, with zero mean and covariance function*

$$Cov[Z(u), Z(v)] = F(H[0, u] \cap F(H[0, v])) - \frac{1}{4} \qquad u, v \in \mathbb{R}^d.$$

In the univariate case, the sample Tukey median is actually the sample median. let $d = 1$, then from the above convergence, we have

$$\sqrt{n}T_n \xrightarrow{d} N\left(0, \frac{1}{4f^2(0)}\right).$$

This result is the same as the asymptotic distribution of the sample median. Of course, for $d = 1$, this result is derived without any assumption of symmetry for the quantile process. Massé (2001) indicates that the condition (a) is pretty restrictive. It implies that any half spaces with a boundary going through the point 0 have the probability of 1/2. This holds if $F$ has a positive density at 0 and $F$ is angular symmetric about 0. Massé extends Nolan's results to a more general case where the distribution $F$ has little or no symmetry at all and the dimensions are higher.

The next proposition shows that all points on $\mathbb{R}^d$ have the minimal half spaces and the minimal directions. This is under the assumption we stated previously, that $F$ is absolutely continuous.

**Proposition 5.1.3.** *Suppose that $F$ gives probability zero to each hyperplane. Then for each $x \in \mathbb{R}^d$,*

**a.** *there exists at least one minimal half space $H[x, u]$ for $x$;*

**b.** *there exists at least one minimal direction* $u \in \mathbb{R}^d$ *at* $x$.

The following two theorems give us some properties of the sample Tukey median.

**Theorem 5.1.4** (Massé (2001))**.** $T_n \to 0$ *almost surely.*

**Proof:** For any points $x \in \mathbb{R}^d$,

$$\begin{aligned}\sup_x |D_n(x) - D(x)| &= \sup_x |\inf_{u \in \mathbb{S}^{d-1}} F_n(H[x,u]) - \inf_{u \in \mathbb{S}^{d-1}} F(H[x,u])| \\ &\leq \sup_x |F_n(H[x,u]) - F(H[x,u])| \to 0 \qquad a.s. \quad as \quad n \to \infty\end{aligned}$$

Therefore, $\sup_x D_n(x) - \sup_x D(x) \to 0$ a.s., i.e. $T_n \to 0$ a.s. as $n \to \infty$. ∎

Next, denote $V(0)$ as all the minimal directions at the Tukey median 0. In Nolan's case, $V(0) = \mathbb{S}^{d-1}$. Besides, recall that a sequence of random variables $\{X_n\}$ is *uniformly tight* if for any $\epsilon > 0$, there exists a positive constant $c$ such that $P(|X_n| \geq c) < \epsilon$ for each n. The next theorem is an extension of that from Nolan (1999).

**Theorem 5.1.5** (Massé (2001))**.** *Suppose that:*

**a.** *there exists* $c > 0$ *such that*

$$\min_{u \in \mathbb{S}^{d-1}} \max_{v \in V(0)} v'u \geq c;$$

**b.** *for each* $v \in V(0)$, $F_v$ *has a bounded positive density* $f_v$ *at 0;*

**c.** *for any sequences* $\{\delta_n\}$ *decreasing to 0,*

$$\sup_{v \in V(0)} \sup_{|s| \leq \delta_n} |f_v(s) - f_v(0)| \to 0 \qquad as \quad n \to \infty.$$

*Then,* $\sqrt{n}T_n$ *is uniformly tight.*

The asymptotic distribution of the sample Tukey median on $\mathbb{R}^d$ is an application of the *continuous mapping theorem.* The detailed proof can be referred to Nolan (1999) and Massé (2001). The next theorem from Massé (2001) shows the asymptotic distribution of the sample Tukey median.

**Theorem 5.1.6** (Massé (2001))**.** *Suppose that:*

**a.** *there exists $c > 0$ such that*

$$\min_{u \in \mathbb{S}^{d-1}} \max_{v \in V(0)} v'u \geq c;$$

**b.** *for each $v \in V(0)$, $F_v$ has a bounded positive density $f_v$ at 0;*

**c.** *for any $u \in \mathbb{S}^{d-1}$, $f_u$ is continuous at 0.*

*Then,*

$$\sqrt{n}T_n \xrightarrow{d} \arg\max_t \inf_{v \in V(0)} \{B(F(H[0,v])) - v'tf_v(0)\}$$

*where $B(\cdot)$ is a Brownian bridge.*

## 5.2 Discussion and Future Work

Given an i.i.d. random sample $X_1, \ldots, X_n$ from an unknown distribution $F$, let $T_n^*$ be the bootstrap sample Tukey median, derived by both regular bootstrap and Bayesian bootstrap. We would like to obtain the consistency of $T_n^*$ by showing

$$\sqrt{n}(T_n^* - T_n) \xrightarrow{d} \arg\max_t \inf_{v \in V(0)} \{B(F(H[0,v])) - v'tf_v(0)\}.$$

In the bootstrap procedures, the empirical distribution $F_n$ of this sample is viewed as the population distribution function. $F_n$ is a discrete function, however, $F$ is required to be absolutely continuous in the proofs of Nolan (1999) or Massé (2001). This is the main obstacle of proving the consistency of the bootstrap sample Tukey median if using the same techniques as Nolan (1999) or Massé (2001). However, there perhaps exists some other techniques to show the consistency which is beyond the scope of this thesis and should be considered in the future research.

# Appendix A

# R Programming

1. Gamma Process (1982)

```

GP_Bond <- function(e){
  a = 10
  w = 0
  n = 1
  uw = 0
  T = E = 0
  SumW = SumW1 = 0
  p = TRUE

  while(p){
    E[n] = rexp(1)
    T[n] = sum(E)
    w[n] = exp(-T[n]/a)*E[n]
    if(w[n] >= e){
      n = n+1
    }
    else p = FALSE
  }

  theta = sort(rnorm(n))

  return(list(n = n, theta = theta, weight = w))
}

> samplepath = GP_Bond(0.0001)
```

```
> plot(stepfun(samplepath$theta,samplepath$weight), do.points = F, col = "blue",
   mgp = c(2,1,0), main = "Gamma␣Process␣GP(10,N(0,1))␣by␣Bondesson(1982)",
  xlab = expression(theta[(i)]), ylab = expression(G[n]^"Bond"))
> for(i in 1:samplepath$n){segments(samplepath$theta[i],0,samplepath$theta[i],
  samplepath$weight[i+1]-samplepath$weight[i])}
> samplepath$n

> samplepath = GP_Bond(0.0001)
> lines(stepfun(samplepath$theta,samplepath$weight), do.points = F, col = "blue"
  )
> samplepath$n
```

## 2. Dirichlet Process (1982)

```
DP_Bond <- function(e){
  a = 10
  w = 0
  n = 1
  uw = 0
  T = E = 0
  SumW = SumW1 = 0
  p = TRUE

  while(p){
    E[n] = rexp(1)
    T[n] = sum(E)
    SumW = SumW + exp(-T[n]/a)*E[n]
    uw[n] = exp(-T[n]/a)*E[n]
    if(uw[n]/SumW >= e){
      n = n + 1
    }
    else p = FALSE
  }

  for( i in 1:n){
    w[i] = uw[i]/SumW
  }

  theta = sort(rnorm(n))
  weight = c(0, cumsum(w))

  return( list(n = n, theta = theta, weight = weight))
}
```

```

> samplepath = DP_Bond(0.0001)
> plot(stepfun(samplepath$theta,samplepath$weight), do.points = F, col = "blue",
   mgp = c(2,1,0), main = "Dirichlet Process DP(10,N(0,1)) by Bondesson(1982)"
  , xlab = expression(theta[(i)]), ylab = expression(P[n]^"Bond"))
> samplepath$n
> for(i in 1:samplepath$n){segments(samplepath$theta[i],0,samplepath$theta[i],
  samplepath$weight[i+1]-samplepath$weight[i])}
```

3. Dirichlet Process (1994)

```

DP_Seth <- function(e){
  a = 10
  n = 1
  pi = 0
  sumpi = 0
  p = TRUE

  while(p){
    pi[n] = rbeta(1,1,a)*(1-sumpi)
    sumpi = sumpi + pi[n]
    if(pi[n]>=e){
      n = n + 1
    }
    else p = FALSE
  }

  theta = sort(rnorm(n))
  weight = c(0, cumsum(pi))

  return(list(n = n, theta = theta, weight = weight))
}
```

4. Gamma Process (Zarepour and Al Labadi (2012))

```
# Compute the function G_n^{-1}(x)
Inv_G <- function(a, n, x){
  q = qgamma (1-x, a/n, 1)
  return(q)
}

```

```r
# Get a Sample of Gamma Process given ε
GP_Sample <- function(e){
  a = 10
  w = 0
  n = 1
  E = rexp(1)
  T = E
  p = TRUE
  while(p){
    E[n+1] = rexp(1)
    T[n+1] = sum(E)
    if (Inv_G (a, n, T[n]/T[n+1]) >= e){
      n = n + 1
    }
    else p = FALSE
  }

  theta = rnorm (n)
  for(i in 1:n){
    w[i] = Inv_G (a, n, T[i]/T[n+1])
  }
  theta_w = cbind (theta, w)
  theta_w_order = theta_w [order(theta_w[,1]), ]
  x = theta_w_order [ ,1]
  y = c (0, cumsum(theta_w_order[ ,2]))

  return (list(n = n, ordered_theta = x, cumulative_weight = y))
  }

 # Draw sample paths and get the values of n's
> samplepath = GP_Sample(0.0001)
> plot (stepfun(samplepath$ordered_theta, samplepath$cumulative_weight), do.points = F, col = "blue", mgp = c(2,1,0), main = "Gamma Process GP(10,N(0,1)) by Zar&Al-Lab", xlab = expression(theta[(i)]), ylab = expression(G[n]^"Zar&Al-Lab"))
> samplepath$n

> samplepath = GP_Sample(0.0001)
> lines(stepfun(samplepath$ordered_theta,samplepath$cumulative_weight ), do.points = F, col = "blue")
> samplepath$n

> for(i in 1:samplepath$n){segments(samplepath$ordered_theta[i],0,samplepath$
```

```
    ordered_theta[i],samplepath$cumulative_weight[i+1]-samplepath$cumulative_
    weight[i])}
```

5. Dirichlet Process (Zarepour and Al Labadi (2012))

```
# Get a Sample of Dirichlet Process given ε
DP_Sample <- function(e){
  a = 10
  w = 0
  n = 1
  E = rexp(1)
  T = E
  SumIG = SumIG1 = 0
  p = TRUE
  while(p){
    E[n+1] = rexp(1)
    T[n+1] = sum(E)
    SumIG = SumIG + Inv_G (a, n, T[n]/T[n+1])
    if (Inv_G (a, n, T[n]/T[n+1])/SumIG >= e){
      n = n + 1
    }
    else p = FALSE
  }

  theta = rnorm (n)
  for(i in 1:n){
    SumIG1 = SumIG1 + Inv_G (a, n, T[i]/T[n+1])
  }

  for(i in 1:n){
    w[i] = Inv_G (a, n, T[i]/T[n+1])/SumIG1
  }
  theta_w = cbind (theta, w)
  theta_w_order = theta_w [order(theta_w[,1]), ]
  x = theta_w_order [ ,1]
  y = c (0, cumsum(theta_w_order[ ,2]))

  return (list(n = n, ordered_theta = x, cumulative_weight = y))
}

# Draw sample paths and get the values of n's
> samplepath=DP_Sample(0.0001)
> plot (stepfun(samplepath$ordered_theta, samplepath$cumulative_weight), do.
```

```
points = F, col = "blue", mgp = c(2,1,0), main = "Dirichlet Process GP(10,N(0,1)) by Zar&Al-Lab", xlab = expression(theta[(i)]), ylab = expression(G[n]^"Zar&Al-Lab"))
> samplepath$n

> samplepath=DP_Sample(0.0001)
> lines(stepfun(samplepath$ordered_theta,samplepath$cumulative_weight ), do.points = F, col = "blue")
> samplepath$n

> for(i in 1:samplepath$n){segments(samplepath$ordered_theta[i],0,samplepath$ordered_theta[i],samplepath$cumulative_weight[i+1]-samplepath$cumulative_weight[i])}
```

6. The posterior Dirichlet Process (Zarepour and Al Labadi (2012))

```
# Dirichlet process prior given n
DP_Prior <- function(n){
  a = 10
  w = 0
  E = rexp(n+1)
  T = cumsum(E)
  SumIG1 = 0

  theta = rnorm (n)
  for(i in 1:n){
    SumIG1 = SumIG1 + Inv_G (a, n, T[i]/T[n+1])
  }

  for(i in 1: n){
    w[i] = Inv_G (a, n, T[i]/T[n+1])/SumIG1
  }
  theta_w = cbind (theta, w)
  theta_w_order = theta_w [order(theta_w[,1]), ]
  x = theta_w_order [ ,1]
  y = c (0, cumsum(theta_w_order[ ,2]))

  return (list(ordered_theta = x, cumulative_weight = y))
}

# Posterior Dirichlet process given n and the size m of data X
DP_Post <- function(n, m){
  a = 10
```

```
  w = 0
  E = rexp(n+1)
  T = cumsum(E)
  SumIG1 = 0
  X = rexp(m, 0.5)
  alpha = a/(a+m)
  theta = 0

  for(i in 1:n){
    u = runif(1)
    if (u < alpha) theta[i] = rnorm(1)
    else theta[i] = sample(X, 1)
  }

  for(i in 1:n){
    SumIG1 = SumIG1 + Inv_G (a+m, n, T[i]/T[n+1])
  }

  for(i in 1: n){
    w[i] = Inv_G (a+m, n, T[i]/T[n+1])/SumIG1
  }
  theta_w = cbind (theta, w)
  theta_w_order = theta_w [order(theta_w[,1]), ]
  x = theta_w_order [ ,1]
  y = c (0, cumsum(theta_w_order[ ,2]))

  return (list(ordered_theta = x, cumulative_weight = y))
}

# Draw paths and make a comparison
> curve(pexp(x,0.5),-5,10,mgp = c(2,1,0), main = "DP(10,N(0,1))␣by␣Zar&Al-Lab␣with␣n=1000␣and␣m=10", xlab = expression(theta[(i)]), ylab=expression(P^"Zar&Al-Lab"))

> curve(pexp(x,0.5),-5,10,mgp = c(2,1,0), main = "DP(10,N(0,1))␣by␣Zar&Al-Lab␣with␣n=1000␣and␣m=50", xlab = expression(theta[(i)]), ylab=expression(P^"Zar&Al-Lab"))

> curve(pexp(x,0.5),-5,10,mgp = c(2,1,0), main = "DP(10,N(0,1))␣by␣Zar&Al-Lab␣with␣n=1000␣and␣m=200", xlab = expression(theta[(i)]), ylab=expression(P^"Zar&Al-Lab"))

> samplepath = DP_Prior(1000)
```

```
> lines (stepfun(samplepath$ordered_theta, samplepath$cumulative_weight), do.points = F, col = ”red”)

> samplepath = DP_Post(1000,200)
> lines (stepfun(samplepath$ordered_theta, samplepath$cumulative_weight), do.points = F, col = ”blue”)
```

## 7. Data Depth

```
##Package
#Ddalpha (for Contour)
#Aplpack (for BagPlot)

# Get a bivariate normal sample
dataset = mvrnorm(n = 500, mu = c(2,3), Sigma = matrix(c(5,1,1,2),2,2))

#CONTOUR

## Tukey depth
#Other depth functions are operated in a similar way.
hd = depth.halfspace(dataset,dataset)
hd_order = sort(hd,decreasing = T)
# Compute the quantiles of 0.2, 0.4, 0.6, 0.8 and the Tukey Median
hd0.8 = hd_order[400]
hd0.6 = hd_order[300]
hd0.4 = hd_order[200]
hd0.2 = hd_order[100]
hd_median = hd_order[1]
# Draw a Contour graph
depth.contours(dataset,depth=”halfspace”,main=”Normal␣contours␣by␣HD”,levels = c(hd0.8,hd0.6,hd0.4,hd0.2,hd_median),frequency = 100,col=”black”)

#BAGPLOT:
 bagplot(dataset, factor = 3, na.rm = FALSE, approx.limit = 300,
+          show.outlier = TRUE, show.whiskers = FALSE,
+          show.looppoints = TRUE, show.bagpoints = TRUE,
+          show.loophull = TRUE, show.baghull = TRUE,
+          create.plot = TRUE, add = FALSE, pch = 16, cex = 0.4,
+          dkmethod = 2, precision = 1, verbose = FALSE,
+          debug.plots = ”no”,   col.loophull=”#aaccff”,
+          col.looppoints=”#3355ff”, col.baghull=”#7799ff”,
+          col.bagpoints=”#000088”, transparency=FALSE, xlab=”␣”, ylab=”␣”)

```

```
airquality
airqual = cbind(airquality$Wind, airquality$Temp)
bagplot(airqual, factor = 3, na.rm = FALSE, approx.limit = 300,
+          show.outlier = TRUE, show.whiskers = FALSE,
+          show.looppoints = TRUE, show.bagpoints = TRUE,
+          show.loophull = TRUE, show.baghull = TRUE,
+          create.plot = TRUE, add = FALSE, pch = 16, cex = 0.4,
+          dkmethod = 2, precision = 1, verbose = FALSE,
+          debug.plots = "no",   col.loophull="#aaccff",
+          col.looppoints="#3355ff", col.baghull="#7799ff",
+          col.bagpoints="#000088", transparency=FALSE,xlab="Wind/mph␣",ylab="Temperature/degrees␣F␣",xlim=c(0,20), ylim= c(50,100),main="Daily␣air␣quality␣measurements␣in␣New␣York")
```

8. Bootstrap for the sample of size 20

```

# Draw a sample of size 20 from normal distribution
# N(c(0,0), matrix(c(1,-.5,-.5,1),2,2))
X20 = mvrnorm(20, c(0,0), matrix(c(1,-.5,-.5,1),2,2))
#X20 = rmvc(20,c(0,0),matrix(c(1,0,0,1),2,2))  ## Standard Cauchy distribution
#X20 = cbind(rexp(20),rexp(20))   ##standard bivariate exponential distribution

# Get a set of medians of size 1000 by regular bootstrap
mb2 = matrix(0,1000,2)
dat2db = matrix(0,20,2)
for(i in 1:1000){
  sploc = sample(1:20, 20, replace = TRUE)
  dat2db[1:20,] = X20[sploc,]
  dat2dbdepth = depth.halfspace(dat2db, dat2db)
  mloc = match(max(dat2dbdepth), dat2dbdepth)
  mb2[i,] = dat2db[mloc,]
  #mloc = which(dat2dbdepth == max(dat2dbdepth))
  #if(length(mloc)==1) mb2[i,] = dat2db[mloc,]
  #else mb2[i,] = apply(dat2db[mloc,],2,mean)
}

# Get a set of medians of size 1000 by bayesian bootstrap
mbb2 = matrix(0,1000,2)
dat2dbb = matrix(0,200,2)
for(i in 1:1000){
  unif = sort(runif(19))
  weight = c(unif,1) - c(0,unif)
```

```
  sploc = sample(1:20, 200, replace = TRUE, prob = weight)
  dat2dbb[1:200,] = X20[sploc,]
  dat2dbbdepth = depth.halfspace(dat2dbb, dat2dbb)
  mloc = match(max(dat2dbbdepth), dat2dbbdepth)
  mbb2[i,] = dat2dbb[mloc,]
  #mloc = which(dat2dbbdepth == max(dat2dbbdepth))
  #if(length(mloc)==1) mbb2[i,] = dat2dbb[mloc,]
  #else mbb2[i,] = apply(dat2dbb[mloc,],2,mean)
}

# Get a Posterior Dirichlet Process path with
# a = 10, H = N(c(0,0), matrix(c(1,-.5,-.5,1),2,2))
# n = 1000
P.DP2 <- function(){
  n = 1000
  a = 5
  t = TRUE
  pi = 0
  sumpi = 0
  path = matrix(0,1000,2)

  for(i in 1:1000){
    u = runif(1)
    if(u<5/25)
      path[i, ] = mvrnorm(1, c(0,0), matrix(c(1,-0.5,-0.5,1),2,2))  # Normal
    # path[i, ] = rmvc(1, c(0,0), matrix(c(1,0,0,1),2,2))  # Cauchy
    # path[i, ] = cbind(rexp(1),rexp(1))  # Exponential
    else
      path[i, ] = X20[sample(c(1:20),1), ]

    pi[i] = rbeta(1,1,a)*(1-sumpi)
    sumpi = sumpi + pi[i]
  }
  return(list(path = path, pi = pi))
}

# Get a set of medians of size 1000 by Dirichlet Process
# Prior is normal N(c(0,0), matrix(c(1,-.5,-.5,1),2,2))
mdp2 = matrix(0,1000,2)
dat2ddp = matrix(0,200,2)
for(i in 1:1000){
  DP = P.DP2()
  sploc = sample(1:1000, 200, replace = TRUE, prob = DP$pi)
```

```
  dat2ddp = DP$path[sploc,]
  dat2ddpdepth = depth.halfspace(dat2ddp, dat2ddp)
  mloc = match(max(dat2ddpdepth), dat2ddpdepth)
  mdp2[i,] = dat2ddp[mloc,]
  #mloc = which(dat2ddpdepth == max(dat2ddpdepth))
  #if(length(mloc)==1) mdp2[i,] = dat2ddp[mloc,]
  #else mdp2[i,] = apply(dat2ddp[mloc,],2,mean)
}


#Graphs
par(mfrow=c(1,2))
plot(mb2,main = "Medians␣by␣Regular␣Bootstrap", sub = "Sample␣Size␣=␣20", xlab =
      "x", ylab = "y" )
depth.contours(mb2,depth="halfspace", main = "Contours␣of␣Tukey␣Medians", sub =
     "Sample␣Size␣=␣20",
                xlab = "x", ylab = "y", level = 5,frequency = 100,col="black",pch
                    = 20)

plot(mbb2,main = "Tukey␣Medians␣by␣Rubin's␣BB", sub = "Sample␣Size␣=␣20", xlab =
      "x", ylab = "y" )
depth.contours(mbb2,depth="halfspace", main = "Contours␣of␣Tukey␣Medians", sub =
      "Sample␣Size␣=␣20",
                xlab = "x", ylab = "y", level = 5,frequency = 100,col="black",pch
                    = 20)

plot(mdp2,main = "Tukey␣Medians␣by␣Lo's␣BB",sub = "Sample␣Size␣=␣20", xlab = "x"
      , ylab = "y" )
depth.contours(mdp2,depth="halfspace", main = "Contours␣of␣Tukey␣Medians", sub =
      "Sample␣Size␣=␣20",
                xlab = "x", ylab = "y", level = 5,frequency = 100,col="black",pch
                    = 20)
```

9. Bootstrap for the sample of size 200 and 500

```
# Draw a sample of size 200 from normal distribution
# N(c(0,0), matrix(c(1,-.5,-.5,1),2,2))

## The sample of size 500 is operated as the same way.

X200 = mvrnorm(200, c(0,0), matrix(c(1,-.5,-.5,1),2,2))
#X200 = rmvc(200,c(0,0),matrix(c(1,0,0,1),2,2)) # Standard Cauchy distribution
# X500 = cbind(rexp(200),rexp(200)) # Exponential
```

```

# Get a set of medians of size 1000 by regular bootstrap
mb = matrix(0,1000,2)
dat2db = matrix(0,200,2)
for(i in 1:1000){
  sploc = sample(1:200, 200, replace = TRUE)
  dat2db[1:200,] = X200[sploc,]
  dat2dbdepth = depth.halfspace(dat2db, dat2db)
  mloc = match(max(dat2dbdepth), dat2dbdepth)
  mb[i,] = dat2db[mloc,]
  #mloc = which(dat2dbdepth == max(dat2dbdepth))
  #if(length(mloc)==1) mb[i,] = dat2db[mloc,]
  #else mb[i,] = apply(dat2db[mloc,],2,mean)
}
# Get a set of medians of size 1000 by bayesian bootstrap
mbb = matrix(0,1000,2)
dat2dbb = matrix(0,500,2)
for(i in 1:1000){
  unif = sort(runif(199))
  weight = c(unif,1) - c(0,unif)
  sploc = sample(1:200, 500, replace = TRUE, prob = weight)
  dat2dbb[1:500,] = X200[sploc,]
  dat2dbbdepth = depth.halfspace(dat2dbb, dat2dbb)
  mloc = match(max(dat2dbbdepth), dat2dbbdepth)
  mbb[i,] = dat2dbb[mloc,]
  #mloc = which(dat2dbbdepth == max(dat2dbbdepth))
  #if(length(mloc)==1) mbb[i,] = dat2dbb[mloc,]
  #else mbb[i,] = apply(dat2dbb[mloc,],2,mean)
}

# Get a Posterior Dirichlet Process path with
# a = 10, H = N(c(0,0), matrix(c(1,-.5,-.5,1),2,2))
# n = 1000
P.DP <- function(){
  n = 1000
  a = 10
  t = TRUE
  pi = 0
  sumpi = 0
  path = matrix(0,1000,2)

  for(i in 1:1000){
    u = runif(1)
```

```
    if(u<10/210)
      path[i, ] = mvrnorm(1, c(0,0), matrix(c(1,-0.5,-0.5,1),2,2))  # Normal
     # path[i, ] = rmvc(1, c(0,0), matrix(c(1,0,0,1),2,2))  # Cauchy
      # path[i, ] = cbind(rexp(1),rexp(1))  # Exponential
    else
      path[i, ] = X200[sample(c(1:200),1), ]

    pi[i] = rbeta(1,1,a)*(1-sumpi)
    sumpi = sumpi + pi[i]
  }
  return(list(path = path, pi = pi))
}

# Get a set of medians of size 1000 by Dirichlet Process
# Prior is normal N(c(0,0), matrix(c(1,-.5,-.5,1),2,2))
mdp = matrix(0,1000,2)
dat2ddp = matrix(0,500,2)
for(i in 1:1000){
  DP = P.DP()
  sploc = sample(1:1000, 500, replace = TRUE, prob = DP$pi)
  dat2ddp = DP$path[sploc,]
  dat2ddpdepth = depth.halfspace(dat2ddp, dat2ddp)
  mloc = match(max(dat2ddpdepth), dat2ddpdepth)
  mdp[i,] = dat2ddp[mloc,]
  #mloc = which(dat2ddpdepth == max(dat2ddpdepth))
  #if(length(mloc)==1) mdp[i,] = dat2ddp[mloc,]
  #else mdp[i,] = apply(dat2ddp[mloc,],2,mean)
}

#Graphs
par(mfrow=c(1,2)) # par(mfrow=c(1,1))
plot(mb,main = ”Medians␣by␣Regular␣Bootstrap”, sub = ”Sample␣Size␣=␣200”, xlab =
     ”x”, ylab = ”y” )
depth.contours(mb,depth=”halfspace”, main = ”Contuors␣of␣Tukey␣Medians”, sub = ”
    Sample␣Size␣=␣200”,
                xlab = ”x”, ylab = ”y”, level = 5,frequency = 100,col=”black”,pch
                     = 20)

plot(mbb,main = ”Tukey␣Medians␣by␣Rubin's␣BB”, sub = ”Sample␣Size␣=␣200”, xlab =
     ”x”, ylab = ”y” )
depth.contours(mbb,depth=”halfspace”, main = ”Contuors␣of␣Tukey␣Medians”, sub =
    ”Sample␣Size␣=␣200”,
                xlab = ”x”, ylab = ”y”, level = 5,frequency = 100,col=”black”,pch
```

```
                                = 20)

plot(mdp,main = "Tukey Medians by Lo's BB",sub = "Sample Size = 200", xlab = "x"
    , ylab = "y" )
depth.contours(mdp,depth="halfspace", main = "Contuors of Tukey Medians", sub =
    "Sample Size = 200",
                    xlab = "x", ylab = "y", level = 5,frequency = 100,col="black",pch
                                = 20)
```

# Bibliography

[1] Al Labadi, L. *On New Constructive Tools in Bayesian Nonparametric Inference.* PhD thesis, University of Ottawa, 2015.

[2] Andrews, D. W. K. Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space. *Econometrica 68*, 2 (2000), 399–405.

[3] Bai, Z., and He, X. Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *The Annals of Statistics 27*, 5 (1999), 1616–1637.

[4] Bickel, P. J., and Freedman, D. A. Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics 9*, 6 (1981), 1196–1217.

[5] Bickel, P. J., Götze, F., and van Zwet, W. R. Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses. *Statistica Sinica 7* (1997), 1–31.

[6] Bickel, P. J., and Ren, J.-J. The m out of n Bootstrap and Goodness of Fit Tests with Double Censored Data. *Robust Statistics, Data Analysis, and Computer Intensive Methods* (1996), 35–47.

[7] Bickel, P. J., and Sakov, A. On the Choice of m in the m Out of n Bootstrap and Confidence Bounds for Extrema. *Statistica Sinica 18*, 3 (2008), 967–985.

[8] Bickel, P. J., and Yahav, J. A. Richardson Extrapolation and the Bootstrap. *Journal of the American Statistical Association 83*, 402 (1988), 387–393.

[9] Bondesson, L. On Simulation from Infinitely Divisible Distributions. *Advances in Applied Probability 14*, 4 (1982), 855–869.

[10] Donoho, D. L., and Gasko, M. Breaking Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. *The Annals of Statistics 20*, 4 (1992), 1803–1827.

[11] Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics 7*, 1 (1979), 1–26.

[12] Ferguson, T. S. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics 1*, 2 (1973), 209–230.

[13] Götze, F., and Račkauskas, A. Adaptive choice of bootstrap sample sizes. *Lecture Notes–Monograph Series 36* (2001), 286–309.

[14] Hall, P. Asymptotic Properties of the Bootstrap for Heavy-Tailed Distributions. *The Annals of Probability 18*, 3 (1990), 1342–1360.

[15] Hall, P., Horowitz, J. L., and Jing, B.-Y. On Blocking Rules for the Bootstrap with Dependent Data. *Biometrika 82*, 3 (1995), 561–574.

[16] Horowitz, J. L. The Bootstrap. *Handbook of Econometrics 5* (2001), 3159–3228.

[17] Ishwaran, H., James, L. F., and Zarepour, M. An Alternative to the m ut of n Bootstrap. *Journal of Statistical Planning and Inference 139* (2009), 788–801.

[18] Ishwaran, H., and Zarepour, M. Exact and Approximate Sum Representations for the Dirichlet Process. *The Canadian Journal of Statistics 30*, 2 (2002), 269–283.

[19] Lee, S. M. S. On a Class of m out of n Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*, 4 (1999), 901–911.

[20] Liu, R. Y. On a Notion of Data Depth Based on Random Simplices. *The Annals of Statistics 18*, 1 (1990), 405–414.

[21] Liu, R. Y., Parelius, J. M., and Singh, K. Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. *The Annals of Statistics 27*, 3 (1999), 783–858.

[22] Lo, A. Y. A Large Sample Study of the Bayesian Bootstrap. *The Annals of Statistics 15*, 1 (1987), 360–375.

[23] Mahalanobis, P. C. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences (India) 2*, 1 (1936), 49–55.

[24] Mammen, E. *When Does Bootstrap Work? Asymptotic Results and Simulations.* Springer, New York, 1992.

[25] Massé, J.-C. Asymptotics for the Tukey Median. *Journal of Multivariate Analysis 81*, 2 (2002), 286–300.

[26] Muliere, P., and Secchi, P. Bayesian Nonparametric Predictive Inference and Bootstrap Techniques. *Annals of the Institute of Statistical Mathematics 48*, 4 (1996), 663–673.

[27] Nolan, D. On Min-Max Majority and Deepest Points. *Statistics and Probability Letters 43*, 4 (1999), 325–333.

[28] Oja, H. Descriptive Statistics for Multivariate Distributions. *Statistics and Probability Letters 1*, 6 (1983), 327–332.

[29] Politis, D. N., Romano, J. P., and Wolf, M. *Subsampling.* Springer Series in Statistics, New York, 1999.

[30] Rousseeuw, P. J., Ruts, I., and Tukey, J. W. The Bagplot: A Bivariate Boxplot. *The American Statistician 53*, 4 (1999), 382–387.

[31] Rubin, D. B. The Bayesian Bootstrap. *The Annals of Statistics 9*, 1 (1981), 130–134.

[32] SAKOV, A. *Using the m out of n Bootstrap in Hypothesis Testing.* PhD thesis, University of California, Berkeley, 1998.

[33] SETHURAMAN, J. A Constructive Definition of Dirichlet Priors. *Statistica Sinica 4* (1994), 639–650.

[34] SINGH, K. On the Asymptotic Accuracy of Efron's Bootstrap. *The Annals of Statistics 9*, 6 (1981), 1187–1195.

[35] SWANEPOEL, J. W. A note on proving that the (modified) bootstrap works. *Communications in Statistics-Theory and Methods 15*, 11 (1986), 3193–3203.

[36] TUKEY, J. W. Mathematics and the Picturing of Data. *Proceedings of the International Congress of Mathematicians (Vancouver) 2* (1975), 523–531.

[37] WILKS, S. S. *Mathematical Statistics.* Wiley, New York, 1962.

[38] ZAREPOUR, M., AND AL LABADI, L. On a Rapid Simulation of the Dirichlet Process. *Statistics and Probability Letters 82*, 5 (2012), 916–924.

[39] ZAREPOUR, M., AND KNIGHT, K. Bootstrapping Point Processes with Some Applications. *Stochastic Processes and Their Applications 84* (1999), 81–90.

[40] ZUO, Y., AND SERFLING, R. General Notions of Statistical Depth Function. *The Annals of Statistics 28*, 2 (2000), 461–482.