Classification of Symmetric α-Stable Noise

A Project Report

submitted by

ALLEN JOB

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS

JUNE 2017

THESIS CERTIFICATE

This is to certify that the thesis titled **Classification of Symmetric α-Stable Noise**, submitted by **Allen Job**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bona fide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Sheetal Kalyani
Research Guide
Associate Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date:

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my guide, Associate Professor, Dr. Sheetal Kalyani, who has been my mentor at IIT Madras. Her constant guidance, valuable insights into the problem at hand and timely advice has made my project a smooth and enriching learning experience.

I am grateful to the faculty members of EE department for the knowledge and wisdom they have shared during coursework and presentations.

I would like to thank HPCE team, P.G.Senapathy center for computing Resources, IIT Madras, for the Virgo Super Cluster which has helped me in running the majority of my simulations.

I am thankful to all the technical and non-technical staff of Electrical Engineering Department for their services.

My sincere thanks to everyone in CSD154. It was great sharing the laboratory with you.

Special mentions to Abhishek, Murali, Vishnu, Manoj, Thulasi, Arun and Sreejith. On several occasions, I had discussed my ideas with them for clarity and had banked on their abilities for tackling problems.

And as always, I'm indebted to my family and friends, whose unconditional support has been instrumental in the success of all my endeavours.

Thanks for all your encouragement!

ABSTRACT

KEYWORDS: Noise; GCLT; Stable distribution; Classification; Sample size; K-L Divergence; L-Kurtosis; FLOM.

Noise modelling in communication systems often decides the level of the efficiency of its operation. Depending on the scenario, the underlying noise source characteristics differ over a large set of probability distributions by being thin tailed to heavy tailed. For noise sources exhibiting a tail decay rate $\sim |x|^{-\alpha}$, as a direct consequence of the Generalized Central Limit Theorem, aggregate noise behaviour turn out to be stable distributed. Symmetric α-Stable noise models are widely used in systems with impulsive noise. Along with these models comes the requirement of classifying the noise to the suitable SαS distribution. Here, different methods to map given noise data to a set of selected pdfs is explored, with focus on SαS pdfs. The performance of these classifiers with sample size is examined. Distance metrics are investigated as tools for classification of the data. For this purpose, divergence calculation methods are examined in detail. A two-stage classification method based on L-Kurtosis and Fractional Lower Order Moments is proposed for assigning noise data to suitable SαS distribution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

K-L	Kullback-Leibler
PDF	Probability density function
GSNR	Geometric signal to noise ratio
SαS	Symmetric α-stable distribution
GCLT	Generalized Central Limit Theorem
WKV_Alg_A	Wang, Kulkarni, Verdu algorithm A
WKV_Alg_C	Wang, Kulkarni, Verdu algorithm C
FPC	Fernando Perez-Cruz algorithm
KDE	Kernel density estimation
NNDE	Nearest neighbour density estimation
KLIEP	Kullback-Leibler importance estimation procedure
GM-KLIEP	Gaussian mixture model based Kullback-Leibler importance estimation procedure
uLSIF	Unconstrained least squares importance fitting
LSTT	Least squares two-samples test
MMD	Maximum mean discrepancy

NOTATION

$p(x)$	PDF of the given noise samples
$q_i(x)$	i^{th} reference PDF
$r_i(x)$	i^{th} density ratio, $\frac{p(x)}{q_i(x)}$
$\{x_j\}_{j=1}^{n}$	Given noise samples $\sim p(x)$
$\{y_j\}_{j=1}^{m}$	Reference samples $\sim q_i(x)$
$\mathcal{T}\{f(.)\}$	Tail decay rate of $f(.)$ given by the asymptotic approximation, $\lim_{x\to\infty} \int_x^{\infty} f(x)dx$
$\mathcal{X}$	Sample set $\{x_j\}_{j=1}^{n}$
$\mathcal{Y}$	Sample set $\{y_j\}_{j=1}^{m}$

CHAPTER 1

INTRODUCTION

In communication system design, one of the fundamental considerations is the noise model adopted. The efficiency of the receiver in extracting the transmitted data with sufficient fidelity, heavily depends on its ability to sift out the non-signal part of the received data. For devising an optimal receiver structure, the noise model characteristics has to be nearly identical to that of the actual random fluctuations the signal undergoes in the channel. In most scenarios, the individual noise sources can be considered independent and identically distributed. Gaussian distributed models are ubiquitous by virtue of Central Limit Theorem and yields good results, provided the noise sources are finite variance or non-impulsive in nature.

But when the noise sources are impulsive in nature, normal distribution does not adequately capture the noise behaviour. With slower tail decay rates for the aggregate noise, the resulting distribution becomes heavy tailed. Proceeding with Gaussian noise assumption in scenarios where heavy-tailed noise is present will lead to erroneous signal extraction and poor system performance. Alternatively, the approach using a heavier-tailed model like Cauchy distribution in the place of Gaussian noise is sub-optimal. So it is imperative that the relevant noise model is selected to ensure proper and optimal operation.

Consider the generic model where the underlying distribution of the noise can vary between any of the various possible distributions at different instances. Since it is impossible for the system to keep track of all those distributions and map each data to the suitable distribution, a compromise can be made by having a set of most probable pdfs, each with different tail decay rates and map the noise sample set to the most suitable among these distributions. Any probability distance metric can be used to do this classification of the received data. The efficiency of the classification will then depend on the distance metric chosen, as well as the metric calculation strategy. In this project, Kullback-Leibler distance, Pearson divergence, and maximum mean discrepancy were explored as candidates for this classification, with primary focus on the K-L divergence.

The generalized Central Limit Theorem states that the only possible non-trivial limit of the sum of n independent and identically distributed random variables tends to be stable distributed as $n \rightarrow \infty$. In the case of impulsive noise sources, the α-stable distribution model characterizes the aggregate noise behaviour better than a Gaussian model. By selecting the suitable characteristic exponent α, heavy-tailed distributions like Cauchy distribution($\alpha = 1$) can be used to model the noise. The stable model also includes the Gaussian model as a limiting case in the form of Symmetric α-Stable model with $\alpha = 2$.

The Stable distribution is specified by its characteristic function as:

$$\Phi(t) = e^{(i\delta t - |\gamma t|^{\alpha} B_{t,\alpha})} \tag{1.1}$$

where

$$B_{t,\alpha} = \begin{cases} 1 - i\beta sgn(t) \, tan\left(\frac{\pi\alpha}{2}\right), & \alpha \neq 1 \\ 1 - i\beta sgn(t) \left(\frac{2}{\pi}\right) log\left(|t|\right), & \alpha = 1 \end{cases} \tag{1.2}$$

with the parameters:

- $\alpha \rightarrow$ Characteristic Exponent
- $\beta \rightarrow$ Skewness
- $\gamma \rightarrow$ Scale
- $\delta \rightarrow$ Location.

Symmetric α-stable distribution is stable distribution with skewness parameter, $\beta = 0$.

When i.i.d. impulsive noise sources are considered, the generic model can be replaced by a set of SαS distributions with varying characteristic exponent. A requirement of large sample sizes is inherent while dealing with heavy-tailed data. Deducing the nature of the underlying distribution from low number of samples is a hard problem. In this project, order statistics based moments known as L-moments, as well as fractional lower order moments (FLOMs), are used for processing the noise data instead of conventional moments. An estimator of low complexity for the scale parameter based on FLOMs is proposed. A two-stage classification algorithm is provided for mapping the samples to SαS pdf with the nearest α value. The first stage is based on the L-Kurtosis

metric of the noise samples which helps in distinguishing between Gaussian and non-Gaussian distributed data. Then the non-Gaussian distributed data is further mapped to the fitting α value using a FLOM based algorithm.

1.1 Thesis Outline

The thesis is organized as follows:

Chapter 2 gives a detailed formulation of the general classification procedure adopted, the assumptions taken on the data samples, the probability distributions of the test data, and the reference set of probability distributions to which the data is to be mapped.

The different classification methods based on distance metrics are discussed in Chapter 3. The simulation results/observations for the classification are also included.

In Chapter 4, estimators for the scale (γ) and exponent (α) parameters of symmetric α-stable distributed data and low complexity methods for classifying given data to a set of SαS distributions are proposed.

1.2 Literature Survey

We first give a brief summary of literature in the context of α-stable PDFs then summarize literature on noise classification.

Gnedenko and Kolmogorov (1968); Zolotarev (1986) presented the generalized central limit theorem and the properties of the stable distribution. McCulloch (1986) gave low complexity, quantile based parameter estimators using lookup tables for α-stable distributions.Kogon and Williams (1998) introduced characteristic function regression based estimator and Kuruoglu (2001) derived closed form estimators for the parameters. Sufficient mathematical background on handling stable distributed data including the Fractional Lower Order Moments, is available in Arce (2005).

Past works have pointed out that non-Gaussian behaviour is observed in many real life scenarios. In various domains, heavy-tailed models based on α-stable distribution

were put forward in Georgiou *et al.* (1999); Briassouli *et al.* (2005); Niranjayan and Beaulieu (2008); Gulati *et al.* (2010).

By modelling node distribution by spatial poisson process and individual interferer amplitudes with spherically symmetric pdfs, Ilow and Hatzinakos (1998); Win *et al.* (2009) showed that the aggregate wireless network interference amplitude followed Symmetric α-Stable distribution. Both proceeded to derive the relation between the parameters of the SαS distributed interference and the network parameters. Meanwhile Gulati *et al.* (2010) had modelled a special case of co-channel interference using SαS models.

In other words, there exists a rich literature on both α-stable distribution and its applications. What we aim to study in this project is that given a set of noise samples, can we classify these as being closest to one of a set of α-stable distributions.

Classification based on K-L divergence involves computation of the divergence value. Wang *et al.* (2005); Perez-Cruz (2008) provided methods based on data partition to calculate the K-L distance. Density ratio estimation techniques were explored in detail by Sugiyama *et al.* (2009). Sugiyama *et al.* (2012) contrasted the existing density estimation methods including kernel density estimation Parzen (1962) and nearest neighbour density estimation methods. Sugiyama *et al.* (2011) compared the Least Squares Two-samples Test (LSTT) with the Maximum mean discrepancy (MMD) based homogeneity test.

CHAPTER 2

PROBLEM SETUP

2.1 Problem

2.1.1 Problem Definition

The noise samples $\{x_j\}_{j=1}^{n}$ from an unknown probability distribution $p(x)$ is available. A set of reference probability distributions $q_i(x)$ of varying levels of tail decay rate $\mathcal{T}\{q_i(x)\}$ are provided to model the noise data. The problem is to map the data to the best/nearest reference probability distribution.The classification of the data to the reference pdf is based on the pdf of the noise samples.

Let H_i be the hypothesis such that the noise sample is best modelled by the reference probability distribution $q_i(x)$.

$$\mathrm{H}_i : \{x_j\}_{j=1}^{n} \sim q_i(x) \tag{2.1}$$

The optimum decision rule in the scenario, $\delta\left(\{x_j\}_{j=1}^{n}\right)$ is such that $p(x)$ is adequately approximated by $q_i(x)$ and $\mathcal{T}\{p(x)\}$ vanishes at a similar or faster rate than $\mathcal{T}\{q_i(x)\}$.

2.1.2 Assumptions

- The real and imaginary parts of the noise data are assumed to be independent.
- Noise probability distribution functions discussed here is for one component (real/imaginary) of the actual noise data.
- From the assumption of independence, noise probability distribution function is one dimensional.
- Noise probability distribution function $p(x)$ is symmetric about $x = 0$[1].

[1] zero mean, zero skewness

2.2 Reference Probability Distributions

The reference distributions considered here are all univariate distributions. Tail decay rate of a PDF $p(x)$ is given by, $\lim_{x\to\infty} \mathrm{P}\left(X > x\right) = \mathcal{T}\left\{p(x)\right\}$.

- *Normal Distribution*:
  $f_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ where $-\infty < \mu < \infty, \quad \sigma > 0$
  $\mathcal{T}\left\{f_N(x)\right\}) \sim \frac{\sqrt{2}e^{-\frac{x^2}{2\sigma^2}}}{x\sqrt{\pi}}$
- *Laplacian Distribution*:
  $f_L(x) = \frac{1}{2\gamma} e^{-\frac{|x-\mu|}{\gamma}}$ where $-\infty < \mu < \infty, \quad \gamma > 0$
  $\mathcal{T}\left\{f_L(x)\right\}) \sim \frac{1}{2} e^{-\frac{x}{\gamma}}$
- *Holtsmark Distribution*:
  $f_H(x) = \mathrm{Stable}\left(\alpha = \frac{3}{2}, \beta = 0, \gamma, \delta\right)$ where $\gamma > 0, \quad -\infty < \delta < \infty$
  $\mathcal{T}\left\{f_H(x)\right\}) \sim \frac{\gamma^{\frac{3}{2}}}{2\sqrt{2\pi}x^{\frac{3}{2}}}$
- *Cauchy Distribution*:
  $f_C(x) = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2+\gamma^2}$ where $-\infty < x_0 < \infty, \quad \gamma > 0$
  $\mathcal{T}\left\{f_C(x)\right\}) \sim \frac{\gamma}{\pi x}$
- *SαS Distribution*:
  $f_{S\alpha S}(x) = \mathrm{Stable}\left(\alpha, \beta = 0, \gamma, \delta\right)$ where $\gamma > 0, \quad -\infty < \delta < \infty$
  $\mathcal{T}\left\{f_{S\alpha S}(x)\right\} \sim |x|^{-\alpha}$

with all distributions having support as $-\infty < x < \infty$. The Normal, Cauchy, Holtsmark distributions are special cases of the symmetric α-stable distribution with $\alpha = 1, 2, 1.5$ respectively. Additionally, SαS pdfs with $\alpha = 1.25, 1.75$ are also used as reference pdfs. For a fair comparison, the distribution parameters are taken so as to make the *Geometric Power*[2] Arce (2005) same for all. Since the noise is modelled as zero mean, the location parameter is selected as zero which eases matching the geometric power.The parameters thus determined are as follows:

- Fix Normal distribution parameters: $\mu = 0, \quad \sigma = \sigma_N$
- Cauchy parameters satisfying the constraint: $x_0 = 0, \quad \gamma = \frac{\sigma_N}{\sqrt{2}} * e^{-\frac{1}{2}C_e}$
- Laplacian parameters satisfying the constraint: $\mu = 0, \quad \gamma = \frac{\sigma_N}{\sqrt{2}} * e^{\frac{1}{2}C_e}$
- Holtsmark parameters satisfying the constraint: $\delta = 0, \quad \gamma = \frac{\sigma_N}{\sqrt{2}} * e^{-\frac{1}{6}C_e}$
- SαS parameters satisfying the constraint: $\delta = 0, \quad \gamma = \frac{\sigma_N}{\sqrt{2}} * e^{-\left(\frac{2-\alpha}{2\alpha}\right)C_e}$

[2] Geometric Power, $S_0(X) = e^{\mathbb{E}[\log|X|]}$

where C_e is the Euler-Mascheroni constant.

The distance between the various distributions can be quantified in terms of K-L divergence. The K-L Divergence from $q(x)$ to $p(x)$ is defined as:

$$\text{`D}_{KL}\left(p||q\right) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx \tag{2.2}$$

This expression can be numerically computed[3]. The values are computed for different choices of σ_N. It is observed that the divergence values are consistent over variations in scale parameter.

Table 2.1: Kullback-Leibler Divergence $(D_{KL}\left(P\,||\,Q\right))$ between probability distributions P & Q

P \ Q	$f_N(x)$	$f_C(x)$	$f_L(x)$	$f_H(x)$
$f_N(x)$	0	0.189195	0.0617442	0.0538536
$f_C(x)$	∞	0	∞	0.0783718
$f_L(x)$	0.174293	0.0912061	0	0.0276048
$f_H(x)$	∞	0.0623275	0.169214	0

The above values are computed based on the actual PDF values. In our problem, the noise sample set $\{x_j\}_{j=1}^{n}$ is available instead of the actual noise PDF. Hence, for classification purposes, the empirical estimates of the distance metrics need to used. There are multiple methods available for computed these empirical values from the samples. The focus of the problem is classification of the sample set, not the accuracy of the distance estimates. Therefore, it is sufficient if the distance estimates can help in proper classification.

2.3 Input Probability Distributions

The input probability distributions used are the same as the reference probability distributions.

[3]Computed using Mathematica

CHAPTER 3

CLASSIFICATION METHODS

3.1 Classification Methods

Distance based methods are used to classify the received data into the reference distribution models. K-L divergence, Pearson divergence, and Maximum mean discrepancy are explored as the distance metrics. Different classification methods exist depending on the way the distance metric is computed. The parameters of Laplace distributed data is calculated using their ML estimators and the scale parameter for SαS distributed data is calculated using Algorithm 2 on page 31. Reference data is generated from the reference PDFs wherever required.

3.2 Classification Based on K-L Divergence

In literature, K-L divergence has been used as a tool to measure the goodness of fit Kapur and Kesavan (1992). The K-L divergence between the given sample distribution and the reference distributions is calculated and the given noise sample is mapped to the reference distribution to which it has minimum divergence. It is framed as a multiple hypothesis testing problem below:

- Let H_i be the hypothesis such that the sample is best modelled by the reference probability distribution $q_i(x)$.

$$\mathrm{H}_i : \{x_j\}_{j=1}^{m} \sim q_i(x) \tag{3.1}$$

- Calculate the K-L divergences, $\mathrm{D}_{KL}\left(p||q_i\right)$ from $q_i(x)$ to $p(x)$
- Decision rule is

$$\delta = \underset{i}{\operatorname{argmin}} \ \left\{\mathrm{D}_{KL}\left(p||q_i\right)\right\} \tag{3.2}$$

K-L Divergence Calculation

We have,

$$\begin{aligned} \mathrm{D}_{KL}\left(p||q\right) &= \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx \\ &= \mathbb{E}_{p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] \\ &= \mathbb{E}_{p(x)}\left[\log\left(r\left(x\right)\right)\right] \end{aligned} \tag{3.3}$$

‘

Hence, the problem of calculating the K-L divergence can be expressed as finding the expectation of the density ratio, $r(x) = \frac{p(x)}{q(x)}$ w.r.to $p(x)$. Empirically,

$$\begin{aligned} \mathbb{E}_{p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] &\approx \sum_{j=1}^{n} p(x_j)\log\left(\hat{r}(x_j)\right) \\ &\approx \frac{1}{n}\sum_{j=1}^{n}\log\left(\hat{r}(x_j)\right) \end{aligned} \tag{3.4}$$

yields the K-L divergence estimate Sugiyama *et al.* (2009, 2012).

3.2.1 Data Partition Methods

Data partition methods require two sets of samples. Here, we have the given noise sample $\{x_j\}_{j=1}^{n}$ and the reference sample $\{y_j\}_{j=1}^{m}$ generated from the reference pdf. The data partition based methods calculate the K-L divergence by partitioning the reference sample and/or the given noise sample into equal number of partitions according to a simple or adaptive rule. The probability density(ratio) values at each partition is assessed from the number of samples that fall into each partition. The ratio of the PDF values thus calculated is used to find the K-L divergence value as follows:
If $\hat{p}(x)$ is the calculated probability distribution of the given noise and $\hat{q}(x)$ is the calculated reference probability distribution,

$$\hat{D}_{KL}(p||q) = D_{KL}(\hat{p}||\hat{q}) = \sum \hat{p}(x)\log(\frac{\hat{p}(x)}{\hat{q}(x)}) \tag{3.5}$$

where the summation is over the partitions.

Wang, Kulkarni, Verdu algorithm A Wang *et al.* (2005)

The reference sample-set $\{y_j\}_{j=1}^{m}$ is partitioned uniformly by dividing into empirically equal segments (i.e., each segment except the ones nearest to $-\infty$ and ∞ contain equal number of samples). Then the number of samples from $\{x_j\}_{j=1}^{n}$ in each segment is tallied and used to calculate the empirical probability of each segment. These empirical probability values are used to calculate the K-L divergence using Equation (3.5).

Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified to nearest SαS distribution. Table 3.1 to Table 3.4 give the results of classification in percentage.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	42.5	10.0	29.6	14.9	2.7	0.3
Laplace	9.8	38.6	10.8	17.5	18.8	4.5
$\alpha = 1.75$	29.7	13.5	28.1	20.2	8.1	0.4
$\alpha = 1.5$	12.6	13.4	27.0	27.3	17.1	2.6
$\alpha = 1.25$	3.6	4.7	15.5	25.9	34.9	15.4
$\alpha = 1$	0.3	0.4	3.1	11.2	29.3	55.7

Table 3.1: WKV_Alg_A, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	55.4	4.1	31.7	8.6	0.2	0.0
Laplace	2.7	54.1	8.8	21.1	12.8	0.5
$\alpha = 1.75$	30.6	7.5	41.0	19.5	1.4	0.0
$\alpha = 1.5$	8.0	8.7	26.0	39.7	17.3	0.3
$\alpha = 1.25$	0.7	0.9	5.1	24.9	54.0	14.4
$\alpha = 1$	0.0	0.0	0.0	1.8	26.1	72.1

Table 3.2: WKV_Alg_A, n = 400

Wang, Kulkarni, Verdu algorithm C Wang *et al.* (2005)

The uniform partitioning in algorthm A is not efficient enough in cases where $\frac{p(x)}{q(x)}$ is high. Algorithm C employs a reasonable partition scheme to improve this by using

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	66.7	0.1	31.8	1.4	0.0	0.0
Laplace	0.7	74.0	3.4	17.7	4.2	0.0
$\alpha = 1.75$	34.4	4.4	48.6	12.4	0.2	0.0
$\alpha = 1.5$	2.7	4.8	25.2	55.0	12.3	0.0
$\alpha = 1.25$	0.0	0.0	0.8	20.5	69.8	8.9
$\alpha = 1$	0.0	0.0	0.0	0.0	13.1	86.9

Table 3.3: WKV_Alg_A, n = 800

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	88.9	0.0	11.1	0.0	0.0	0.0
Laplace	0.0	93.4	0.4	6.2	0.0	0.0
$\alpha = 1.75$	18.3	1.6	74.1	6.0	0.0	0.0
$\alpha = 1.5$	0.0	1.1	11.3	82.0	5.6	0.0
$\alpha = 1.25$	0.0	0.0	0.0	7.9	90.4	1.7
$\alpha = 1$	0.0	0.0	0.0	0.0	2.5	97.5

Table 3.4: WKV_Alg_A, n = 1600

finer partitions where the rate of change is high, and coarser partitions where the rate of change is low.

Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified to nearest SαS distribution. Table 3.5 to Table 3.8 give the results of classification in percentage.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	52.8	9.7	26.5	9.4	1.5	0.1
Laplace	37.9	30.9	12.8	10.3	7.0	1.1
$\alpha = 1.75$	43.9	13.3	26.2	13.0	3.5	0.1
$\alpha = 1.5$	34.7	10.7	24.4	21.4	8.2	0.6
$\alpha = 1.25$	19.0	4.0	21.8	23.2	22.4	9.6
$\alpha = 1$	10.2	0.2	13.1	16.1	26.5	33.9

Table 3.5: WKV_Alg_C, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	57.5	4.2	31.8	6.5	0.0	0.0
Laplace	31.8	41.9	5.4	12.0	8.6	0.3
$\alpha = 1.75$	53.6	6.8	28.8	10.3	0.5	0.0
$\alpha = 1.5$	41.5	6.1	23.5	21.0	7.9	0.0
$\alpha = 1.25$	23.4	0.7	16.2	20.4	30.2	9.1
$\alpha = 1$	8.2	0.0	4.8	5.5	26.4	55.1

Table 3.6: WKV_Alg_C, n = 400

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	79.4	0.4	20.0	0.2	0.0	0.0
Laplace	12.9	74.9	3.3	8.5	0.4	0.0
$\alpha = 1.75$	61.5	3.0	29.6	5.9	0.0	0.0
$\alpha = 1.5$	37.4	5.0	17.1	34.4	6.1	0.0
$\alpha = 1.25$	10.5	0.1	3.5	20.8	59.7	5.4
$\alpha = 1$	5.0	0.0	0.0	0.8	16.7	77.5

Table 3.7: WKV_Alg_C, n = 800

Fernando Perez-Cruz algorithm Perez-Cruz (2008)

Another method is by constructing a piecewise linear CDF function from the samples empirically and use the constructed ECDFs to evaluate the density ratio at the input sample points. The evaluated ratio can then be used to get a K-L divergence estimate.

Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified to nearest SαS distribution. Table 3.9 to Table 3.12 give the results of classification in percentage.

Inferences

The algorithms A and C from Wang *et al.* (2005) shows classification accuracy that gets better with sample size. It is observed that FPC based K-L divergence classification performs poorly, even though it yields the best K-L divergence values among the three, at high sample sizes ($\sim 10^6$). The algorithm A performs best among the data partition methods discussed.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	88.3	0.0	11.7	0.0	0.0	0.0
Laplace	2.3	92.5	0.5	4.7	0.0	0.0
$\alpha = 1.75$	67.1	0.6	30.2	2.1	0.0	0.0
$\alpha = 1.5$	30.2	1.6	11.5	53.0	3.7	0.0
$\alpha = 1.25$	1.5	0.0	0.1	11.9	85.6	0.9
$\alpha = 1$	2.9	0.0	0.0	0.0	3.2	93.9

Table 3.8: WKV_Alg_C, n = 1600

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	56.3	14.0	17.4	9.3	2.9	0.1
Laplace	84.1	5.9	7.2	1.7	0.8	0.3
$\alpha = 1.75$	70.4	9.0	13.4	5.5	1.6	0.1
$\alpha = 1.5$	75.9	5.6	13.2	3.8	1.5	0.0
$\alpha = 1.25$	76.7	0.5	16.6	4.3	1.6	0.3
$\alpha = 1$	67.2	0.1	22.5	6.8	2.9	0.5

Table 3.9: FPC, n = 200

3.2.2 Non-parametric Methods for Density Estimation

The individual probability densities of the given noise and reference sample can be evaluated using one of the non-parametric metrics and can be used to calculate an empirical approximation of the K-L divergence. The non-parametric density estimation approach is as follows:

- Let $\{x_j\}_{j=1}^n \in \mathcal{D}$ (domain) and $\{x_j\}_{j=1}^n \sim p(x)$
- Then for any region $\mathcal{R}_\mathcal{D} \in \mathcal{D}$ of volume V, probability of $x \in \mathcal{R}_\mathcal{D}$ can be approximated as,
$$P_{\mathcal{R}_\mathcal{D}} \approx V * p(x') \tag{3.6}$$
where x' is any point in the region $\mathcal{R}_\mathcal{D}$
- If m samples fall in the region $\mathcal{R}_\mathcal{D}$ out of the n samples, then probability of $x \in \mathcal{R}_\mathcal{D}$ can again be approximated as,
$$P_{\mathcal{R}_\mathcal{D}} \approx \frac{m}{n} \tag{3.7}$$
Using Equation (3.6) and Equation (3.7), we have an estimate of the PDF as:
$$p(x) \approx \frac{m}{nV} \tag{3.8}$$
Since the PDF estimate depends on how m and V is chosen, the quality of the approximation depends on the choice of $\mathcal{R}_\mathcal{D}$.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	59.9	10.2	21.6	7.0	1.2	0.1
Laplace	91.4	5.2	1.7	1.2	0.4	0.1
$\alpha = 1.75$	81.5	6.7	8.0	3.1	0.7	0.0
$\alpha = 1.5$	88.4	3.3	5.9	1.8	0.5	0.1
$\alpha = 1.25$	87.9	0.5	8.1	2.6	0.7	0.2
$\alpha = 1$	75.1	0.0	14.7	6.7	2.6	0.9

Table 3.10: FPC, n = 400

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	69.0	5.2	21.8	3.7	0.3	0.0
Laplace	92.7	5.7	0.6	0.9	0.1	0.0
$\alpha = 1.75$	90.5	3.8	4.5	1.0	0.2	0.0
$\alpha = 1.5$	95.7	1.2	2.0	0.8	0.3	0.0
$\alpha = 1.25$	92.5	0.0	4.3	1.6	1.1	0.5
$\alpha = 1$	80.2	0.0	11.9	4.4	2.5	1.0

Table 3.11: FPC, n = 800

Non-parametric methods explained here can be used to obtain the values $\{p(x_j)\}_{j=1}^n$ and $\{r(x_j)\}_{j=1}^n$.

Kernel Density Estimation

Kernel density estimation (KDE) Parzen (1962); Sugiyama *et al.* (2012) is a non-parametric approach for approximating the probability density function of a sample set. The PDF of the underlying distribution is obtained as the mean of the $kernel$ functions centred at the sample points and of a suitable bandwidth. The kernel controls the weights of assigning samples to different regions $\mathcal{R}_\mathcal{D} \in \mathcal{D}$ based on proximity to other samples. Hence, each sample x_j contributes to every region $\mathcal{R}_\mathcal{D}$ with different weight.

For a sample set $\{x_j\}_{j=1}^n$ and a normalized kernel $K_B(x, x')$ with bandwidth B, the KDE based PDF estimate is given by:

$$\hat{p}_{KDE}(x) := \frac{1}{n}\sum_{j=1}^{n} K_B(x, x_j) \tag{3.9}$$

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	74.2	3.9	19.6	2.3	0.0	0.0
Laplace	92.3	6.1	0.6	0.8	0.2	0.0
$\alpha = 1.75$	96.6	1.4	1.8	0.2	0.0	0.0
$\alpha = 1.5$	98.6	0.1	0.8	0.5	0.0	0.0
$\alpha = 1.25$	95.8	0.0	2.2	1.0	0.9	0.1
$\alpha = 1$	78.5	0.0	9.3	4.1	4.6	3.5

Table 3.12: FPC, n = 1600

For Gaussian Kernel,

$$K_\sigma(x, x') = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{\left(\frac{(x-x')^2}{2\sigma^2}\right)}$$
$$\hat{p}_{DE}(x) = \frac{1}{n\sigma\sqrt{2\pi}} \sum_{j=1}^{n} \mathrm{e}^{\left(\frac{(x-x')^2}{2\sigma^2}\right)} \tag{3.10}$$

Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified to the nearest SαS distribution. Table 3.13 to Table 3.16 give the results of classification in percentage.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	95.6	2.4	2.0	0.0	0.0	0.0
Laplace	0.1	97.4	2.0	0.5	0.0	0.0
$\alpha = 1.75$	4.9	9.7	74.6	10.8	0.0	0.0
$\alpha = 1.5$	0.4	6.2	12.0	71.9	9.5	0.0
$\alpha = 1.25$	0.0	0.8	0.1	12.9	78.8	7.4
$\alpha = 1$	0.0	0.0	0.0	0.0	8.4	91.6

Table 3.13: KDE, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	99.7	0.0	0.3	0.0	0.0	0.0
Laplace	0.0	99.3	0.6	0.1	0.0	0.0
$\alpha = 1.75$	0.6	2.3	92.1	5.0	0.0	0.0
$\alpha = 1.5$	0.0	1.2	6.1	88.1	4.6	0.0
$\alpha = 1.25$	0.0	0.0	0.0	5.8	91.9	2.3
$\alpha = 1$	0.0	0.0	0.0	0.0	4.5	95.5

Table 3.14: KDE, n = 400

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	99.9	0.0	0.1	0.0	0.0	0.0
Laplace	0.0	99.9	0.1	0.0	0.0	0.0
$\alpha = 1.75$	0.0	0.2	98.1	1.7	0.0	0.0
$\alpha = 1.5$	0.0	0.0	1.1	98.1	0.8	0.0
$\alpha = 1.25$	0.0	0.0	0.0	1.0	98.9	0.1
$\alpha = 1$	0.0	0.0	0.0	0.0	0.5	99.5

Table 3.15: KDE, n = 800

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	100.0	0.0	0.0	0.0	0.0	0.0
Laplace	0.0	100.0	0.0	0.0	0.0	0.0
$\alpha = 1.75$	0.0	0.0	100.0	0.0	0.0	0.0
$\alpha = 1.5$	0.0	0.0	0.1	99.9	0.0	0.0
$\alpha = 1.25$	0.0	0.0	0.0	0.0	100.0	0.0
$\alpha = 1$	0.0	0.0	0.0	0.0	0.0	100.0

Table 3.16: KDE, n = 1600

Nearest Neighbour Density Estimation

For a domain $\mathcal{D}$ with dimension d, NNDE Sugiyama *et al.* (2012) uses hyperspheres with radius τ as the region $\mathcal{R}_{\mathcal{D}}$. The volume V of $\mathcal{R}_{\mathcal{D}}$ is given by

$$V = \frac{\pi^{d/2}\tau^d}{\Gamma\left(\frac{d}{2}+1\right)} \tag{3.11}$$

The PDF is then calculated using Equation (3.8).

Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified to the nearest SαS distribution. Table 3.17 to Table 3.20 give the results of classification in percentage.

Inferences

The KDE and NNDE based computation of the K-L divergence yields good results for the sample sizes considered. This performance does come with the penalty of reference

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	96.1	1.2	2.7	0.0	0.0	0.0
Laplace	0.1	96.0	2.7	1.1	0.1	0.0
$\alpha = 1.75$	5.0	11.1	72.5	11.3	0.1	0.0
$\alpha = 1.5$	0.2	6.9	11.9	72.0	9.0	0.0
$\alpha = 1.25$	0.0	0.9	0.0	12.9	79.7	6.5
$\alpha = 1$	0.0	0.0	0.0	0.0	10.4	89.6

Table 3.17: NNDE, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	99.3	0.3	0.4	0.0	0.0	0.0
Laplace	0.0	98.9	1.0	0.1	0.0	0.0
$\alpha = 1.75$	0.3	3.0	91.9	4.8	0.0	0.0
$\alpha = 1.5$	0.0	1.1	6.1	87.8	5.0	0.0
$\alpha = 1.25$	0.0	0.0	0.0	6.6	91.5	1.9
$\alpha = 1$	0.0	0.0	0.0	0.0	3.5	96.5

Table 3.18: NNDE, n = 400

PDF value generation which requires additional time and memory. Also, KDE requires kernel matrix evaluation.

3.2.3 Density Ratio Estimation

One way to use Equation (3.3) without going through the hassle of computing the individual densities is by computing the density ratio, $r(x)$ from the interference data points $\{x_j\}_{j=1}^{n}$ such that a convenient K-L divergence approximation is obtained. Equation (3.4) can be used to empirically evaluate the divergence metric with $\hat{r}(x)$ computed using a density ratio estimation algorithm Sugiyama *et al.* (2009, 2012).

Kullback-Leibler Importance Estimation Procedure (KLIEP)

The density ratio of $p(x)$ and $q(x)$ is given by $r(x) = \frac{p(x)}{q(x)}$. Supposing we have an estimate for the density ratio $\hat{r}(x)$, an estimate for $p(x)$ is obtained as:

$$\hat{p}(x) = \hat{r}(x) * q(x) \tag{3.12}$$

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	99.9	0.1	0.0	0.0	0.0	0.0
Laplace	0.0	100.0	0.0	0.0	0.0	0.0
$\alpha = 1.75$	0.1	0.1	98.4	1.4	0.0	0.0
$\alpha = 1.5$	0.0	0.0	0.5	99.0	0.5	0.0
$\alpha = 1.25$	0.0	0.0	0.0	0.7	99.0	0.3
$\alpha = 1$	0.0	0.0	0.0	0.0	0.5	99.5

Table 3.19: NNDE, n = 800

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	100.0	0.0	0.0	0.0	0.0	0.0
Laplace	0.0	100.0	0.0	0.0	0.0	0.0
$\alpha = 1.75$	0.0	0.0	100.0	0.0	0.0	0.0
$\alpha = 1.5$	0.0	0.0	0.1	99.7	0.2	0.0
$\alpha = 1.25$	0.0	0.0	0.0	0.2	99.7	0.1
$\alpha = 1$	0.0	0.0	0.0	0.0	0.0	100.0

Table 3.20: NNDE, n = 1600

Minimizing the K-L divergence from $\hat{p}(x)$ to $p(x)$ is one way of obtaining a good estimate $\hat{p}(x)$.

$$\begin{aligned} D_{KL}\left(p(x)||\hat{p}(x)\right) &= \int p(x)\log\left(\frac{p(x)}{\hat{p}(x)}\right)dx \\ &= \int p(x)\log\left(\frac{p(x)}{\hat{r}(x)*q(x)}\right)dx \\ &= \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx - \int p(x)\log\left(\hat{r}(x)\right)dx \\ &= D_{KL}\left(p(x)||q(x)\right) - \int p(x)\log\left(\hat{r}(x)\right)dx \\ &= C - \int p(x)\log\left(\hat{r}(x)\right)dx \end{aligned} \tag{3.13}$$

where C is a constant since the K-L divergence evaluates to the same constant independent of the estimators $\hat{p}(x)$ or $\hat{r}(x)$ chosen. The problem then becomes maximizing the term $\int p(x)\log\left(\hat{r}(x)\right)dx$. $\hat{r}(x)$ is then empirically approximated as follows:

$$\widehat{r}(x) = \underset{\hat{r}(x)}{\operatorname{argmax}} \frac{1}{n}\sum_{j=1}^{n}\log\left(\hat{r}(x_j)\right) \tag{3.14}$$

with the additional implicit constraints $\int \hat{p}(x)dx = 1$ and $r(x) \geq 0$ empirically imposed as:

$$\frac{1}{m}\sum_{j=1}^{m}\hat{r}(y_j) = 1 \tag{3.15}$$

and

$$\hat{r}(x) \geq 0 \quad \forall x \tag{3.16}$$

Linear $\left(\sum_{l=1}^{b}\theta_l\psi_l(x)\right)$ and kernel $\left(\sum_{l=1}^{b}\theta_l K(x, c_l)\right)$ models can be used to estimate $r(x)$ while implementing KLIEP.

GMM based Kullback-Leibler Importance Estimation Procedure

The optimization problem defining GM-KLIEP is the same as that of KLIEP. Here, the major difference is modelling the density ratio as a Gaussian mixture:

$$r(\mathbf{x}) = \sum_{\mathbf{k=1}}^{\mathbf{c}}\theta_{\mathbf{k}}\mathbf{N}\left(\mathbf{x}; \mu_{\mathbf{k}}, \mathbf{\Sigma_k}\right) \tag{3.17}$$

Since the noise data is assumed to be one dimensional, using GM-KLIEP is the same as KLIEP with Gaussian kernel.

Inferences

KLIEP is computationally intensive because of the kernel matrix evaluation, gradient ascent convergence time and cross validation for model selection. It is inconvenient for implementation in real time systems.

3.3 Classification Based on Pearson Divergence

K-L divergence minimization is equivalent to minimizing the log error. Similarly, minimizing the squared error yields the Pearson divergence. The Pearson divergence is

given by:

$$\begin{aligned} \mathrm{PE}(p||q) &= \frac{1}{2}\int \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x)dx \\ &= \frac{1}{2}\int r(x)p(x) - \int r(x)q(x) + \frac{1}{2} \end{aligned} \tag{3.18}$$

3.3.1 Density Ratio Estimation

Hence, the problem of calculating the Pearson divergence is equivalent to empirically computing Equation (3.18). Similar to what was discussed for K-L divergence, we can compute the density ratio empirically at convenient data points and use it for arriving at the empirical approximation Sugiyama *et al.* (2009, 2012).

Assuming a kernel model for r(x):

$$\begin{aligned} r(x) &= \alpha_0 + \sum_{l=1}^{b} \alpha_i K\left(\mathbf{x}, \mathbf{x_i}\right) \\ &= \alpha^{\mathbf{T}}\mathbf{k}(\mathbf{x}) \end{aligned} \tag{3.19}$$

Pearson divergence is evaluated empirically from a density ratio estimate $\hat{r}(x)$ as:

$$\hat{PE}(\{x_j\}_{j=1}^n, \{y_j\}_{j=1}^m) = \frac{1}{2n}\sum_{j=1}^{n}\hat{r}(x_i) - \frac{1}{m}\sum_{j=1}^{m}\hat{r}(y_i) + \frac{1}{2} \tag{3.20}$$

The Pearson divergence between the given sample distribution and the reference distributions is calculated and the given noise sample is mapped to the reference distribution to which it has minimum divergence. It is framed as a multiple hypothesis testing problem below:

- Let H_i be the hypothesis such that the noise sample is best modeled by the reference probability distribution $q_i(x)$.

$$\mathrm{H}_i : \{x_j\}_{j=1}^m \sim q_i(x) \tag{3.21}$$

- Calculate the divergences, $\mathrm{PE}\left(p||q_i\right)$ from $q_i(x)$ to $p(x)$
- Decision rule is

$$\delta = \underset{i}{\operatorname{argmin}}\ \left\{\mathrm{PE}\left(p||q_i\right)\right\} \tag{3.22}$$

where $\hat{r}(x)$ is computed using a density ratio estimation algorithm.

Least Squares Importance Fitting Sugiyama *et al.* (2009)

The formulation of least squares importance fitting (LSIF) is similar to that of KLIEP. The density ratio of $p(x)$ and $q(x)$ is given by $r(x) = \frac{p(x)}{q(x)}$. Supposing we have an estimate for the density ratio $\hat{r}(x)$, the squared error in the estimate is obtained as:

$$\begin{aligned} SE'\left(\hat{r}(x)\right) &= \frac{1}{2}\int \left(r(x) - \hat{r}(x)\right)^2 q(x)dx \\ &= \frac{1}{2}\int \hat{r}(x)^2 q(x)dx - \int \hat{r}(x)p(x)dx + \frac{1}{2}\int r(x)p(x)dx \end{aligned} \tag{3.23}$$

Here, the last term is a constant independent of our choice of $\hat{r}(x)$ and can be ignored. Empirically approximating the remaining terms, we get

$$SE\left(\hat{r}(x)\right) = \frac{1}{2m}\sum_{j=1}^{m} \hat{r}(y_j)^2 - \frac{1}{n}\sum_{j=1}^{n} \hat{r}(x_j) \tag{3.24}$$

The optimization problem which yields the density ratio estimate is now:

$$\widehat{r}(x) = \underset{\hat{r}(x)}{\operatorname{argmin}}\, SE\left(\hat{r}(x)\right) \tag{3.25}$$

Additional constraints of non-negativity $\hat{r}(x)$ is also imposed empirically. $r(x)$ is modeled as linear - $\left(\sum_{l=1}^{b} \theta_l \psi_l(x) = \psi(x)^T\theta\right)$ - for estimation purposes and θ is then estimated.

Unconstrained Least Squares Importance Fitting Sugiyama *et al.* (2009)

Unconstrained least squares importance fitting (uLSIF) is LSIF without non-negativity constraint for the density ratio. The negative $\hat{r}(x)$ values are rounded to zero by having a $\hat{\theta_k} = \min\left(\theta_k, 0_b\right)$ step in θ update.

3.3.2 Least Squares Two-samples Test

The least squares two-samples test (LSTT) is a homogeneity test making use of empirical Pearson divergence estimates Sugiyama *et al.* (2011) based on uLSIF algo-

rithm. Consider the input sample set $\mathcal{X} = \{x_j\}_{j=1}^{n} \sim p(x)$ and reference sample set $\mathcal{Y} = \{y_j\}_{j=1}^{m} \sim q(x)$ with $m = n$. Denoting the distribution of $\widehat{PE}(\mathcal{X}, \mathcal{Y})$ as F, let

$$\beta = \sup \{t \in \mathbb{R} | F(t) \leq 1 - \alpha\} \tag{3.26}$$

be the upper 100α-percentile point of F. If $p(x) = q(x)$, we have

$$P\left(\widehat{PE}(\mathcal{X}, \mathcal{Y}) > \beta\right) \leq \alpha \tag{3.27}$$

The two-samples test is based on the permutation test Efron and Tibshirani (1994). The test procedure is as follows:

- The hypotheses of the test are:
  H_0 : $\mathcal{X}$ and $\mathcal{Y}$ are from populations with same distribution. i.e. $p(x) \equiv q(x)$
  H_1 : $\mathcal{X}$ and $\mathcal{Y}$ are not from populations with same distribution. i.e. $p(x) \not\equiv q(x)$

- Calculate the Pearson divergence estimate for the original datasets $\mathcal{X}$ and $\mathcal{Y}$.

$$PE_0 = \widehat{PE}(\mathcal{X}, \mathcal{Y}) \tag{3.28}$$

- Randomly permute the $|\mathcal{X} \cup \mathcal{Y}|$ samples, assign first $|\mathcal{X}|$ samples to $\mathcal{X}_i$ and the rest $|\mathcal{Y}|$ samples to $\mathcal{Y}_i$. Calculate the divergence between the new sets.

$$PE_i = \widehat{PE}(\mathcal{X}_i, \mathcal{Y}_i) \tag{3.29}$$

- Repeat random shuffling and divergence calculation many number of times (T) to construct a distribution of $\widehat{PE}$ under the null hypothesis.
- Approximate the p-value as the relative ranking of PE_0 among $\{PE_i\}_{i=1}^{T}$

$$\widehat{p-\text{value}} = \frac{1}{T} \sum_{i=1}^{T} I(PE_i > PE_0) \tag{3.30}$$

  where $I(\text{true}) = 1$ and $I(\text{false}) = 0$
- Since Pearson divergence is not symmetric, the divergence in the opposite order is also calculated and the procedure is carried out. The final p-value is determined as the minimum of the p-values obtained from both procedures.

Inferences

As demonstrated in Sugiyama *et al.* (2011), the LSTT works well as a homogeneity test. But extending the test to map a given data to a set of multiple reference distributions by

comparing the p-values of pairwise tests with each reference sample does not work. The reason is that the procedure uses naive bootstrap which is ill-suited for handling heavy-tailed data. Additionally, the execution time is often long due to the uLSIF algorithm, which comprises kernel matrix evaluation, matrix inversion and cross validation steps.

3.4 Classification Based on Integral Probability Metric

Using the K-L divergence and Pearson divergence as metrics for classification required estimating the individual densities or the density ratio. Maximum mean discrepancy (MMD) is a statistical distance which maximizes the difference expected values of a function defined on a universal reproducing kernel Hilbert space (RKHS). The key idea is to use empirically calculated mean values instead of estimating the density to measure the distance.

3.4.1 Maximum Mean Discrepancy

Maximum mean discrepancy Sugiyama *et al.* (2011) is an integral probability metric between two distributions $p(x)$ and $q(x)$ defined on a function class $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}$ given by

$$\mathrm{MMD}\left(\mathcal{H}, p(x), q(x)\right) = \sup_{f \in \mathcal{H}} \left[\int f(x)p(x)dx - \int f(x)q(x)dx \right] \tag{3.31}$$

If the function class $\mathcal{H}$ is a unit ball in a universal reproducing kernel Hilbert space defined on a compact metric space, then the MMD vanishes if and only if $p(x) \equiv q(x)$. Gaussian kernels are universal RKHSs. Reproducing property of the kernel allows us to obtain the value of a function at a point x provided we know the inner product of the function and the kernel centred at x.

$$f(x) = \langle f(.), K\left(x, .\right) \rangle_{\mathcal{H}} \tag{3.32}$$

Using the reproducing property of the RKHS, we can arrive upon an empirical approximation of MMD^2. If we have $\mathcal{X} = \{x_j\}_{j=1}^n \sim p(x)$ and $\mathcal{Y} = \{y_j\}_{j=1}^m \sim q(x)$, then

$$\widehat{\mathrm{MMD}^2}\left(\mathcal{H}, p(x), q(x)\right) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}K\left(x_j, x_k\right) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}K\left(y_j, y_k\right) - \frac{2}{nm}\sum_{j=1}^{n}\sum_{k=1}^{m}K\left(x_j, y_k\right) \quad (3.33)$$

Extending to Multiple Reference Distributions

The metric $\widehat{\mathrm{MMD}^2}\left(\mathcal{H}, p(x), q_i(x)\right)$ are calculated for all reference samples with $q_i(x)$ and $p(x)$ is mapped to the $q_i(x)$ which yields the smallest $\widehat{\mathrm{MMD}}^2$ estimate.

Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified to nearest SαS distribution. Table 3.21 to Table 3.24 gives the classification accuracy of each class in percentage.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	30.4	7.7	30.4	21.7	8.4	1.4
Laplace	6.7	21.2	12.3	17.7	22.4	19.7
$\alpha = 1.75$	21.2	12.1	23.1	21.7	16.3	5.6
$\alpha = 1.5$	10.3	11.3	19.8	21.6	21.6	15.4
$\alpha = 1.25$	2.5	3.8	11.1	19.3	33.6	29.7
$\alpha = 1$	0.2	2.2	5.4	13.7	28.2	50.3

Table 3.21: MMD, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	39.4	2.1	41.4	16.4	0.7	0.0
Laplace	4.6	34.4	10.5	20.7	19.0	10.8
$\alpha = 1.75$	23.8	10.6	29.5	25.9	7.7	2.5
$\alpha = 1.5$	7.4	8.0	18.7	30.7	24.1	11.1
$\alpha = 1.25$	0.3	1.8	6.6	23.4	40.6	27.3
$\alpha = 1$	0.0	1.6	1.0	8.3	32.7	56.4

Table 3.22: MMD, n = 400

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	56.5	0.0	38.7	4.8	0.0	0.0
Laplace	0.7	63.2	3.9	14.7	12.3	5.2
$\alpha = 1.75$	27.8	4.0	41.5	24.4	2.3	0.0
$\alpha = 1.5$	2.5	3.1	23.6	43.0	22.9	4.9
$\alpha = 1.25$	0.0	0.9	1.6	20.1	52.5	24.9
$\alpha = 1$	0.0	0.6	0.0	3.8	23.6	72.0

Table 3.23: MMD, n = 800

$p(x)$ \ $q_i(x)$	$\alpha = 2$	Laplace	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	67.4	0.0	32.3	0.3	0.0	0.0
Laplace	0.0	91.3	0.7	4.5	3.3	0.2
$\alpha = 1.75$	26.8	1.4	55.1	16.4	0.3	0.0
$\alpha = 1.5$	0.0	0.3	16.3	69.6	13.4	0.4
$\alpha = 1.25$	0.0	0.2	0.0	12.1	68.5	19.2
$\alpha = 1$	0.0	0.6	0.0	0.2	19.0	80.2

Table 3.24: MMD, n = 1600

Inference

MMD classification results improve as the sample size increases. MMD involves multiple kernel matrix evaluations which increase the time complexity. The calculations take a significant amount of time which makes this test ill suited for implementing in a real time system.

3.4.2 Summary of Classification Methods

In the classification methods discussed so far, the classification based on K-L divergence using KDE and NNDE performs the best. They have the upper hand due to the reference PDF being used rather than reference data as in other methods. The empirical methods proposed by Wang *et al* are computationally faster but gives less classification accuracy. Density ratio estimation and MMD are computationally expensive. The LSTT and K-L divergence based on FPC algorithm gives poor classification results. For low complexity with good classification results, a different approach is required.

CHAPTER 4

SαS MODEL SPECIFIC ESTIMATION AND CLASSIFICATION

One of the challenges in selecting the appropriate SαS model for $p(x)$ is accurately identifying Gaussian data from near-Gaussian data. From Monte-Carlo simulations to classify Gaussian ($\alpha = 2$) and near-Gaussian ($\alpha = 1.75$) data, a tendency to classify near-Gaussian data as Gaussian was observed. An initial test which can segregate the Gaussian data from the other stable models may then increase the classification performance. A 2-stage classification approach using order statistics and fractional lower order moments is presented here. A scale parameter estimator with better performance than existing low complexity estimator is also proposed.

4.1 Identifying Gaussian data from Near-Gaussian data

Various normality tests are available in literature which are capable of distinguishing between Gaussian and non-Gaussian data but not necessarily Near-Gaussian data. Simple normality tests like G-Kurtosis test, Jarque-Bera test and empirical CDF based normality tests like one-sample Kolmogorov-Smirnov Test and Anderson-Darling test were used to classify the data as Gaussian or not by using different sample sizes. For standard threshold values, all tests classified SαS,$\alpha = 1.75$ distributed data wrongly as Gaussian ($\alpha = 2$).

Since we are more concerned with the shape of the probability distribution here, L-moments - quantities analogous to conventional moments but based on order statistics - can be used. It has already been shown that L-moments and their ratios can be used to characterize the PDF Hosking (2006). L-moments are calculated from order statistics of the data as follows:

$$\lambda_r \equiv r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}\left[X_{r-k:r}\right] \tag{4.1}$$

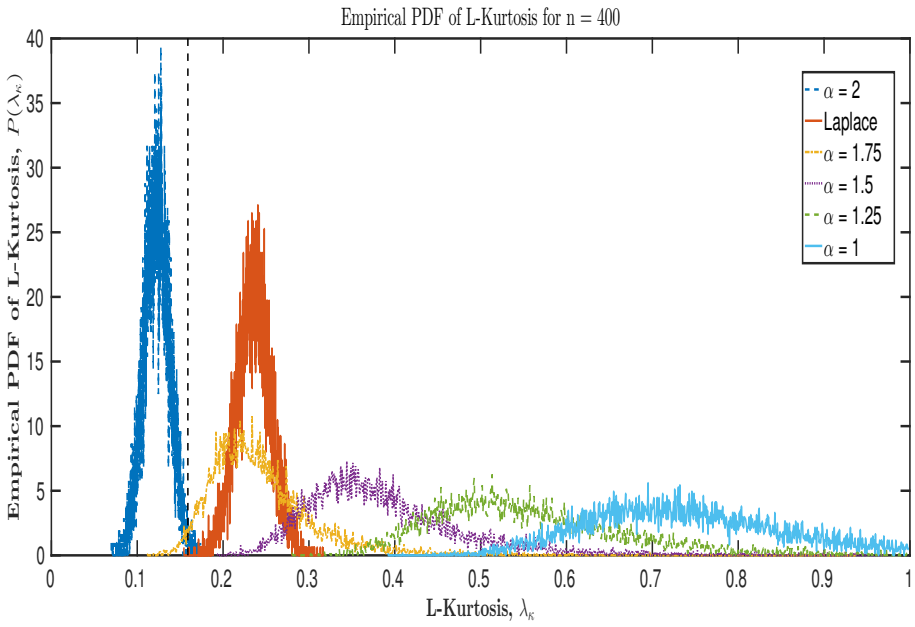

Figure 4.1: Empirical PDF of λ_κ for sample size, n = 400 and scale $\gamma = 100$

The dotted line represent the threshold($\tau = 0.159$) for the test

where $X_{k:n}$ is the k^{th} order statistic of a sample of X with size n. The L-Kurtosis λ_κ is defined as:

$$\lambda_\kappa = \frac{\lambda_4}{\lambda_2} \tag{4.2}$$

The direct sample based estimators Hosking (1990) can be used to get estimates for λ_2 and λ_4.

The threshold to which the test statistic should be compared is obtained empirically by fixing the probability of wrong classification and running Monte-Carlo simulations to find the threshold that achieves that probability. The threshold chosen here, $\tau = 0.159$, is obtained by fixing the probability of wrong classification at 0.005 for a sample size of $n = 800$. Figure 4.1 and Figure 4.2 show the empirical PDFs of the L-Kurtosis metric for 2 different sample sizes. Algorithm 1 explains the test procedure.

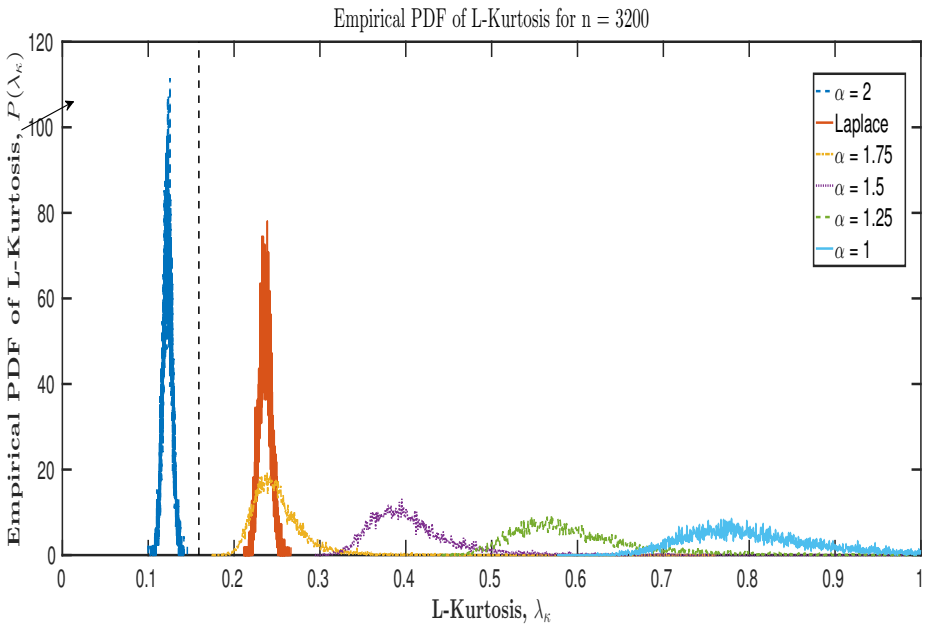

Figure 4.2: Empirical PDF of λ_κ for sample size, n = 3200 and scale $\gamma = 100$

The dotted line represent the threshold($\tau = 0.159$) for the test

---

Algorithm 1: Classification of $p(x)$ using L-Kurtosis

Data: Samples $\{x_j\}_{j=1}^n \sim p(x)$

Result: Gaussian or non-Gaussian (Heavier-tailed)

1 Sort $\{x_j\}_{j=1}^n$ and obtain the order statistics $X_{i:n}$ from the sample

2 Calculate $\hat{\lambda}_2$ and $\hat{\lambda}_4$

$$\hat{\lambda}_2 = \frac{1}{2\binom{n}{2}} \sum_{i=1}^{n} \left(\binom{i-1}{1} - \binom{n-i}{1} \right) X_{i:n}$$

$$\hat{\lambda}_4 = \frac{1}{4\binom{n}{4}} \sum_{i=1}^{n} \left(\binom{i-1}{3} - 3\binom{i-1}{2}\binom{n-i}{1} + 3\binom{i-1}{1}\binom{n-i}{2} - \binom{n-i}{3} \right) X_{i:n}$$

3 Calculate L-Kurtosis as $\hat{\lambda}_\kappa = \frac{\hat{\lambda}_4}{\hat{\lambda}_2}$

4 Compare with $\hat{\lambda}_\kappa$ threshold:

if $\lambda_\kappa < \tau$ **then**

$\hat{p}(x) \equiv$ Gaussian

else

$\hat{p}(x) \equiv$ non-Gaussian (Heavier-tailed)

4.1.1 Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 1000 iterations and classified. L-kurtosis is calculated from the order statistics and compared against the threshold τ as calculated earlier, to classify the samples as Gaussian/Non-Gaussian. Table 4.1 to Table 4.4 show the results of classification in percentage for sample sizes from 200 to 1600.

$p(x)$ \ $q_i(x)$	Gaussian	Non-Gaussian
$\alpha = 2$	98.3	1.7
Laplace	0.3	99.7
$\alpha = 1.75$	12.4	87.6
$\alpha = 1.5$	0.2	99.8
$\alpha = 1.25$	0.0	100.0
$\alpha = 1$	0.0	100.0

Table 4.1: Algorithm 1, n = 200

$p(x)$ \ $q_i(x)$	Gaussian	Non-Gaussian
$\alpha = 2$	99.8	0.2
Laplace	0.1	99.9
$\alpha = 1.75$	3.7	96.3
$\alpha = 1.5$	0.0	100.0
$\alpha = 1.25$	0.0	100.0
$\alpha = 1$	0.0	100.0

Table 4.2: Algorithm 1, n = 400

$p(x)$ \ $q_i(x)$	Gaussian	Non-Gaussian
$\alpha = 2$	100.0	0.0
Laplace	0.0	100.0
$\alpha = 1.75$	0.5	99.5
$\alpha = 1.5$	0.0	100.0
$\alpha = 1.25$	0.0	100.0
$\alpha = 1$	0.0	100.0

Table 4.3: Algorithm 1, n = 800

$p(x)$ \ $q_i(x)$	Gaussian	Non-Gaussian
$\alpha = 2$	100.0	0.0
Laplace	0.0	100.0
$\alpha = 1.75$	0.0	100.0
$\alpha = 1.5$	0.0	100.0
$\alpha = 1.25$	0.0	100.0
$\alpha = 1$	0.0	100.0

Table 4.4: Algorithm 1, n = 1600

4.2 Estimating Scale Parameter from Fractional Lower Order Moments

Fractional Lower Order Moments of a Symmetric α-Stable variable with zero location parameter and scale parameter γ is given by Zolotarev (1986) as

$$\text{FLOM}(p) = \mathbb{E}\left(|X|^p\right) = \begin{cases} C\left(p,\alpha\right)\gamma^p & 0 < p < \alpha \\ \text{Doesn't exist} & p \geq \alpha \end{cases} \tag{4.3}$$

where

$$C\left(p,\alpha\right) = \frac{2^{p+1}\Gamma\left(\frac{p+1}{2}\right)\Gamma\left(\frac{-p}{\alpha}\right)}{\alpha\sqrt{\pi}\Gamma\left(\frac{-p}{2}\right)} \tag{4.4}$$

From Equation (4.3), for a given α and p with $0 < p < \alpha$, $\mathbb{E}\left(|X|^p\right)$ is solely a function of γ. A FLOM based estimator for γ - provided α is known - can than be formulated from the above relations. For fixed α and p,

$$\hat{\gamma^p} = \frac{\text{FLOM}(p)}{C\left(\alpha,p\right)} \tag{4.5}$$

It is empirically observed that calculating γ based on above relation for a range of p values and taking the mean gives a reasonable estimate of γ. For a suitable selection of $\mathcal{P} = \{p_k\}_{k=1}^{n_p}$, this method outperforms the popular scale parameter estimator

McCulloch (1986) in terms of bias and Mean Square Error performance.

Algorithm 2: Estimate scale parameter γ from Fractional Lower Order Moments

Data: Samples $\{x_j\}_{j=1}^{n}$, fractional powers $\mathcal{P} = \{p_k\}_{k=1}^{n_p}$, alpha value α

Result: Estimate of scale parameter, $\hat{\gamma}$

Init: $\hat{\gamma} = 0$

foreach *k* **do**

$$\mathrm{FLOM}\left(p_k\right) = \frac{1}{N}\sum_{j=1}^{n}|x_j|^{p_k}$$

$$C\left(\alpha, p_k\right) = \frac{2^{p_k+1}\Gamma\left(\frac{p_k+1}{2}\right)\Gamma\left(\frac{-p_k}{\alpha}\right)}{\alpha\sqrt{\pi}\Gamma\left(\frac{-p_k}{2}\right)}$$

$$\hat{\gamma} \leftarrow \hat{\gamma} + \frac{1}{n_p}\left(\frac{FLOM(k)}{C_{i,k}}\right)^{\frac{1}{p_k}}$$

4.3 Estimating Characteristic Exponent from Fractional Lower Order Moments from a Discrete Set of α

Building upon the scale parameter estimate specified in the previous section, a procedure for selecting the most apt characteristic exponent α from a set of alpha values $\mathcal{A} = \{\alpha_i\}_{i=1}^{M}$ for the given data is proposed by minimizing the mean square error over p across α in estimating $\hat{\gamma}$. The error function $J(\alpha)$ thus proposed, when empirically calculated and plotted, exhibits a minima at the α value closest to that of the actual PDF of the data. Figure 4.3 to Figure 4.6 show the convex nature of $J(\alpha)$.

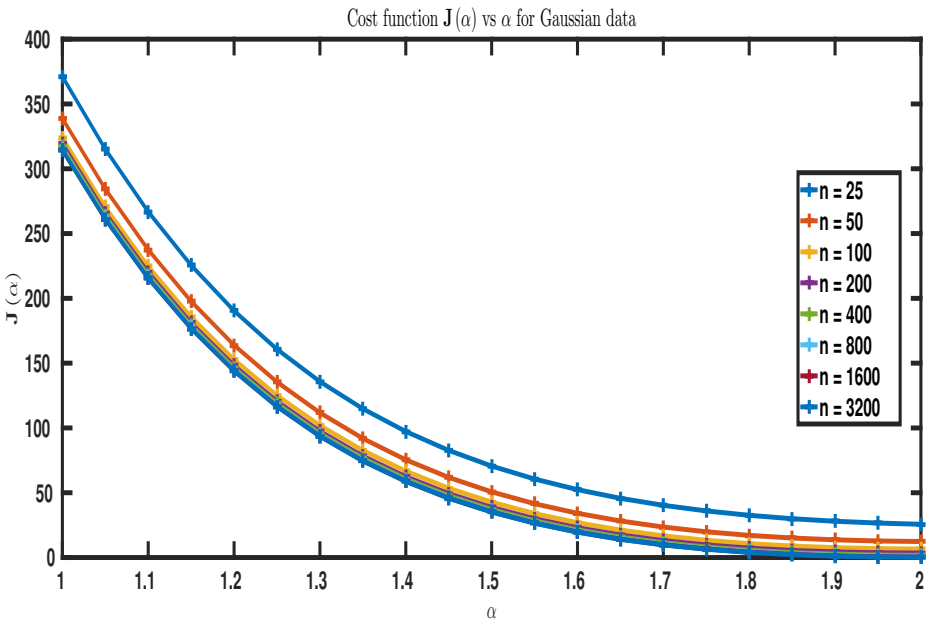

Figure 4.3: Cost function $J(\alpha)$ vs α for Gaussian data

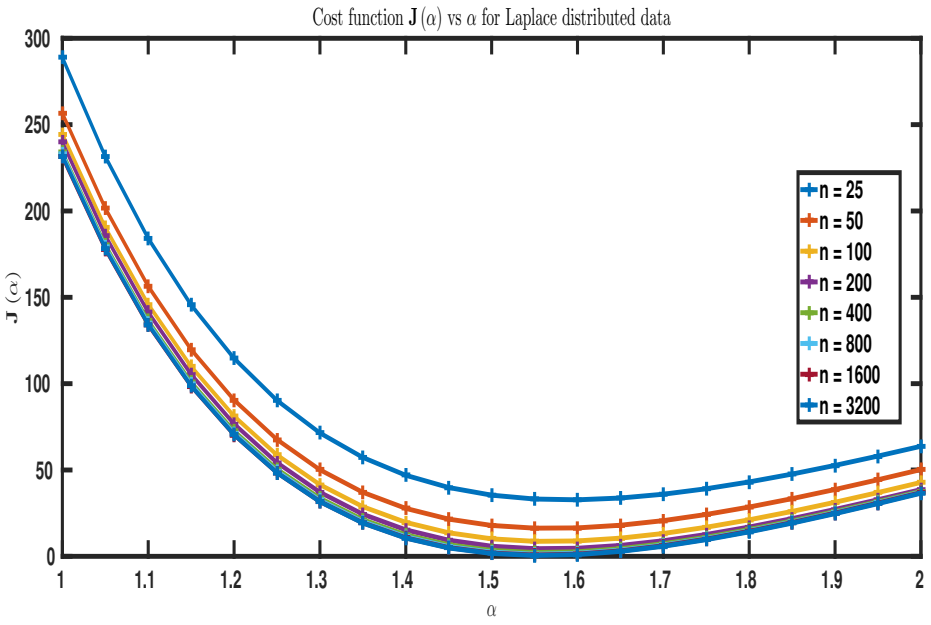

Figure 4.4: Cost function $J(\alpha)$ vs α for Laplace data

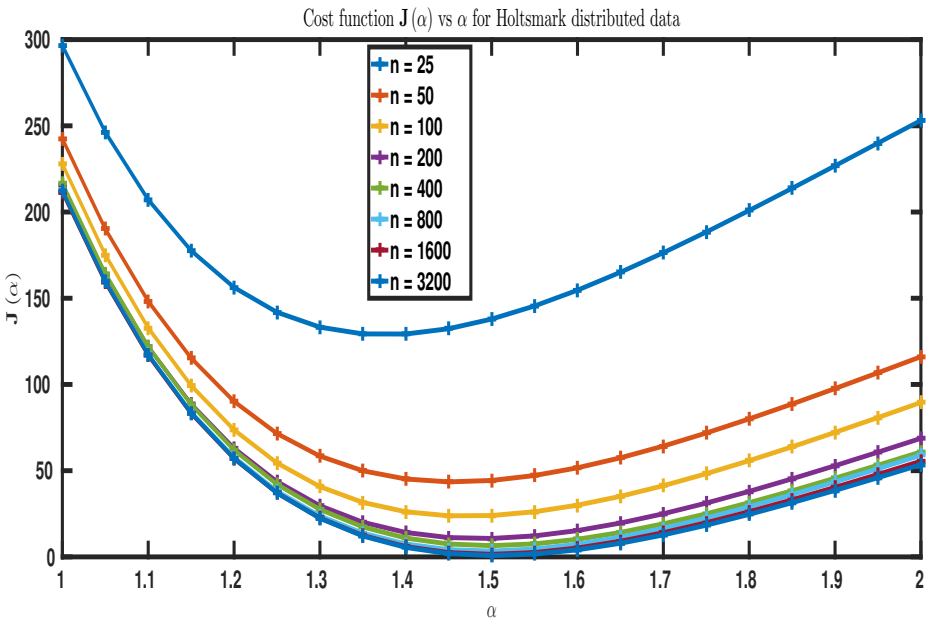

Figure 4.5: Cost function $J(\alpha)$ vs α for Holtsmark data

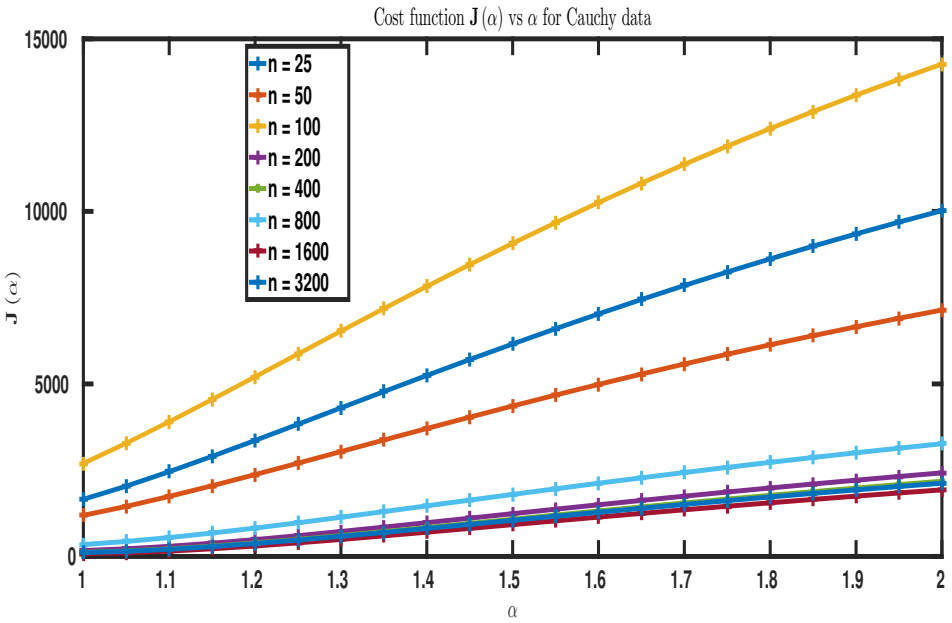

Figure 4.6: Cost function $J(\alpha)$ vs α for Cauchy data

Algorithm 3: Classify $p(x)$ based on Fractional Lower Order Moments

Data: Samples $\{x_j\}_{j=1}^{n}$, fractional powers $\mathcal{P} = \{p_k\}_{k=1}^{n_p}$, alpha values $\mathcal{A} = \{\alpha_i\}_{i=1}^{M}$, $q_i(x) = S_{\alpha_i}(0, \gamma, 0)$ $\quad \max_k p_k < \min_i \alpha_i$

Result: $\hat{p}\,(x)$

foreach *k* **do**

$\mathrm{FLOM}\,(k) = \frac{1}{N}\sum_{j=1}^{n} |x_j|^{p_k}$

foreach *i* **do**

$C_{i,k} = \frac{2^{p_k+1}\Gamma\left(\frac{p_k+1}{2}\right)\Gamma\left(\frac{-p_k}{\alpha_i}\right)}{\alpha_i\sqrt{\pi}\Gamma\left(\frac{-p_k}{2}\right)}$

$\hat{\gamma}_{i,k} = \left(\frac{FLOM(k)}{C_{i,k}}\right)^{\frac{1}{p_k}}$

foreach *i* **do**

$J\,(\alpha_i) = Var\left(\{\hat{\gamma}_{i,k}\}_{k=1}^{n_p}\right)$

$\hat{i} = \underset{i}{\operatorname{argmin}}\, J\,(\alpha_i)$

$\hat{p}(x) \equiv q_{\hat{i}}\,(x)$

4.3.1 Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 10000 iterations and classified. Gamma estimates are calculated from the data samples with assumption on α, and the α value yielding the lowest error is chosen. Table 4.5 to Table 4.8 show the results of classification in percentage for sample sizes from 200 to 1600.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	78.4	20.7	0.9	0.0	0.0
Laplace	1.5	32.5	64.2	1.7	0.0
$\alpha = 1.75$	31.4	52.2	15.8	0.6	0.0
$\alpha = 1.5$	2.5	25.9	58.9	12.3	0.3
$\alpha = 1.25$	0.0	0.9	26.6	63.5	9.0
$\alpha = 1$	0.0	0.0	0.2	24.2	75.6

Table 4.5: Algorithm 3, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	85.0	15.0	0.0	0.0	0.0
Laplace	0.1	27.2	72.5	0.1	0.0
$\alpha = 1.75$	22.0	67.0	10.8	0.2	0.0
$\alpha = 1.5$	0.3	19.1	73.1	7.4	0.1
$\alpha = 1.25$	0.0	0.0	16.1	78.0	5.9
$\alpha = 1$	0.0	0.0	0.0	13.4	86.6

Table 4.6: Algorithm 3, n = 400

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	92.7	7.3	0.0	0.0	0.0
Laplace	0.0	19.6	80.4	0.0	0.0
$\alpha = 1.75$	11.7	83.5	4.8	0.0	0.0
$\alpha = 1.5$	0.0	7.7	88.1	4.2	0.0
$\alpha = 1.25$	0.0	0.0	6.4	90.7	2.9
$\alpha = 1$	0.0	0.0	0.0	5.5	94.5

Table 4.7: Algorithm 3, n = 800

4.4 Estimating Characteristic Exponent from Fractional Lower Order Moments from a Given Range of α

If we have a continuous range of α instead of a discrete set $\mathcal{A} = \{\alpha_i\}_{i=1}^{M}$, a minimization algorithm[1] can be used on $J(\alpha)$ to get the estimate of the characteristic exponent. The resulting estimator operation is similar to Algorithm 3 except that the error function values are calculated on successive iterations rather than being precomputed.

4.4.1 Simulation Results

SαS distributed data of different sample sizes are generated over 10000 iterations and SαS parameter estimates are obtained. Figure 4.7 gives the Mean Square Error (MSE) performance of the estimator for different SαS input. The performance is contrasted with existing FLOM based estimator Kuruoglu (2001) for α.

[1] fminbnd() in MATLAB uses golden section search and parabolic interpolation

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	97.9	2.1	0.0	0.0	0.0
Laplace	0.0	9.1	90.9	0.0	0.0
$\alpha = 1.75$	4.8	93.6	1.6	0.0	0.0
$\alpha = 1.5$	0.0	2.8	95.8	1.3	0.0
$\alpha = 1.25$	0.0	0.0	1.4	97.1	1.5
$\alpha = 1$	0.0	0.0	0.0	0.8	99.2

Table 4.8: Algorithm 3, n = 1600

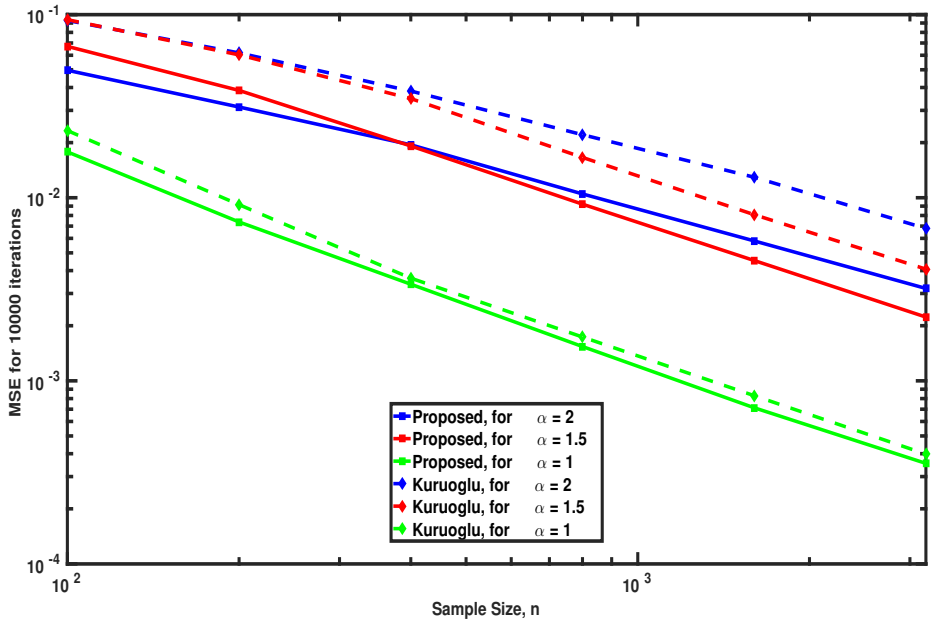

Figure 4.7: Proposed $\hat{\alpha}$ estimator performance for different SαS data

4.5 SαS Model Selection

Using Algorithm 1 to do initial classification of data into Gaussian or Non-Gaussian, followed by Algorithm 3 to identify the nearest characteristic exponent yields better classification results.

4.5.1 Simulation Results

SαS and Laplace distributed data of different sample sizes are generated over 10000 iterations and classified to nearest SαS distribution. Table 4.9 to Table 4.12 give the results of classification in percentage.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	96.6	3.2	0.2	0.0	0.0
Laplace	0.5	34.4	63.4	1.7	0.0
$\alpha = 1.75$	9.9	72.4	17.2	0.4	0.0
$\alpha = 1.5$	0.1	28.4	58.7	12.4	0.5
$\alpha = 1.25$	0.0	1.1	26.5	62.9	9.5
$\alpha = 1$	0.0	0.0	0.2	24.9	74.9

Table 4.9: SαS model selection, n = 200

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	99.5	0.5	0.0	0.0	0.0
Laplace	0.0	26.8	73.1	0.1	0.0
$\alpha = 1.75$	2.5	86.4	10.9	0.2	0.0
$\alpha = 1.5$	0.0	19.4	72.6	7.9	0.1
$\alpha = 1.25$	0.0	0.0	15.6	78.2	6.2
$\alpha = 1$	0.0	0.0	0.0	13.7	86.3

Table 4.10: SαS model selection, n = 400

4.5.2 Inferences

The simulation results for Algorithm 1 shows that it is capable of minimizing the occurrences of $\alpha = 1.75$ from classified as $\alpha = 2$. Algorithm 3 simulation results show that there is a significant amount of cases where near-Gaussian is classified as Gaussian. The continuous version of the classifier has better MSE performance than the existing FLOM based estimator for α. By making use of the two-stage classification, the classification accuracy improves and is found to be second only to the K-L divergence based classification using non-parametric methods.

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	100.0	0.0	0.0	0.0	0.0
Laplace	0.0	17.5	82.5	0.0	0.0
$\alpha = 1.75$	0.2	94.7	5.1	0.0	0.0
$\alpha = 1.5$	0.0	9.7	86.2	4.0	0.0
$\alpha = 1.25$	0.0	0.0	7.1	89.5	3.4
$\alpha = 1$	0.0	0.0	0.0	4.9	95.1

Table 4.11: SαS model selection, n = 800

$p(x)$ \ $q_i(x)$	$\alpha = 2$	$\alpha = 1.75$	$\alpha = 1.5$	$\alpha = 1.25$	$\alpha = 1$
$\alpha = 2$	100.0	0.0	0.0	0.0	0.0
Laplace	0.0	9.3	90.7	0.0	0.0
$\alpha = 1.75$	0.0	98.4	1.6	0.0	0.0
$\alpha = 1.5$	0.0	3.0	95.9	1.1	0.0
$\alpha = 1.25$	0.0	0.0	1.5	97.2	1.3
$\alpha = 1$	0.0	0.0	0.0	0.9	99.1

Table 4.12: SαS model selection, n = 1600

CHAPTER 5

CONCLUSIONS AND FUTURE SCOPE

5.1 Conclusions

In this project, methods of classifying data with unknown characteristics to a set of known distributions were explored and contrasted. As one would expect, among distance based classifiers, the Kernel Density Estimation and Nearest Neighbour Density Estimation based methods which use reference probability distribution functions, outperformed the methods using reference data. But, both have the computational overhead of generating the reference pdf values. Moving on from a generic model to a SαS model, FLOM based scale and characteristic exponent estimators were proposed. Also, a 2-stage classification method was introduced. The initial stage served as a separator for Gaussian from non-Gaussian and the final stage mapped the suitable symmetric α-stable model to the data.

5.2 Future Scope

Focussing on speeding up the classification process, with significant accuracy for smaller samples sizes, will help in realizing algorithms which can cater to the requirements of real life systems. Such an algorithm can raise the performance of the system as the noise modelling will be more accurate. Also, existence of a reliable classifier will help in faster adoption of the flexible α-stable models for industrial design.

REFERENCES

1. **Arce, G. R.**, *Nonlinear signal processing : a statistical approach*. Wiley-Interscience, 2005. ISBN 0471691844,9780471691846.

2. **Briassouli, A.**, **P. Tsakalides**, and **A. Stouraitis** (2005). Hidden messages in heavy-tails: Dct-domain watermark detection using alpha-stable models. *IEEE Transactions on Multimedia*, **7**(4), 700–715. ISSN 1520-9210.

3. **Efron, B.** and **R. J. Tibshirani**, *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1994. ISBN 0412042312.

4. **Georgiou, P. G.**, **P. Tsakalides**, and **C. Kyriakakis** (1999). Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE Transactions on Multimedia*, **1**(3), 291–301. ISSN 1520-9210.

5. **Gnedenko, B.** and **A. Kolmogorov**, *Limit distributions for sums of independent random variables*. Addison-Wesley series in statistics. Addison-Wesley, 1968. URL `https://books.google.co.in/books?id=rYsZAQAAIAAJ`.

6. **Gulati, K.**, **B. L. Evans**, **J. G. Andrews**, and **K. R. Tinsley** (2010). Statistics of co-channel interference in a field of poisson and poisson-poisson clustered interferers. *IEEE Transactions on Signal Processing*, **58**(12), 6207–6222. ISSN 1053-587X.

7. **Hosking, J.** (2006). On the characterization of distributions by their l-moments. *Journal of Statistical Planning and Inference*, **136**(1), 193 – 198. ISSN 0378-3758. URL `http://www.sciencedirect.com/science/article/pii/S0378375804002514`.

8. **Hosking, J. R. M.** (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, **52**(1), 105–124. ISSN 00359246. URL `http://www.jstor.org/stable/2345653`.

9. **Ilow, J.** and **D. Hatzinakos** (1998). Analytic alpha-stable noise modeling in a poisson field of interferers or scatterers. *IEEE Transactions on Signal Processing*, **46**(6), 1601–1611. ISSN 1053-587X.

10. **Kapur, J. N.** and **H. K. Kesavan**, *Entropy Optimization Principles and Their Applications*. Springer Netherlands, Dordrecht, 1992. ISBN 978-94-011-2430-0, 3–20. URL `http://dx.doi.org/10.1007/978-94-011-2430-0_1`.

11. **Kogon, S. M.** and **D. B. Williams**, A practical guide to heavy tails. chapter Characteristic Function Based Estimation of Stable Distribution Parameters. Birkhauser Boston Inc., Cambridge, MA, USA, 1998. ISBN 0-8176-3951-9, 311–335. URL `http://dl.acm.org/citation.cfm?id=292595.292617`.

12. **Kuruoglu, E.** (2001). Density parameter estimation of skewed alpha-stable distributions. *Trans. Sig. Proc.*, **49**(10), 2192–2201. ISSN 1053-587X. URL http://dx.doi.org/10.1109/78.950775.

13. **McCulloch, J. H.** (1986). Simple consistent estimators of stable distribution parameters. *Communications in Statistics: Simulation and Computation*, **15**, 1109–1136.

14. **Niranjayan, S.** and **N. C. Beaulieu**, A myriad filter detector for uwb multiuser communication. *In 2008 IEEE International Conference on Communications*. 2008. ISSN 1550-3607.

15. **Parzen, E.** (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**(3), 1065–1076. ISSN 00034851. URL http://www.jstor.org/stable/2237880.

16. **Perez-Cruz, F.**, Kullback-leibler divergence estimation of continuous distributions. *In 2008 IEEE International Symposium on Information Theory*. 2008. ISSN 2157-8095.

17. **Sugiyama, M.**, **T. Kanamori**, **T. Suzuki**, **S. Hido**, **J. Sese**, **I. Takeuchi**, and **L. Wang** (2009). A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, **1**, 183–208.

18. **Sugiyama, M.**, **T. Suzuki**, **Y. Itoh**, **T. Kanamori**, and **M. Kimura** (2011). Least-squares two-sample test. *Neural Networks*, **24**(7), 735 – 751. ISSN 0893-6080. URL http://www.sciencedirect.com/science/article/pii/S0893608011001262.

19. **Sugiyama, M.**, **T. Suzuki**, and **T. Kanamori**, *Density Ratio Estimation in Machine Learning*. Cambridge University Press, New York, NY, USA, 2012, 1st edition. ISBN 0521190177, 9780521190176.

20. **Wang, Q.**, **S. R. Kulkarni**, and **S. Verdu** (2005). Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, **51**(9), 3064–3074. ISSN 0018-9448.

21. **Win, M. Z.**, **P. C. Pinto**, and **L. A. Shepp** (2009). A mathematical theory of network interference and its applications. *Proceedings of the IEEE*, **97**(2), 205–230. ISSN 0018-9219.

22. **Zolotarev, V.**, *One-dimensional Stable Distributions*. Translations of mathematical monographs. American Mathematical Society, 1986. ISBN 9780821898154. URL https://books.google.co.in/books?id=ydwt9SotnN0C.