UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Contemplating Statistics: estimation and regression according to arc lengths

by

Mattheüs Theodor Loots

Submitted in partial fulfilment of the requirements for the degree
Philosophiae Doctor (Mathematical Statistics)
in the Faculty of Natural and Agricultural Sciences
University of Pretoria, Pretoria

September 2017

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Publication data:

Mattheüs Theodor Loots. Contemplating Statistics: estimation and regression according to arc lengths. Doctoral thesis, University of Pretoria, Department of Statistics, Pretoria, South Africa, September 2017.

Electronic, hyperlinked versions of this thesis are available online, as Adobe PDF files, at:

http://upetd.up.ac.za/UPeTD.htm

© University of Pretoria

Contemplating Statistics: estimation and regression according to arc lengths

by

Mattheüs Theodor Loots
E-mail: theodor.loots@up.ac.za

Abstract

Advances in computing has undoubtfully been one of the main catalysts in the formation of the discipline always known as Statistics. A fundamental question addressed here is whether computing facilities, such as parallel or high performance computing, could assist in the development of methodologies that render stronger results, based on some predetermined optimality criterion. The candidate at the hand of which this enquiry is made, is the arc length of some statistical function. Estimation, goodness-of-fit, linear regression and non-linear regression, which may all be considered as central themes in Statistics, are revisited, and redefined in terms of this new measure. The results resulting from these arc length methodologies are obtained from simulation, as well as from real case studies, and contrasted to that obtained using their classical counterparts. Mathematical premises for the proposed methods are provided, together with the documentation accompanying the companion R package, along with the data utilised for the applications.

Keywords: Arc length, goodness-of-fit, non-linear regression, parameter estimation, regression.

Supervisor : Prof. A. Bekker
Department : Department of Statistics
Degree : Philosophiae Doctor

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

> “Use?” replied Reepicheep. “Use, Captain? If by use you mean filling our bellies or our purses, I confess it will be no use at all. So far as I know we did not set sail to look for things useful but to seek honour and adventures. And here is as great an adventure as ever I heard of, and here, if we turn back, no little impeachment of all our honours.”

Lewis, C. S. “The Voyage of the Dawn Treader. 1952.” The Complete Chronicles of Narnia (1980): 286 – 369.

> “He who loves practise without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast.”

Leonardo da Vinci

To Katryn, Katelan, Matthew and Galen, with love.

© University of Pretoria

Acknowledgements

Without the various contributions described in the following list, this work would not have been able to stand as it is today:

- To God, my creator: Thank You for providing the breath to sing Your song; You alone shall have my heart;
- My home, Katryn, Katelan, Matthew and Galen: Thank you for all the motivation and inspiration, and for keeping my feet on the earth;
- My natural and spiritual family: You kept me going;
- Mrs. Hestelle Viljoen: There once was a Mathematics teacher who tought me that potential only describes the void of things to come. Thank you for believing;
- Prof. Andriëtte Bekker: Thank you for sharing a journey very few others would have embarked upon;
- Colleagues at the Department of Statistics: I cannot list you all, but thank you for fourteen years of dedication and support;
- My co-authors, Dr. Steven Hussey and Prof. Walter Focke: Your problems made this thesis come to life;
- Prof. Brenda Wingfield: The mug;
- Examinors: Your attention to detail completed this work;
- The Centre for High Performance Computing: Thank you for providing access to Lengau, and especially to Mr. Dane Kennedy for your enthusiasm and support;
- This work is based on the research supported in part by the "National Research Foundation" of South Africa for the grant, Unique Grant No. 94108. Any opinion, finding and conclusion or recommendation expressed in this material is that of the author(s) and the NRF does not accept any liability in this regard.

© University of Pretoria

Contents

© University of Pretoria

© University of Pretoria

© University of Pretoria

© University of Pretoria

List of Figures

© University of Pretoria

© University of Pretoria

List of Tables

© University of Pretoria

© University of Pretoria

Chapter 1

Introduction

Statisticians have always been interested in the development and refinement of procedures, i.e. those having greater statistical power, efficiency, maximum entropy, largest r^2*, smallest variance, ..., and the list goes on. The ingredient providing the edge, of one technique over another, may for instance be quantified using measures from information theory, which simply states that the procedure contributing more "information" is preferred above the other. It is this prospecting for more information, that has developed into the subject of the current enquiry. Such developments have been heightened in the current age, where statisticians find themselves in a balancing game of theory and computation, as described by Efron and Hastie (2016) [25].*

The approach here is therefore to use a mathematically intractable measure, namely the arc length, as pensieve by means of which musings on the positioning of classical Statistics within the broader field of Data Science, are provided. To this end, parameter estimation, goodness-of-fit, regression and non-linear regression will be revisited, and redefined in terms of the arc length of suitable statistical functions. Reflections along these lines are found in "Data Science: The End of Statistics" by Larry Wasserman[1]*, and on the exercising of new-found cloud powers by Jonathan Rosenblatt*[2] *which are brought about by high performance computing.*

[1] https://normaldeviate.wordpress.com/2013/04/13/data-science-the-end-of-statistics/

[2] https://www.r-statistics.com/2013/07/analyzing-your-data-on-the-aws-cloud-with-r/

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Holst and Rao (1980) [43] provides the description of the problem of random spacings in everyday English, when the lengths of a stick, broken at random, are discussed. This problem is however much older, and was considered by Pearson (1902) [68], when he studied the mean value of a spacing in response to a problem posed by Francis Galton. Many of the earlier approaches to the random division of an interval are surveyed by Moran (1947) [65], in a response to Greenwood (1946) [38]. Yet another review of the earlier contributions to random spacings was made by Pyke (1965 and 1972) [71, 72] and discussions found therein.

Arc lengths has been considered, and explicitly mentioned, by Rao (1976) [74], in the context of the distance between observations on the circle, i.e. spacings in the circular world, where the goodness-of-fit (GoF) and two-sample problem were treated. Tung and Rao (2012) [79] used U-statistics (see Hoeffding (1948) [41]) for developing U-statistics based on spacings. They proposed yet another test for uniformity on the circle in (2013) [80], by using Gini's mean difference statistic for sample arc lengths. These circular sample arc lengths correspond to sample spacings on the real line. Since many applications of spacings are in directional statistics, or since the probability integral transform is often used in changing a problem on the real line into a problem on the circle, it is worth remarking the contributions by Mardia and Jupp (2009) [57], and Jammalamadaka and Sengupta (2001) [47]. In their work, the developments, and general theory of directional statistics is reviewed, and the theory for distributions wrapped around the circle is discussed, but is also generalised to other manifolds.

Cressie (1976) [17] generalised first order spacings to m-order spacings, and used the sum of their logarithms in testing for uniformity. For this, he also showed asymptotic normality under the null hypothesis. Holst (1979) [42] showed the asymptotic normality of functions of spacings, involving sums. The asymptotic theory of functions of spacings was further studied by Holst and Rao (1980, 1981) [43, 44], where they showed its application to the two-sample case. Here the non-parametric flavour of these tests are emphasised. Gatto and Rao (1999) [34] provide a saddlepoint approximation for an m-statistic conditional on another. A special case of this is that dependent spacings may be written in terms of independent and identically distributed (IID) random variables, conditioned on their sum. The sample arc length statistics proposed here, involve the summation of spacings, of order one, and the mapping of that into the range of the statistical function, i.e. creating a dependence structure.

© University of Pretoria

Saddlepoint approximations were introduced into the statistical literature by Daniels (1954) [18], and is frequently encountered in the theory of spacings. Ma and Robinson (1999) [56] used saddlepoint approximations in deriving approximate distributions for the difference of order statistics. In other words, saddlepoint approximations were used in finding approximate distributions of spacings. The book by Butler (2007) [11] provides an overview of its use, and extensions to the multivariate case, as well as a discussion on how it may be used with other base distributions than the normal. Ghosh and Rao (1998) [35] used saddlepoint approximations in deriving approximate distributions for statistics of spacings involving small samples, and also tabled some values for distributions that can't be derived explicitly.

Despite the rich literature on the theory of spacings, and the use of saddlepoint approximations, no exact or approximate distributions for the sample arc length statistics presented here, could be obtained. However, as mentioned above, the tabulation of critical values are often encountered for sample statistics involving spacings. An extended table for Rao's Spacing Test is for instance given by Russell and Levitin (1995) [75]. Chen (2004) [13], used simulation when a test for a difference in scale between two populations, based on spacings, was proposed.

Since explicit solutions to arc length formulae are seldom obtained, the voyage of finding the most suitable candidate distribution for the problem, is abandoned from the outset. The point of departure, then, is approximation and simulation. This notion is supported by McElreath (2016) [58] in commenting on computational statistics, that a purely mathematical approach is anyways insufficient.

1.1 Motivation

The two questions posed by Larry Wasserman (2013) [84], provides the fuel for the work presented here:

1. Why do Statisticians find themselves left out of the Data Science discussion?

2. What can Statisticians do about it?

With regards to the first, assuming that computational capabilities are infinite, can Statisticians offer theoretically sound methodologies, which improve on those existing in the literature? When the day dawns that the disclaimer "These are computationally

© University of Pretoria

intensive methods requiring supercomputing facilities…" can safely be left out of newly proposed work, what will the Statistician have to offer? Heeding the warning of Leonardo da Vinci quoted at the outset of this work, and echoed by Azzalini and Scarpa (2012) [7] (amongst others), a theoretical framework is proposed for the computationally intensive methods presented here.

Secondly, making use of parallel and high performance computing, the methods presented here, should be refined, and the available options, selected as to maximise their performance. These options include optimisation methodologies, and numerical analysis routines, such as numerical integration.

1.2 Objectives

- Develop a method for parameter estimation using arc lengths, namely the method of arc lengths.
- Develop a GoF test using arc lengths.
- Present a framework for arc length regression (ALR).
- Extend the ALR framework to the non-linear setting, namely non-linear arc length regression (NALR).
- Formulate the mathematical premises for arc length based techniques.

1.3 Contributions

The novel contributions this thesis sets out to make are:

- The development of an estimation method, that renders biased estimates, with smaller root-mean-square error (rMSE) and biases than existing methods.
- Determine conditions under which rejection regions for a newly proposed GoF test are larger than that of competitive tests, i.e. rendering a test with greater sensitivity.

© University of Pretoria

- The presentation of a flexible linear regression technique, that utilises tuning parameters which yields larger r^2 values, and smaller sample divergences, using the same model specification as ordinary least squares (OLS).
- Propose a non-linear regression technique for fitting a sigmoidal function, under non-linear parameter constraints, as an alternative to non-linear least squares (NLS).
- Develop an R [73] package for enhancing the reproducibility of all methods proposed here.

1.4 Thesis Outline

- **Chapter 2** provides the necessary background for developing the basic intuition for the arc length based techniques to follow in the consecutive chapters.
- **Chapter 3** focuses on the classical estimation problem in Statistics. The method of arc lengths, along with the arc length test for GoF is presented.
- **Chapter 4** extends the idea of a sequence of data values, to a set of dependent and independent observations. Here, arc length regression and the method of moments (regression) are introduced.
- **Chapter 5** generalises the linear model assumption of Chapter 4 to the non-linear setting.
- **Chapter 6** proposes the mathematical premises for the techniques presented in this thesis.
- **Chapter 7** summarises the main findings of this thesis.

The set of appendices include:

- **Appendix A** describes the “alR” [54] package documentation for the functions used in this thesis.
- **Appendix B** provides additional simulation results for the estimation technique, applied to the normal distribution, introduced in Chapter 3.

© University of Pretoria

- **Appendix** C includes comparative analyses from two popular machine learning techniques, for the data used in Chapter 4.
- **Appendix** D provides a list of the acronyms used throughout this thesis, as well as their associated definitions.
- **Appendix** E lists and defines the mathematical symbols used in this work, categorised according to the relevant chapter in which they appear.
- **Appendix** F lists the publications derived from, and that is associated with this work.

An index of some of the key terms used in this thesis, begin on page 129.

© University of Pretoria

Chapter 2

Preliminaries

In this chapter, the mathematical background is reviewed in Section 2.1, required for the development of a basic intuition for the methods proposed in later chapters. Section 2.2 describes the implementation infrastructure required for bringing all this together.

2.1 Mathematical Background

Kent, Mardia and Rao (1979) [49] proved a characterisation of the uniform distribution on the circle. It is proposed that distributions may be characterised, not only on the circle, but also on the real line, by the lengths of arcs of statistical functions, over partitions of its domain. It makes for instance intuitive sense that the arc length of the probability density function (PDF), of a unimodal random variable, will be halved by the median, i.e. have equal lengths to the left and right of the fiftieth percentile. The formal definition of an arc length is now given in Definition 2.1.1.

Definition 2.1.1. Define the arc length of the function $f(x,\theta)$, on the interval $[a,b]$ (on which it is continuous) as

$$\mathcal{S}_f^{[a,b]}(\theta) = \int\limits_a^b \sqrt{1+(f'(x,\theta))^2}dx \tag{2.1.1}$$

where $f'(\cdot)$ denotes the derivative of $f(\cdot)$ with respect to x, θ is additional parameters of the function $f(\cdot)$, and where $a \leq b$ is in the domain of $f(\cdot)$.

For the normal distribution, the PDF with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$

© University of Pretoria

is for instance given by

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

in the usual way, so that (2.1.1) becomes

$$\begin{aligned} \mathcal{S}_f^{[a,b]}(\mu, \sigma) &= \int\limits_a^b \sqrt{1 + (f'(x, \mu, \sigma))^2} dx \\ &= \int\limits_a^b \sqrt{1 + \left(-\frac{(x-\mu)}{\sigma^2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right)^2} dx \\ &= \int\limits_a^b \sqrt{1 + \frac{(x-\mu)^2}{2\pi\sigma^6} e^{-\frac{1}{\sigma^2}(x-\mu)^2}} dx. \end{aligned} \tag{2.1.2}$$

In Chapter 3 for instance, the discussion is concerned with the arc length of the PDF, so that the subscript may be dropped, i.e. $\mathcal{S}^{[a,b]}(\mu, \sigma)$ is used throughout.

The optimal choice of the interval $[a, b]$, remains an open question, and is viewed as tuning parameters. However, selection of an interval, covering the greater portion of the range of observed values, seems like a fair strategy. For example, $[a, b] = [q_{0.025}, q_{0.975}]$ results in an arc spanning the 0.025^{th} to the 0.975^{th} sample quantiles; thus excluding 5% of the data values. This is done, since the arc length over an infinite domain, would also be infinite.

2.2 Implementation Infrastructure

An “R” (see [73]) package “alR” (see Loots (2017) [54]) was developed for the methods proposed here. Code has been written in c++ as far as possible using “Rcpp” by Eddelbuettel and François (2011) [22], and Eddelbuettel (2013) [21] and “RcppArmadillo” by Eddelbuettel and Sanderson (2014) [23].

Because of the computational intensive nature of the methods proposed, the Lengau supercomputer from the CHPC[1] was used, which is a Dell, Linux, homogeneous cluster, comprising Intel 5th generation CPUs. The amount of compute nodes used for a particular application will be noted where applicable. The job distribution was performed using task-pull parallelism from the “pbdMPI” package by Chen et al (2012) [12]).

[1] http://www.chpc.ac.za/

© University of Pretoria

Further technical requirements specific to a particular chapter will be provided where necessary.

2.3 Summary

The fundamental mathematical concepts for the development of the methods proposed in the following chapters were reviewed in Section 2.1. The technical specifications regarding their implementation and execution were presented in Section 2.2.

In Chapter 3, the first musings on the use of the arc length measure is presented for cross-sectional data, for estimation and goodness-of-fit problems.

© University of Pretoria

Chapter 3

Estimation

In this chapter, the method of arc lengths applied to the estimation of the parameters of normally distributed variables, is presented, based on the arc length of the PDF. It is furthermore shown how arc lengths may be used for GoF problems (as the arc length test).

Here, the lengths of a broken stick aren't considered explicitly, as by Holst and Rao (1980) [43] (as mentioned in Chapter 1) but rather the lengths of a PDF, "broken" at random, i.e. the arc lengths of a density function over sub-domains of its support. After providing a basic definition and describing the methods utilised here in Section 3.1, a sample statistic is formulated in terms of spacings, and functions thereof, in Section 3.2. Apart from the review on spacings given in Chapter 1, it is noted here that spacings have also been used in the construction of GoF tests, by Ghosh and Jammalamadaka (2001) [36], Ekström (2008) [26], and Mirakhmedov and Jammalamadaka (2013) [64].

The method of arc lengths proposed here is similar to that of the minimum distance estimators (or maximum goodness-of-fit estimators) as described by Wolfowitz (1953) [86], Kac, Kiefer and Wolfowitz (1955) [48], Pollard (1980) [69], and more recently by Luceño (2006) [55], in that:

1. it may be used in parameter estimation (Section 3.3), and
2. it may be used as a measure of GoF for which critical values are simulated via a parametric bootstrap (Section 3.4).

Two special distances that will receive attention here, for comparative purposes, are Cramér-von Mises (CvM), which may be regarded as a special case of Anderson-Darling

© University of Pretoria

(AD) (see Anderson and Darling (1952) [4]).

3.1 Implementation Infrastructure

In addition to the specifications set out in Chapter 2, the following applies to this chapter.

The violin plots in Figures 3.1 and 3.2 were constructed using the "vioplot" package (see Adler (2005) [2]), based on Hintze and Nelson (1998) [40]; and the heatmaps in Figures 3.4 and 3.6 using "heatmap.2" from "gplots" (see Warnes et al (2016) [83]).

The construction of these heatmaps required the Lengau supercomputer from the CHPC. For Figure 3.4, 6 compute nodes with 24 cores and 128 GiB memory each were utilised, but was down-scaled to 5 compute nodes with 24 cores each for Figure 3.6.

Parameter estimation and GoF were respectively carried out using the packages "fitdistrplus" by Delignette and Dutang (2015) [19] and "goftest" by Faraway et al (2015) [29]. The random samples from the skew normal distribution and the power exponential distribution were respectively simulated using "sn" by Azzalini (2016) [6] and "normalp" by Mineo (2014) [62].

3.2 Sample Arc Length Statistics

Suppose that $x_1, \ldots, x_n$ is an observed sample from a distribution, having PDF $f(x)$, with corresponding ordered values $x_{(1)}, \ldots, x_{(n)}$. Suppose further that $[a, b]$ is some interval in the domain of f, such that $a \leq x^*_{(1)} \leq \ldots \leq x^*_{(m)} \leq b$. Here $x^*_{(i)}, i = 1, \ldots, m$ denote the m ordered observations in the interval $[a, b]$. (2.1.2) is then estimated by a finite sum of lengths of straight lines, as

$$ {}_1f^{[a,b]} = \sum_{i=2}^{m} \sqrt{\left(x^*_{(i)} - x^*_{(i-1)}\right)^2 + \left(\hat{f}\left(x^*_{(i)}\right) - \hat{f}\left(x^*_{(i-1)}\right)\right)^2}, \tag{3.2.1} $$

where $\hat{f}(\cdot)$ is the kernel density estimate (KDE) of $f(\cdot)$, based on the Gaussian kernel, and is given by

$$ \hat{f}(x) = \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi} hn} e^{-\frac{(x-x_i)^2}{2h^2}}, $$

and where

$$ h = s_x \left(\frac{4}{3n}\right)^{1/5}, $$

is the smoothing parameter suggested by Silverman (1986) [77]. The choice of the kernel and smoothing parameter correspond to the hypothesis, that the observed data is normally distributed.

Since the KDE has already been mentioned, and is a PDF in its own right, another estimator resulting from this, is

$$ {}_2\int^{[a,b]} = \int_a^b \sqrt{1 + \left(\hat{f}'(x)\right)^2}\,dx, \tag{3.2.2}$$

which is the continuous analogue of (3.2.1).

Since explicit distributions of the sample arc length statistics aren't derived explicitly, 1,000 samples of size 1,000 were simulated from the $N(2, 3.5)$ distribution, and (3.2.1) and (3.2.2) calculated, in providing an empirical distribution for each. The violin plots for the measurements generated by (3.2.1) and (3.2.2), are given in Figures 3.1 and 3.2, for three interval choices:

(a) $[a, b] = [q_{0.01}, q_{0.99}] = [-6.1422175591, 10.1422175591]$,

(b) $[a, b] = [q_{0.05}, q_{0.95}] = [-3.7569876943, 7.7569876943]$, and

(c) $[a, b] = [q_{0.10}, q_{0.90}] = [-2.4854304794, 6.4854304794]$.

The final row in Table 3.1 and 3.2 indicates the quantile at which the following "theoretical" arc lengths (2.1.2) were respectively observed (sij, $i = 1, 2$, and $j = a, b, c$ refer to the estimators and interval choices respectively):

(a) $\mathcal{S}^{[-6.1422175591,10.1422175591]}(2, 3.5) = 16.2860589511$,

(b) $\mathcal{S}^{[-3.7569876943,7.7569876943]}(2, 3.5) = 11.5153831966$, and

(c) $\mathcal{S}^{[-2.4854304794,6.4854304794]}(2, 3.5) = 8.9719304133$.

Note that for the discrete sample statistic, given in (3.2.1), the "theoretical" value was only observed in the range of the empirical distribution once the segment over which the arc length was computed included less tail observations.

© University of Pretoria

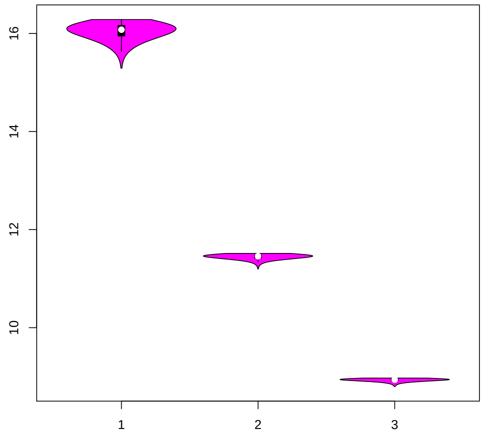


Figure 3.1: Violin plots for the empirical distributions of discrete arc length statistics: ${}_1\int^{[-6.1422175591,10.1422175591]}(2,3.5)$, ${}_1\int^{[-3.7569876943,7.7569876943]}(2,3.5)$, and ${}_1\int^{[-2.4854304794,6.4854304794]}(2,3.5)$.

	s1a	s1b	s1c
Min	15.16233	11.16418	8.76773
Q1	15.96189	11.42711	8.91885
Mean	16.04576	11.44779	8.93275
Median	16.07521	11.45979	8.93825
Q3	16.16260	11.48252	8.95326
Max	16.27826	11.51517	8.97279
F(S)	1.00000	1.00000	0.99600

Table 3.1: Summary statistics for the distributions of discrete sample arc length statistics.

© University of Pretoria

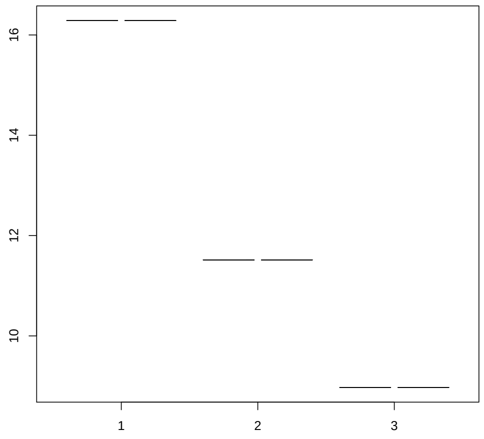


Figure 3.2: Violin plots for the empirical distributions of continuous arc length statistics: ${}_{2}\!\int^{[-6.1422175591,10.1422175591]}(2, 3.5)$, ${}_{2}\!\int^{[-3.7569876943,7.7569876943]}(2, 3.5)$, and ${}_{2}\!\int^{[-2.4854304794,6.4854304794]}(2, 3.5)$.

	s2a	s2b	s2c
Min	16.28563	11.51479	8.97138
Q1	16.28588	11.51518	8.97172
Mean	16.28598	11.51529	8.97184
Median	16.28597	11.51529	8.97184
Q3	16.28607	11.51539	8.97196
Max	16.28652	11.51577	8.97261
F(S)	0.71700	0.73700	0.69300

Table 3.2: Summary statistics for the distributions of continuous sample arc length statistics.

From Figure 3.2, and the summary statistics in Table 3.2, the robustness of the continuous sample arc length statistic is evident, given the small variation in the observed values.

© University of Pretoria

3.3 Parameter Estimation

Generally, a set of observations is tested under the null hypothesis of being generated by a normal distribution, for which the parameters must be estimated, i.e. let $X_1, \dots, X_n \sim N(\mu, \sigma)$ be IID, with μ and σ unknown. The $(\hat{\mu}, \hat{\sigma})$ combination that satisfies

$$\mathcal{S}^{[a,b]}(\mu, \sigma) = \int^{[a,b]},$$

where $\int^{[\cdot]}$ is one of the sample arc length statistics defined in Section 3.2. This leads to the following objective function, based on the Euclidean distance:

$$\sqrt{\left(\mathcal{S}^{[a,b]}(\mu, \sigma) - \int^{[a,b]}\right)^2}. \tag{3.3.1}$$

$(\hat{\mu}, \hat{\sigma})$ is therefore the (μ, σ) pair that simultaneously minimises this objective function. In this case, simplification of (3.3.1) results in the absolute difference, however, the degree of fit may be controlled by the inclusion of additional subintervals, and expansion of the Euclidean distance in the usual fashion. This however complicates the matter of choosing optimal tuning parameters, i.e. selecting integration intervals for which this Euclidean distance is minimised.

Tables 3.3 and 3.4 summarise the estimated bias and rMSE values (in ascending order of rMSE) for the median values of the parameters estimated from 1,000 samples of size 1,000, from a $N(2, 3.5)$ distribution. The method of arc lengths (resulting from the two proposed sample arc length statistics given in (3.2.1) and (3.2.2)), maximum-likelihood (ML), CvM, and AD estimation techniques are compared. For each of the proposed sample arc length statistics, five variants were included depending on the interval $[a, b]$, where the two end points are specified in terms of the sample quantiles:

(a) $[0.01, 0.99]$,

(b) $[0.025, 0.975]$,

(c) $[0.05, 0.95]$,

(d) $[0.075, 0.925]$, and

(e) $[0.1, 0.9]$.

	bias	rMSE
MLE	-0.00735	0.07379
AD	-0.00819	0.07610
CvM	-0.00804	0.07935
s1b	0.08767	0.12631
s1c	0.14413	0.17853
s1a	0.15275	0.18645
s2e	0.20622	0.21606
s2a	0.22728	0.23129
s2d	0.22842	0.23211
s2c	0.23264	0.23988
s2b	0.24070	0.24381
s1e	0.26803	0.26803
s1d	0.27966	0.28752

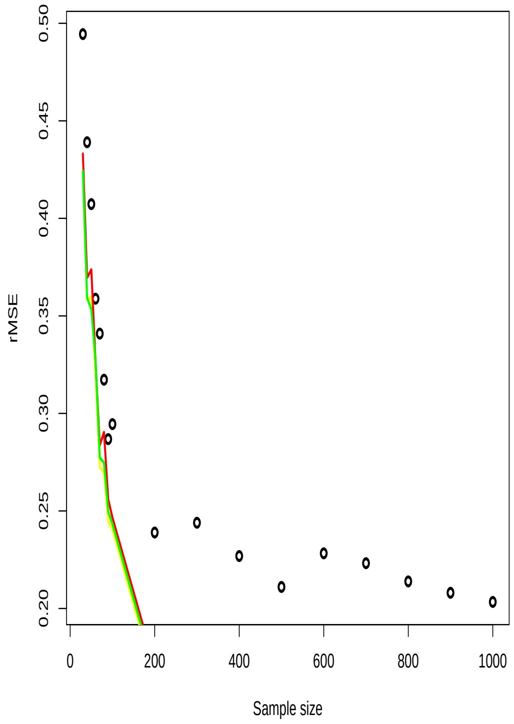
Table 3.3: Estimated bias and rMSE for μ.

Using the method of arc lengths with ${}_1\int^{[q_{0.01},q_{0.99}]}$ yielded the smallest bias and rMSE values for simultaneously estimating μ and σ. All sample arc length statistics performed worse than the maximum-likelihood estimation (MLE), AD and CvM methods for estimating μ. The same holds true for the estimation of σ, accept for ${}_1\int^{[q_{0.01},q_{0.99}]}$ and CvM, for which the ordering are interchanged in certain cases (see Appendix B). In Figure 3.3 these results are reproduced using only ${}_1\int^{[q_{0.01},q_{0.99}]}$, for varying sample sizes.

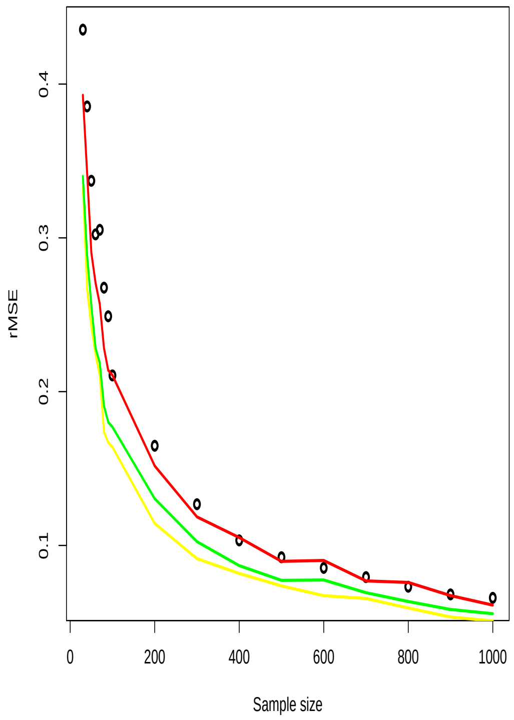
From Figure 3.3, it is clear that the rMSE values for the method of arc lengths with ${}_1\int^{[q_{0.01},q_{0.99}]}$ decrease for increasing sample sizes. For estimating σ, this corresponds well to the values obtained for the CvM method, but are not as strong as that obtained for the MLE and AD methods.

Appendix B provides simulation results for a number of μ and σ choices, and for varying sample sizes, with similar results given here.

© University of Pretoria



(a) Estimated rMSE for μ.



(b) Estimated rMSE for σ.

Figure 3.3: Estimated rMSE for μ and σ for varying sample sizes.

© University of Pretoria

	bias	rMSE
MLE	-0.00256	0.05460
AD	-0.00118	0.05638
s1a	-0.00956	0.06043
CvM	-0.00705	0.06375
s2a	0.06186	0.08131
s2b	0.06544	0.08554
s2c	0.07228	0.09172
s2d	0.07086	0.09981
s2e	0.07402	0.10673
s1b	-0.13979	0.13993
s1c	-0.35670	0.35670
s1d	-0.56492	0.56492
s1e	-0.75129	0.75129

Table 3.4: Estimated bias and rMSE for σ.

3.4 Goodness-Of-Fit

After obtaining estimates $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ, from a sample $\underset{n\times 1}{\underline{x}} = x_1, \ldots, x_n$, a GoF procedure may be setup using ${}_2\!\int$.[1] The procedure for the arc length test is as follows:

1. Under the null hypothesis: $H_0 : \underline{x} \sim N(\hat{\mu}, \hat{\sigma})$.
2. Calculate ${}_2\!\int$ from $\underline{x}$ as a test statistic.
3. Simulate m samples of size n from the $N(\hat{\mu}, \hat{\sigma})$ distribution.
4. Calculate ${}_2\!\int$ in each case, i.e. a parametric bootstrap.
5. A two-sided p-value may then be obtained from this bootstrap distribution in the usual fashion.

The choice of the interval $[a, b]$ yield tests of varying strengths, and a number of these will be investigated. This class of GoF procedures is now compared to the CvM and AD methods in detecting departures from normality, first against skew alternatives, and

[1] Due to the discrete nature of ${}_1\!\int$, it is not particularly useful in a GoF setting.

© University of Pretoria

secondly against alternatives with mesokurtic deviations. For the arc length test, the parameters are estimated using ${}_1\int$ from (3.2.1), and ${}_2\int$ from (3.2.2), and for the CvM and AD tests, the CvM and AD estimation methods are respectively used. In other words, four GoF procedures will be compared.

3.4.1 Skewness Considerations

The skew normal distribution proposed by Azzalini (1985) [5] has PDF

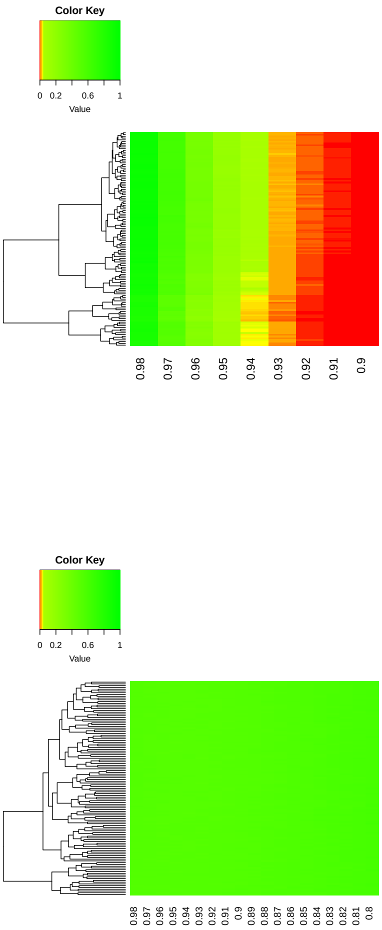
$$f(x,\mu,\sigma,\alpha) = \frac{1}{\pi\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \int\limits_{-\infty}^{\alpha\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt$$

where $x \in \mathbb{R}$, and has location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$, and shape parameter $\alpha \in \mathbb{R}$, and includes the normal distribution as a special case, when the "skewness" parameter $\alpha = 0$. It is desired to test how sensitive the proposed arc length method is in detecting skewed departures from normality. Random samples of size 1,000 were simulated from the skew normal distribution, for the skewness parameter in the interval $[-3, 3]$, and in each case:

1. it was assumed that the data was generated by a normal distribution,
2. $(\hat{\mu}, \hat{\sigma})$ were estimated from the sample,
3. the p-values for the respective GoF methods were recorded, and
4. the entire experiment was repeated 100 times, and the median p-value at each skewness level recorded, in stabilising results.

The axis labels in the heatmaps indicate the range of the sub-domains over which the arc lengths were calculated, with equal distances from each endpoint. E.g. ${}_1\int^{[q_{0.03},q_{0.97}]}$ will be included with the $0.97 - 0.03 = 0.94$ label.

© University of Pretoria



(a) Estimation using ${}_1\int$.

(b) Estimation using ${}_2\int$.

Figure 3.4: Heatmaps of p-values for GoF using arc lengths (skewness considerations).

Significance levels often range between 0.01 and 0.05 and the p-values corresponding to this are indicated by the yellow regions in the heatmaps, while p-values smaller than 0.01 are coloured red. Using the discrete sample arc length statistic ${}_1\int^{[q_{0.03},q_{0.97}]}$ seems to be the optimal choice at a 5% level of significance, since any larger intervals will never reject the null hypothesis, while smaller intervals, will almost always lead to its rejection. From the green heatmap represented by ${}_2\int$, it is clear that although the parametric bootstrap distribution under the null hypothesis, is simulated using this continuous sample arc length statistic, it never leads to its rejection.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

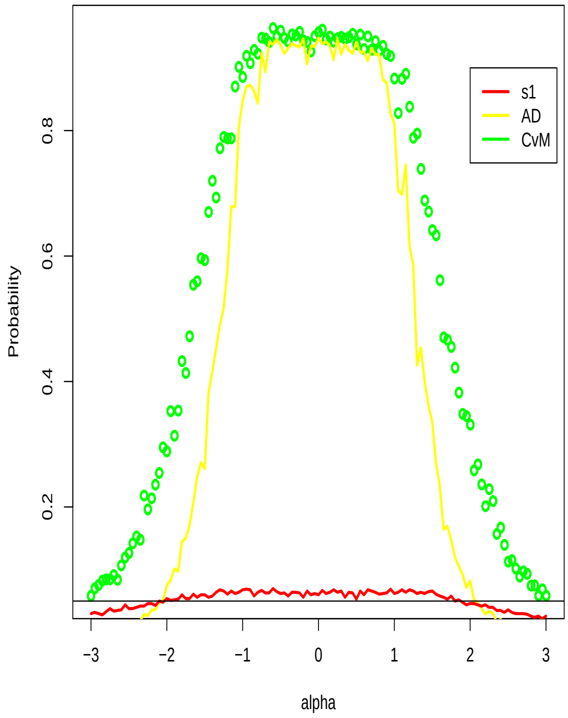


Figure 3.5: GoF comparisons for skew alternatives.

From Figure 3.5 it is clear that the CvM test never leads to the rejection of the null hypothesis for the skewness parameter in the interval $[-3, 3]$. The arc length test (using $_{1}f^{[q_{0.03}, q_{0.97}]}$ rejects the null hypothesis for values smaller than -2.1, (ignoring the jump at -2.05) and then again for values greater than 1.85 (inclusive). The AD test stops rejecting the null hypothesis at roughly the same point as the arc length test, but only starts rejecting again at 2.1. Thus although the arc length test rejects for a larger range of positively skewed alternatives, the AD test does so with greater decisiveness (i.e. a larger peak).

© University of Pretoria

3.4.2 Mesokurtic Considerations

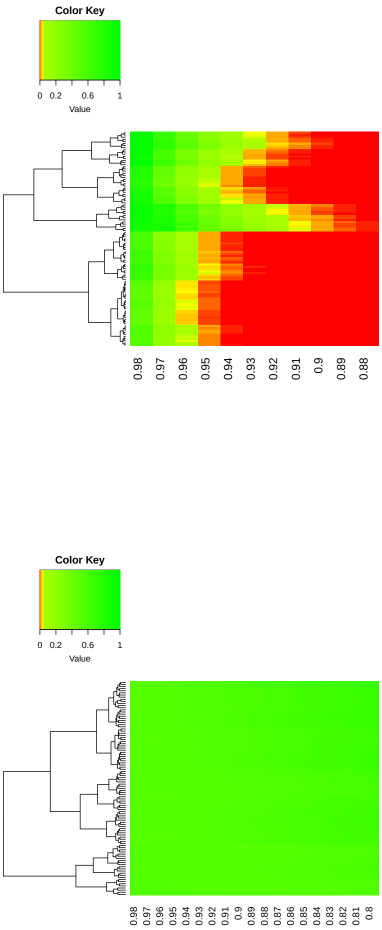
The exponential power distribution is a generalisation of the normal distribution, for which a third parameter p is introduced in rendering mesokurtic deviances. It has PDF

$$f(x, \mu, \sigma, p) = \frac{1}{2p^{\frac{1}{p}}\Gamma(1+\frac{1}{p})\sigma} e^{-\frac{|x-\mu|^p}{p\sigma^p}}$$

for $x \in \mathbb{R}$, location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma > 0$, and shape parameter $p > 0$. For a history on this distribution, and the various names it goes by, see Mineo et al (2005) [63]. This distribution has as special cases the Laplace distribution for $p = 1$ (leptokurtic), normal distribution for $p = 2$, and tends to the uniform (platykurtic), as $p \to \infty$.

Random samples of size 1,000 were simulated from the exponential power distribution, for p in the interval $[1, 5]$, with the remainder of the experiment following that outlined in Section 3.4.1.

© University of Pretoria



(a) Estimation using ${}_1\!\int$.

(b) Estimation using ${}_2\!\int$.

Figure 3.6: Heatmaps of p-values for GoF using arc lengths (mesokurtic deviations).

As with the skew alternatives considered in Figure 3.4, using the discrete sample arc length statistic ${}_1\!\int^{[q_{0.03},q_{0.97}]}$ seems to be the optimal choice at a 5% level of significance, since any larger intervals will decrease the rejection region, while smaller intervals increase the Type I error. The case for the continuous sample arc length statistic ${}_2\!\int$, is similar to that described for skew alternatives.

© University of Pretoria

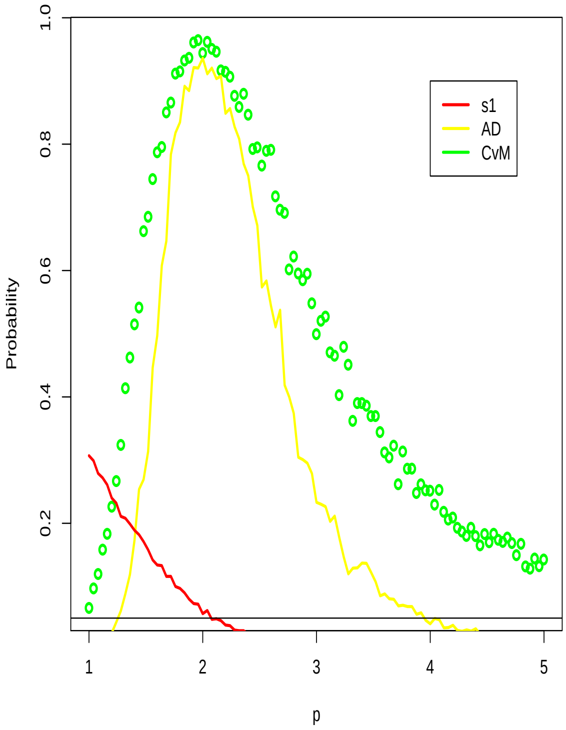


Figure 3.7: GoF comparisons for various mesokurtic deviances.

From Figure 3.7 it is clear that the CvM test never leads to the rejection of the null hypothesis for $p \in [1, 5]$. The arc length test (using ${}_1\int^{[q_{0.03}, q_{0.97}]}$) rejects the null hypothesis for values larger than 2.04, but not for any leptokurtic alternatives. The AD test stops rejecting the null hypothesis at $p = 1.28$, but only starts rejecting again at 3.96. Thus the AD test is more sensitive to leptokurtic alternatives, while the arc length test has greater sensitivity against platykurtic alternatives.

3.5 Summary

The method of arc lengths, for parameter estimation, was introduced here, and it was shown that although it being consistent, generally leads to parameter estimates with greater bias and rMSE than ML, CvM, and AD methods. The discrete arc length sample statistic, ${}_1\int^{[q_{0.01}, q_{0.99}]}$ proved to be particularly effective for estimating σ, and offers results

© University of Pretoria

comparable to that obtained by the CvM method.

The arc length GoF test is constructed using the continuous sample arc length statistic ${}_2\!\int$, for which a parametric bootstrap yields a null distribution. A GoF test constructed with ${}_2\!\int^{[q_{0.03},q_{0.97}]}$, with parameters estimated using ${}_1\!\int^{q_{0.03},q_{0.97}}$ yielded the most powerful tests. It was shown that the arc length test is more powerful against positively skewed and platykurtic alternatives than both the CvM and AD tests.

In the following chapter, the single sequence of data considered here, is extended to sets of dependent and independent observations, as the focus shifts to linear regression.

© University of Pretoria

Chapter 4

Arc Length Regression

Chapter 3, considered a single sequence of data, and discussed parameter estimation along with GoF. In this chapter, two related sets of data are considered, one being a dependent outcome of the other.

The geometric interpretation of OLS, in a regression context, is the process by which the set of coefficients are sought, that minimises the sum of the vertical distances between a data point, and the regression line assumed to describe the data. For a historical review, see Seal (1967) [76]. Although numerous extensions and generalisations of the basic regression model have been proposed, a particular variation, having a geometric interpretation, involving distances, are worth mentioning.

Orthogonal regression seeks to minimise the sum of the orthogonal distances from a point to the line, and not the vertical distance (see Coolidge (1913) [16] and Minda and Phelps (2008) [61]). This is a special case of what is now referred to as Deming regression (see Adcock (1878) [1], Kummell (1879) [50], Deming (1943) [20]), a type of errors-in-variables model, obtained when equality of the errors in the dependent and independent variables are assumed. This on its part, is again a special case of total least squares (TLS), for which no general solution exists (see Golub and Van Loan(1980) [37]).

The importance of geometric- as opposed to algebraic methods, in recent applications, are emphasised by Ahn (2008) [3], Liu and Wang (2008) [53], and Chernov (2011) [14]. It makes sense that a geometrical approach yields beneficial results in fitting curves to objects in space, but would there be any advantage in applying such techniques directly to PDF's? After some preliminaries in Section 4.1, the framework for addressing this question is developed in Sections 4.2 and 4.3.

© University of Pretoria

A problem in biology is then discussed as an application in Section 4.5, and the findings summarised in Section 4.6.

4.1 Introduction

Suppose that $\underset{n\times 1}{\underline{y}}$ is a vector of n dependent observations, and that $\underset{n\times p}{\mathbf{X}} = \left(\underset{n\times 1}{\underline{x}_1}, \ldots, \underset{n\times 1}{\underline{x}_p}\right)$ is a set of p independent variables (where a column may or may not be included for an intercept). In OLS regression the following is assumed:

$$\underline{y} = \mathbf{X}\underline{\beta} + \underline{\epsilon}$$

with $\underset{p\times 1}{\underline{\beta}}$ the set of coefficients to be determined, and $\underset{n\times 1}{\underline{\epsilon}}$ a vector of stochastic error terms with $E[\underline{\epsilon}] = \underline{0}$, and $var(\underline{\epsilon}) = \sigma^2\mathbf{I}_n$, where $\underset{n\times n}{\mathbf{I}_n}$ is the identity matrix. The stochastic mechanism by which modelling takes place, is then through

$$\begin{aligned} E[\underline{y}] =& E[\mathbf{X}\underline{\beta} + \underline{\epsilon}] \\ =& \mathbf{X}\underline{\beta}, \text{ and} \\ var(\underline{y}) =& var(\mathbf{X}\underline{\beta} + \underline{\epsilon}) \\ =& var(\underline{\epsilon}) \\ =& \sigma^2\mathbf{I}_n, \end{aligned}$$

since $\mathbf{X}\underline{\beta}$ is assumed fixed. The OLS solutions $\underline{\hat{\beta}}$ are unbiased, having minimum variance, and if the further assumption is made that

$$\underline{\epsilon} \sim N(\underline{0}, \sigma\mathbf{I}_n)$$

i.e. following a normal distribution, then the coefficient estimates $\underline{\hat{\beta}}$ are given by the ML solutions with known distributions.

A closer look at this model shows that an equivalent formulation is to assume that

$$\underline{\gamma} \sim N(\mathbf{X}\underline{\beta}, \sigma\mathbf{I}_n)$$

with $\underset{n\times 1}{\underline{\gamma}}$, so that the model may be rewritten as

$$\underline{y} = \underline{\gamma}$$

© University of Pretoria

with

$$\begin{aligned} E[\underline{y}] =& E[\underline{\gamma}] \\ =& \mathbf{X}\underline{\beta}, \text{ and} \\ var(\underline{y}) =& var(\underline{\gamma}) \\ =& \sigma^2 \mathbf{I}_n. \end{aligned}$$

Thus, instead of modelling the residuals, the dependent variable may be modelled along with the residuals, seeking coefficient estimates $\underline{\hat{\beta}}$ that will yield a characteristic that is equal for both the $\underline{y}$ and $\underline{\gamma}$ observations.

The assumption of normality is also redundant in this case, and a more general, non-parametric KDE, with Gaussian kernel, will be assumed for $\underline{y}$ and $\underline{\gamma}$. Thus

$$\hat{f}_y(y) = \frac{1}{nh_y} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-y_i}{h_y}\right)^2} \tag{4.1.1}$$

with

$$h_y = s_y \left(\frac{4}{3n}\right)^{1/5},$$

being the smoothing parameter suggested by Silverman (1986) [77]. Similarly, the KDE for $\underline{\gamma}$ is given by

$$\begin{aligned} \hat{f}_\gamma(\gamma) =& \frac{1}{nh_\gamma} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\gamma-\gamma_i}{h_\gamma}\right)^2} \\ =& \frac{1}{nh_\gamma} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\gamma-\sum_{j=1}^{p} x_{ij}\beta_j}{h_\gamma}\right)^2}, \end{aligned} \tag{4.1.2}$$

with

$$h_\gamma = s_\gamma \left(\frac{4}{3n}\right)^{1/5}.$$

4.2 Equating Moments

Non-parametric regression using kernel density estimators has long been proposed by Nadaraya (1964) [66], and Watson (1964) [85], and a general framework for moment

© University of Pretoria

matching by Song et al (2008) [78]. Here, it is proposed that the moments of $\underline{y}$ and $\underline{\gamma}$ be matched in order to solve for $\underline{\beta}$. Thus for $\nu \in \mathbb{N}$, it follows from (4.1.1) that

$$\begin{aligned}
E_y[y^\nu] =& \int_{-\infty}^{\infty} y^\nu \hat{f}_y(y)dy \\
=& \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} y^\nu \frac{1}{\sqrt{2\pi}h_y} e^{-\frac{1}{2}\left(\frac{y-y_i}{h_y}\right)^2} dy \\
=& \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} y^\nu \frac{1}{\sqrt{2\pi}h_y} e^{-\frac{1}{2}\left(\frac{y-y_i}{h_y}\right)^2} dy \\
=& \frac{1}{n}\sum_{i=1}^{n} m'_\nu(y_i, h_y),
\end{aligned}$$

where $m'_\nu(y_i, h_y)$ denotes the raw moments of a $N(y_i, h_y)$ variable. In particular

$$\begin{aligned}
E_y[y] =& \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \\
E_y[y^2] =& \frac{1}{n}\sum_{i=1}^{n} y_i^2 + h_y^2 \\
E_y[y^3] =& \frac{1}{n}\sum_{i=1}^{n} y_i^3 + 3y_i h_y^2,
\end{aligned}$$

and so on. Similarly, from (4.1.2), it follows that

$$\begin{aligned}
E_\gamma[\gamma^\nu] =& \int_{-\infty}^{\infty} \gamma^\nu \hat{f}_\gamma(\gamma)d\gamma \\
=& \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} \gamma^\nu \frac{1}{\sqrt{2\pi}h_\gamma} e^{-\frac{1}{2}\left(\frac{\gamma-\sum_{j=1}^{p} x_{ij}\beta_j}{h_\gamma}\right)^2} d\gamma \\
=& \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} \gamma^\nu \frac{1}{\sqrt{2\pi}h_\gamma} e^{-\frac{1}{2}\left(\frac{\gamma-\sum_{j=1}^{p} x_{ij}\beta_j}{h_\gamma}\right)^2} d\gamma \\
=& \frac{1}{n}\sum_{i=1}^{n} m'_\nu(\sum_{j=1}^{p} x_{ij}\beta_j, h_\gamma),
\end{aligned}$$

© University of Pretoria

where $m'_\nu\left(\sum_{j=1}^{p} x_{ij}\beta_j, h_\gamma\right)$ denotes the raw moments of a $N\left(\sum_{j=1}^{p} x_{ij}\beta_j, h_\gamma\right)$ variable. In particular

$$\begin{aligned}
E_\gamma[\gamma] =& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}\beta_j \\
E_\gamma[\gamma^2] =& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + h_\gamma^2 \\
E_\gamma[\gamma^3] =& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)^3 + 3\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) h_\gamma^2,
\end{aligned}$$

and so on.

An objective function may then be constructed as the minimum squared Euclidean distance between the first p moments of the kernel densities of y and γ, i.e.

$$\underline{\hat{\beta}} = \min_{\underline{\beta}} \sum_{\nu=1}^{p} \left(E_y[y^\nu] - E_\gamma[\gamma^\nu]\right)^2 . \tag{4.2.1}$$

For instance, for a model having an intercept term, and two independent variables, the following three equations are implicitly solved

$$\begin{aligned}
E_y[y] =& E_\gamma[\gamma] \\
\therefore \sum_{i=1}^{n} y_i =& \sum_{i=1}^{n}\sum_{j=1}^{3} x_{ij}\beta_j && (4.2.2) \\
E_y[y^2] =& E_\gamma[\gamma^2] \\
\therefore \sum_{i=1}^{n}\left[y_i^2 + h_y^2\right] =& \sum_{i=1}^{n}\left[\left(\sum_{j=1}^{3} x_{ij}\beta_j\right)^2 + h_\gamma^2\right] && (4.2.3) \\
E_y[y^3] =& E_\gamma[\gamma^3] \\
\therefore \sum_{i=1}^{n}\left[y_i^3 + 3y_i h_y^2\right] =& \sum_{i=1}^{n}\left[\left(\sum_{j=1}^{3} x_{ij}\beta_j\right)^3 + 3\left(\sum_{j=1}^{3} x_{ij}\beta_j\right) h_\gamma^2\right]. && (4.2.4)
\end{aligned}$$

The left-hand-sides of these equations are all values obtained directly from the data, but the right-hand-sides are non-linear functions in the three unknowns β_1, β_2, and β_3. Note that although a Gaussian kernel is assumed for the KDE, it is not required to actually construct this in obtaining estimates for the model.

4.3 Arc Length Regression

Instead of finding the set of $\underline{\hat{\beta}}$ values that yield

$$E[\underline{y}] = E[\underline{\gamma}]$$

(as in the OLS case), observe the arc lengths of the two kernel density functions, (4.1.1) and (4.1.2), on the interval $[a, b]$, assumed to be in both their domains

$$\mathcal{S}_y^{[a,b]} = \int\limits_a^b \sqrt{1 + \left(\hat{f}'_y(y)\right)^2}\, dy$$

$$\mathcal{S}_\gamma^{[a,b]}(\underline{\beta}) = \int\limits_a^b \sqrt{1 + \left(\hat{f}'_\gamma(\gamma)\right)^2}\, d\gamma.$$

An objective function may be constructed by partitioning the interval $[a, b]$ into p subintervals, i.e. let $a = i_1 < \ldots < i_{p+1} = b$, so that $\underline{\hat{\beta}}$ is given by the set of $\underline{\beta}$ values that simultaneously minimises the p equations

$$\mathcal{S}_y^{[i_1,i_2]} = \mathcal{S}_\gamma^{[i_1,i_2]}(\underline{\beta}) \tag{4.3.1}$$

$$\vdots\,\vdots \tag{4.3.2}$$

$$\mathcal{S}_y^{[i_p,i_{p+1}]} = \mathcal{S}_\gamma^{[i_p,i_{p+1}]}(\underline{\beta}), \tag{4.3.3}$$

or equivalently, that minimises the squared Euclidean distance between the two sets of p arc lengths:

$$\underline{\hat{\beta}} = \min_{\underline{\beta}} \sum_{j=1}^{p} \left(\mathcal{S}_y^{[i_j,i_{j+1}]} - \mathcal{S}_\gamma^{[i_j,i_{j+1}]}(\underline{\beta})\right)^2. \tag{4.3.4}$$

4.4 Implementation Infrastructure

In order to obtain solutions to (4.2.1) and (4.3.4) for the moment matching method (MMM) and ALR methods respectively, the Nelder-Mead simplex method by Nelder and Mead (1965) [67] from the "optim" function in "R" was used, with initial values resulting from OLS.

For calculation of standard errors, confidence intervals, p-values and biases, bootstrap resampling proposed by Efron (1982) [24] was implemented. These methods are computationally intensive, and for ALR, supercomputing facilities were required. For

© University of Pretoria

this, the Lengau super computer from the CHPC was used, with 90 compute nodes with 24 cores and 128 GiB memory each.

4.5 Application

Hussey et al (2017) [46] were interested in calculating the percentage variation explained in expression data captured in "Raw_FPKM" signals by two histone modifications, "H3K4me3" and "H3K27me3". Statistically speaking, this is a model with a single dependent variable, two independent variables, and an intercept. The data is summarised below:

```
       Gene_ID  Raw_FPKM  H3K4me3_signal_bin25  H3K27me3_signal_bin21
1  Eucgr.A00007        0              -16.6477               -10.1981
2  Eucgr.A00043        0               10.3002                25.5719
3  Eucgr.A00047        0              -11.8096                18.5719
4  Eucgr.A00088        0              -16.9475                -3.9692
5  Eucgr.A00090        0               -9.8643                -8.3181
6  Eucgr.A00098        0                1.6385                47.6019
```

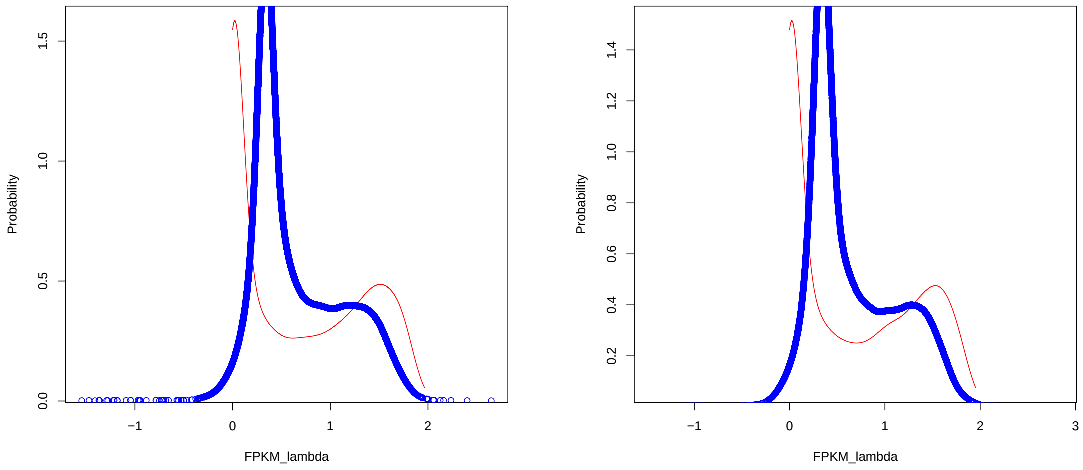
A 60%–40% split of the data yielded a training set (of 21427 observations) and testing set (of 14284 observations) respectively. The Yeo-Johnson transformation, using the "car" package by Fox and Weisberg (2011) [33], has been applied to the "Raw_FPKM" signal, with optimal choice $\lambda = -0.503013133065168$. An OLS model has been fitted to the transformed variable, and the results reproduced here.

```

Call:
lm(formula = FPKM_lambda ~ H3K4me3_signal_bin25 + H3K27me3_signal_bin21
    ,
    data = training)

Residuals:
   Min      1Q  Median      3Q     Max
-1.776  -0.329  -0.109   0.267   1.899

Coefficients:
                        Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)             0.3629992  0.0042044     86.3    <2e-16
H3K4me3_signal_bin25    0.0046198  0.0000334    138.4    <2e-16
```

© University of Pretoria

```
H3K27me3_signal_bin21 -0.0048987   0.0001261    -38.9    <2e-16

Residual standard error: 0.457 on 21424 degrees of freedom
Multiple R^2:  0.503,     Adjusted R^2:  0.503
F-statistic: 1.08e+04 on 2 and 21424 DF,  p-value: <2e-16
```

The fitted KDE's, using Gaussian kernels, with Silverman's bandwidth estimator, to the transformed "Raw_FPKM", fitted model resulting from OLS estimates, and these results applied to the testing set, along with the respective bandwidth estimators, are illustrated in Figure 4.1. The solid (red) curves represent the KDE's for the actual, and the blue curves the KDE's for the estimated signals. For the training set $r^2 = 0.50266$, and for the testing set $r^2 = 0.49657$.



(a) KDE's for transformed FPKM signal and the OLS fit thereof, with bandwidths 0.09333 and 0.06617 respectively (training set).

(b) KDE's for transformed FPKM signal and the OLS fit thereof, with bandwidths 0.10109 and 0.07233 respectively (testing set).

Figure 4.1: Plots for transformed expression data resulting from OLS.

4.5.1 Moment Matching

Solving for the three coefficients above using MMM set out in Section 4.2, requires finding solutions to (4.2.2) – (4.2.4). For this, the "mmKDEboot" function from the

© University of Pretoria

"alR" package is used in solving for $\underline{\beta}$. Here, bootstrapping is employed for estimating standard errors, confidence intervals, biases and p-values for the estimated coefficients. The results now follow.

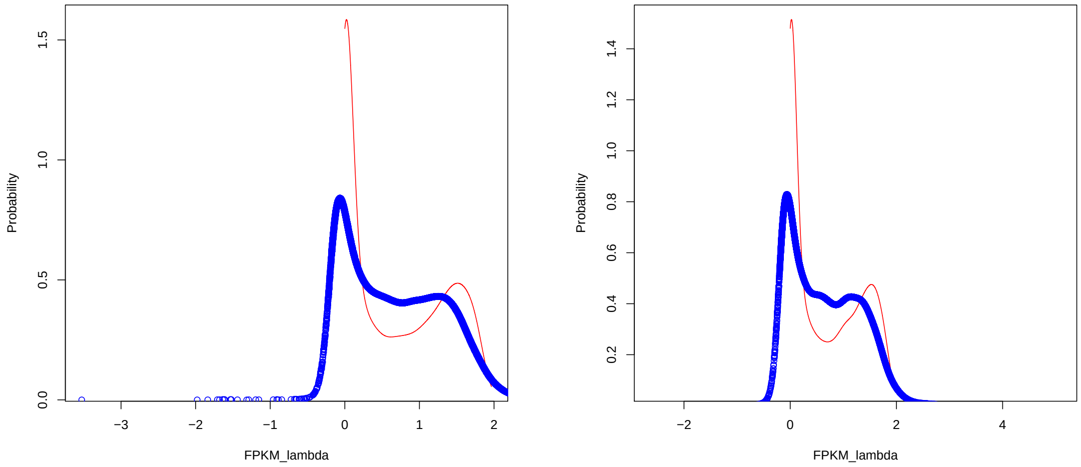
```

Residuals:
     Min.    1st Qu.     Median    3rd Qu.       Max.
-3.925800  -0.328780  0.043892  0.351300  3.527000

KDE Moments:
       LHS     RHS
1   0.66191 0.66191
2   0.86605 0.86605
3   1.24962 1.24962
BW  0.09333 0.09333

Value of objective function:
      Min.     1st Qu.      Median     3rd Qu.        Max.
0.0000e+00 5.9787e-16 9.1551e-16 1.8611e-15 5.5212e-03

                      Estimate StdErr p.value
Intercept              0.08251   0.17    0.77
H3K4me3_signal_bin25   0.00587   0.09  <2e-16
H3K27me3_signal_bin21  0.01558   0.09  <2e-16

Residual standard error: 0.686 on 21424 degrees of freedom
Multiple R^2: 0.471,     Adjusted R^2: 0.471
Value of objective function: 0.000

Elapsed time:
  min   mean   max
11490 11490 11490
```

From this, it can be seen that the MMM estimators, $\hat{\beta}$, causes the bandwidth (standard deviation) estimator for FPKM_lambda and that of the KDE cast over $\mathbf{X}\underline{\beta}$, and the values of (4.2.2) – (4.2.4), to be equal. The fitted KDE's, using Gaussian kernels, with Silverman's bandwidth estimator, to the transformed "Raw_FPKM" signal, fitted method of moments estimates, and these results applied to the testing set, along with the respective bandwidth estimators, are illustrated in Figure 4.2. For the training set

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

$r^2 = 0.47082$, and for the testing set $r^2 = 0.47088$. The r^2value as well as the first three moments for the KDE's of the testing set seem to be stable. This is depicted in the following table:

	LHS	RHS
1	0.65208	0.65158
2	0.85361	0.85480
3	1.23187	1.23051
Silverman BW	0.10109	0.10131



(a) KDE's for transformed FPKM signal and the method of moments fit thereof, with bandwidths 0.09333 and 0.09333 respectively (training set).

(b) KDE's for transformed FPKM signal and the method of moments fit thereof, with bandwidths 0.10109 and 0.10131 respectively (testing set).

Figure 4.2: Plots for transformed expression data resulting from method of moments.

4.5.2 Arc Length Matching

Solving for $\underline{\beta}$ using ALR set out in Section 4.3, requires finding solutions to (4.3.1) – (4.3.3). For this, the "alKDEboot" function from the "alR" package is used in solving for $\underline{\beta}$. Again, bootstrapping is employed for estimating standard errors, confidence intervals, biases and p-values for the estimated coefficients. The results now follow.

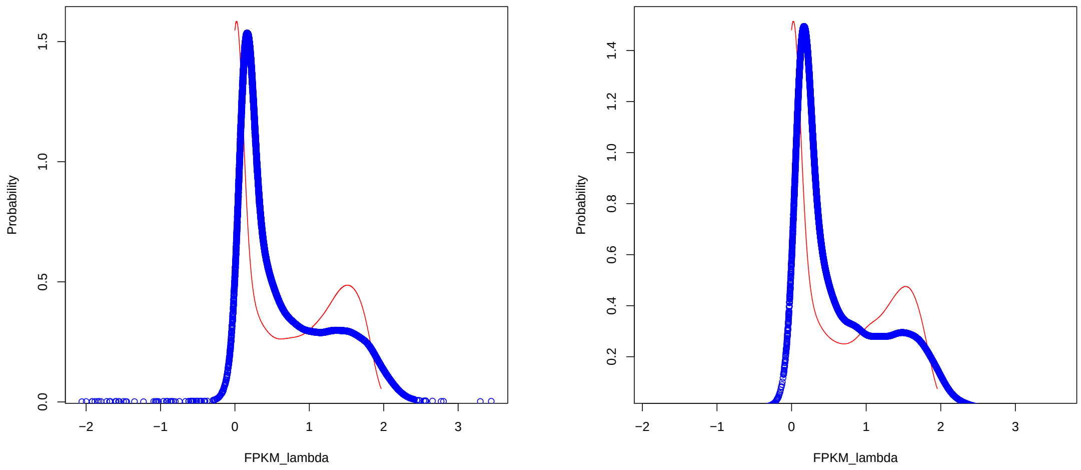
© University of Pretoria

```
Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-2.17800 -0.27935 -0.11239  0.19860  2.59920

KDE Arc Lengths:
                        LHS      RHS
[-0.12825, 0.43317] 2.50550 2.505496
[0.43317, 1.75541]  1.41384 1.413842
BW                  0.09333 0.087619

Value of objective function:
      Min.    1st Qu.     Median    3rd Qu.       Max.
0.0000e+00 4.4409e-16 6.6613e-16 1.3506e-15 1.5469e-01

                      Estimate StdErr p.value
Intercept              0.21554   0.16  <2e-16
H3K4me3_signal_bin25   0.00641   0.11  <2e-16
H3K27me3_signal_bin21 -0.00231   0.11    0.19

Residual standard error: 0.489 on 21424 degrees of freedom
Multiple $R^2$: 0.607,      Adjusted $R^2$: 0.607
Value of objective function: 0.000

Elapsed time:
   min   mean    max
1377.5 1380.6 1381.9
```

The ALR estimators, $\hat{\underline{\beta}}$, do not necessarily cause the bandwidth estimator for the dependent observations, "FPKM_lambda", and that of the KDE cast over $\mathbf{X}\underline{\beta}$ to be equal. The fitted KDE's, using Gaussian kernels, with Silverman's bandwidth estimator, to the transformed "Raw_FPKM" signal, fitted ALR estimates, and these results applied to the testing set, along with the respective bandwidth estimators, are illustrated in Figure 4.3. For the training set $r^2 = 0.60716$, and for the testing set $r^2 = 0.63292$. The r^2value as well as the arc length segments for the KDE's of the testing set seem to be stable. This is depicted in the following table:

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

	LHS	RHS
$[-0.1409, 0.40889]$	2.36790	2.37996
$[0.40889, 1.75679]$	1.44474	1.47711
BW	0.10109	0.09567



(a) KDE's for transformed FPKM signal and the arc length regression fit thereof, with bandwidths 0.09333 and 0.08762 respectively (training set).

(b) KDE's for transformed FPKM signal and the arc length regression fit thereof, with bandwidths 0.10109 and 0.09567 respectively (testing set).

Figure 4.3: Plots for transformed expression data resulting from arc length regression.

4.6 Model Evaluation

The parameter estimates, along with their properties for the three models fitted to the data, are summarised in Table 4.1. For assessing the respective fit of the models proposed, the r^2 values are compared, including a r^2 resulting from cross-validation, as well as D^2, a measure of divergence between the $\underline{y}$ and $\mathbf{X}\underline{\hat{\beta}}$ vectors. This test is based on the assumption that the two vectors have a common distribution, and similar size, and was proposed by Bhattacharyya (1946) [8]. This assumption seems reasonable, since $\underline{\hat{\beta}}$ is the set of parameters that provide equal properties for the KDE's of the two vectors. Since the sample size in this case is "large", 100 proportions were compared, and consequently, the χ^2 approximation is adequate for approximating the exact distribution derived by Bhattacharyya. The results are shown in Table 4.2.

	Intercept	H3K4me3_signal_bin25	H3K27me3_signal_bin21
OLS			
Estimate	0.3630	0.0046	-0.0049
StdErr	0.0042	0.0000	0.0001
LCI	0.3548	0.0046	-0.0051
UCI	0.3712	0.0047	-0.0047
p.value	0.0000	0.0000	0.0000
MMM			
Estimate	0.0825	0.0059	0.0156
StdErr	0.1711	0.0935	0.0948
LCI	0.0749	0.0056	-0.0110
UCI	0.3114	0.0062	0.0168
b.value	0.1679	0.0060	0.0052
Bias	0.0854	0.0001	-0.0103
p.value	0.7708	0.0000	0.0000
ALR			
Estimate	0.2155	0.0064	-0.0023
StdErr	0.1603	0.1069	0.1105
LCI	0.0859	0.0056	-0.0041
UCI	0.2549	0.0066	0.0018
b.value	0.1912	0.0061	-0.0000
Bias	-0.0244	-0.0003	0.0023
p.value	0.0000	0.0000	0.1939

Table 4.1: Comparison of bootstrap results for OLS, MMM and ALR methods.

From Table 4.1 it can be seen that the parameter standard errors resulting from OLS are smaller than that of the two newly proposed methods. This is in accordance with the properties of OLS estimates, being those with the smallest variance. The "b.values" is the estimated values with the bias removed, however, these do not correspond to the respective objective functions being minimised. Only the OLS method resulted

© University of Pretoria

	OLS	MMM	ALR
r2 (Train)	0.5027	0.4708	0.6072
r2 (Test)	0.4966	0.4709	0.6329
RSS	0.4566	0.6865	0.4890
D2 (Train)	0.3239	0.2566	0.2654
D2 (Test)	0.3294	0.2575	0.2710

Table 4.2: Comparison of performance results for OLS, MMM and ALR methods.

in significant (at the 5% level) p-values for all parameters; however, these are based on strong underlying model assumptions, whereas the other two methods are non-parametric in nature.

From Table 4.2 it is clear that OLS yielded the smallest residual standard error (indicated by "RSS"), with ALR resulting in the largest r^2 values. The MMM resulted in the smallest divergences, closely followed by ALR. The r^2 and $D2$ measures kept steady under cross-validation, and therefore signify stable results. Although the p-values of the Bhattacharyya test are all very small (< 0.0001 not given here), the divergence measures on which the test is based, may still be used for comparison purposes.

4.7 Summary

Two new regression techniques were proposed here based on equating properties of KDE's, based on the set of dependent observations $\underline{y}$, and the set of independent observations $\mathbf{X}\underline{\beta}$. Both these methods didn't tamper with the linear model configuration, but instead proposed new objective functions. Standard errors, confidence intervals and p-values were obtained using the bootstrap, and the results obtained compared to that obtained from OLS regression.

It was seen that although OLS regression resulted in the smallest residual standard errors, ALR could predict a larger percentage explained variation as measured by r^2, but that the MMM, had the smallest divergence between the observed, and predicted value distributions. ALR therefore balances the residual standard error, and sample divergence, while resulting in a larger r^2.

In the following chapter, the linear model specification will be generalised, to allow for a non-linear regression setting. It will be seen how the arc length may be utilised in

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

parameter estimation under such assumptions.

© University of Pretoria

Chapter 5

Non-Linear Arc Length Regression

The linear model assumption from Chapter 4 is extended here to the non-linear setting, and the comparison to OLS, naturally replaced by NLS. In this chapter, a NALR technique is proposed, and used to fit the cumulative distribution function (CDF), of a subset of the four-parameter kappa distribution, to time-conductivity curves.

In Section 5.1, the usual model specification for non-linear regression, along with NLS estimation are presented. NALR is then introduced by modification of the NLS objective function.

It is shown that the practical problem in Section 5.2, may be cast in the form of a typical non-linear regression setting, solved in Section 5.3, using a subset of the four-parameter kappa CDF. This is obtained by enforcing a particular relationship between its two shape parameters. Finally, the NLS and NALR techniques are demonstrated in Section 5.4, using data resulting from actual experiments.

5.1 Introduction

Consider n pairs of observations, (x_1, y_1), (x_2, y_2), . . ., (x_n, y_n) for which

$$y_i = F(x_i, \underline{\beta})$$

describes the error-free relationship between a set of dependent and independent observations $\underset{n\times 1}{\underline{y}}$, and $\underset{n\times 1}{\underline{x}}$, respectively. Here, $F(\cdot)$ is a non-linear function in the p parameters $\underset{p\times 1}{\underline{\beta}}$, say.

© University of Pretoria

In NLS, the estimated parameters $\underline{\hat{\beta}}$, is the set of parameter values that minimises the objective function

$$\text{SSE} = \sum \left(\underline{y} - F(\underline{x}, \underline{\hat{\beta}})\right)^2.$$

Define

$$\mathcal{S}(a, b, \underline{\beta}) = \int_a^b \sqrt{1 + \left(f(x, \underline{\beta})\right)^2}dx$$

as the arc length of the function $F(x, \underline{\beta})$ on the interval $[a, b]$. It is assumed that

$$f(x, \underline{\beta}) = F'(x, \underline{\beta})$$

exists and is well defined. Also define

$$\smallint(a, b) = \sum_{i=1}^{n-1} \sqrt{\left(x_{(i+1)} - x_{(i)}\right)^2 + \left(y_{(i+1)} - y_{(i)}\right)^2} \tag{5.1.1}$$

to be the sample arc length statistic. Here, $(x_{(i)}, y_{(i)})$ denotes the observed pairs, ordered in $\underline{x}$, i.e. $x_{(1)} \leq \ldots \leq x_{(n)}$. This is done in order to honour the relationship between $\underline{x}$ and $\underline{y}$. The estimated parameters $\underline{\hat{\beta}}$ is the set of parameters that minimises the objective function

$$\left(\mathcal{S}(a, b, \underline{\hat{\beta}}) - \smallint(a, b)\right)^2.$$

The interval $[a, b]$ may be thought of as tuning parameters, and may even correspond to vectors, so that the arc lengths over a collection of intervals are matched. The objective function for a collection of n arc length segments then becomes

$$\sqrt{\sum_{i=1}^{n} \left(\mathcal{S}(a_i, b_i, \underline{\hat{\beta}}) - \smallint(a_i, b_i)\right)^2}.$$

The optimisation process is therefore similar to that of NLS, with a change in only the objective function.

5.2 Problem Description

Currently the reference method for the measurement of the oxidative stability of biodiesel is the EUROPEAN STANDARD EN 14112 (2003) [27]. It prescribes the Rancimat method by Hadorn (1974) [39] and Läubli and Bruttel (1986) [51] for measuring the

© University of Pretoria

induction time of the biodiesel. In this method autoxidation is accelerated by passing a constant flow of air through the biodiesel sample and controlling the temperature at an elevated level, i.e. 110 $°C$. The oxidation process is driven by radical reactions that involve the unsaturated fatty acids. During an initial induction phase virtually no secondary products are formed. This is abruptly followed by an oxidation phase characterised by a rapid increase in peroxide value and the formation of volatile products. The Rancimat method relies on the fact that the greater part of these volatile matter consists of formic acid. This is trapped by passing the air through distilled water and its accumulation is recorded conductometrically. The length of the induction period (IP) is taken as a measure of oxidative stability. The EUROPEAN STANDARD EN 14214 (2008) [28] describes two methods for the evaluation of the IP. Furthermore there is a tacit assumption that the two methods yield comparable if not identical results.

Both these methods utilises the conductivity vs. time curve, say $F(x)$. Focke, van der Westhuizen and Oosthuizen (2016) [32] showed that this process can be automated by curve fitting using a range of sigmoidal functions, however, that analysis demonstrated that the IP values resulting from the two methods differ. The question now is to establish conditions under which the two methods yield equivalent results.

Let X be a unimodal random variable with CDF and PDF respectively $F(X)$ and $f(x)$. By definition

$$f(x) = F'(x).$$

The mode of X is attained at

$$f'(x) = F''(x) = 0,$$

say at c, i.e. the inflection point of $F(x)$ or global maximum of $f(x)$.

For the first method, $F'(x) = f(x)$ is the gradient of $F(x)$ at any point x. The formula for the gradient of the straight line at the point c is given by $f(c)$. c corresponds to the mode of X, and may be found as the inflection point $(c, 0)$, say, of $F(x)$, i.e. where

$$F''(c) = 0. \tag{5.2.1}$$

The formula for the straight line, $T(x)$, through the point $(c, F(c))$, with gradient $f(c)$

© University of Pretoria

is therefore given by

$$\begin{aligned} F(c) =& f(c)c + c_0 \\ \therefore c_0 =& F(c) - f(c)c \\ \therefore T(x) =& f(c)\,(x-c) + F(c). \end{aligned} \tag{5.2.2}$$

The IP value is then the point $(a, 0)$ where this intersects the x-axis (see EUROPEAN STANDARD EN 14214 (2008) [28]), and is given by

$$\begin{aligned} T(a) =& 0 \\ \therefore 0 =& f(c)\,(a-c) + F(c) \\ \therefore a =& c - \frac{F(c)}{f(c)}. \end{aligned} \tag{5.2.3}$$

For the second method, suppose that the IP value is $(b, 0)$ which is the smallest solution of

$$F'''(b) = 0, \tag{5.2.4}$$

(see EUROPEAN STANDARD EN 14214 (2008) [28]).

In addressing the research question, a CDF, $F(x)$, is therefore sought that satisfies

$$\begin{aligned} a =& c - \frac{F(c)}{f(c)} \\ =& b, \end{aligned}$$

or if $a = b = d$, say, then from (5.2.1), (5.2.3), and (5.2.4):

$$\begin{aligned} F''(c) =& F'''(d) \\ =& F'''\left(c - \frac{F(c)}{f(c)}\right). \end{aligned}$$

As an example, consider the standard normal distribution, which has mode $c = 0$. For method 1, the line through the point $(c, F(c)) = (0, 0.5)$, with gradient $f(c) = f(0) = \frac{1}{\sqrt{2\pi}}$ is given by

$$T(x) = \frac{x}{\sqrt{2\pi}} + 0.5,$$

with corresponding IP value using (5.2.3)

$$a = -\sqrt{\frac{\pi}{2}} = -1.2533141373.$$

© University of Pretoria

For the second method, the calculations reduce to

$$\begin{aligned} F'''(x) =& f''(x) \\ f'(x) =& -xf(x) \\ \therefore f''(x) =& -f(x) - xf'(x) \\ =& -f(x) + x^2 f(x) \\ =& f(x)\left(x^2-1\right) \\ \therefore F'''(x) =& 0 \\ \Leftrightarrow 0 =& f(x)\left(x^2-1\right) \\ \therefore x =& \pm 1. \end{aligned}$$

b is now the smallest of these two roots, namely $b = -1$.

Thus, a corresponds to the point $(-1.2533141373, 0)$ which is less than b corresponding to the point $(-1, 0)$. Clearly the standard normal distribution isn't a solution to this problem. This solution is depicted in Figure 5.1.

© University of Pretoria

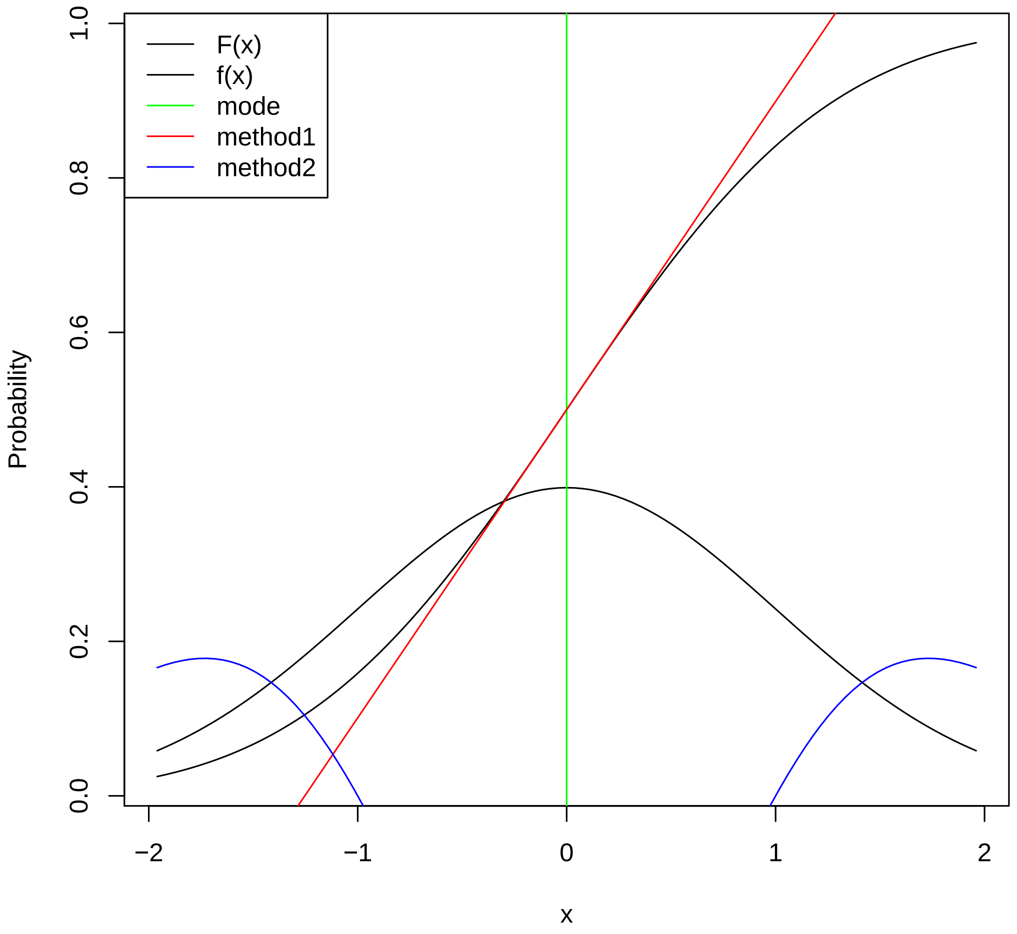


Figure 5.1: Illustration of the two methods for the evaluation of the IP.

In summary then, taking into account (5.2.1), (5.2.3), and (5.2.4), a suitable probability distribution will exhibit the following properties:

- $f''\left(c - \frac{F(c)}{f(c)}\right) = 0$, where
- $F''(c) = 0$, i.e. c is the mode of the distribution.

© University of Pretoria

5.3 The Four-Parameter Kappa Distribution

The four-parameter kappa distribution was introduced by Hosking (1994) [45] as a generalisation of the three-parameter version, introduced by Mielke Jr (1973) [60], by making use of a suitable re-parameterisation. In order to simplify notation, let

$$\begin{aligned} g(x) =& 1 - k\frac{x-\mu}{\sigma} \\ \therefore g'(x) =& -\frac{k}{\sigma}, \end{aligned}$$

then the quantile function (QF), CDF, and PDF are respectively given by

$$Q(p) = \mu + \frac{\sigma}{k}\left(1 - \left(\frac{1-p^h}{h}\right)^k\right), \; 0 \leq p \leq 1,$$

$$\begin{aligned} F(x) =& \left(1 - h\left(1 - k\frac{x-\mu}{\sigma}\right)^{1/k}\right)^{1/h} \\ =& \left(1 - h\left(g(x)\right)^{1/k}\right)^{1/h} \end{aligned}$$

$$\begin{aligned} f(x) =& -\frac{1}{k}F(x)/\left(1 - h\left(g(x)\right)^{1/k}\right)\left(g(x)\right)^{1/k-1}g'(x) \\ =& \frac{1}{\sigma}F(x)/\left(1 - h\left(g(x)\right)^{1/k}\right)\left(g(x)\right)^{1/k-1} \\ =& \frac{1}{\sigma}\left(1 - h\left(g(x)\right)^{1/k}\right)^{1/h-1}\left(g(x)\right)^{1/k-1} \\ =& \frac{1}{\sigma}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-1}. \end{aligned}$$

Here $\mu \in \mathbb{R}$ is a location parameter, $\sigma > 0$ is a scale parameter, and $h, k \in \mathbb{R}$ are shape parameters. The notation used here is ment to include the limiting cases for $h = 0$ and

$k = 0$. The first derivative is then

$$\begin{aligned}
f'(x) =&(1-h)\left(F(x)\right)^{-1}\left(f(x)\right)^{2} - \frac{1-k}{\sigma}f(x)\left(g(x)\right)^{-1}\\
=&f(x)\left[(1-h)\left(F(x)\right)^{-1}f(x) - \frac{1-k}{\sigma}\left(g(x)\right)^{-1}\right]\\
=&f(x)\left[(1-h)\left(F(x)\right)^{-1}\frac{1}{\sigma}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-1} - \frac{1-k}{\sigma}\left(g(x)\right)^{-1}\right]\\
=&\frac{1}{\sigma}f(x)\left(g(x)\right)^{-1}\left[(1-h)\left(F(x)\right)^{-h}\left(g(x)\right)^{1/k} - (1-k)\right]\\
=&\frac{1}{\sigma}f(x)\left(g(x)\right)^{-1}\left[(1-h)\left(1-h\left(g(x)\right)^{1/k}\right)^{-1}\left(g(x)\right)^{1/k} - (1-k)\right]\\
=&\frac{1}{\sigma}f(x)\left(g(x)\right)^{-1}\left[(1-h)\left(\left(g(x)\right)^{-1/k} - h\right)^{-1} - (1-k)\right].
\end{aligned}$$

The mode is therefore given by

$$\begin{aligned}
f'(x) =&0\\
\Leftrightarrow \left(\frac{1-k}{1-hk}\right)^{k} =&g(x)\\
\therefore x =&\mu + \frac{\sigma}{k}\left(1-\left(\frac{1-k}{1-hk}\right)^{k}\right) = c, \text{ say.}
\end{aligned} \tag{5.3.1}$$

© University of Pretoria

The second derivative is given by

$$\begin{aligned}
f''(x) =&(1-h)\left[-\left(F(x)\right)^{-2}\left(f(x)\right)^{3}+2\left(F(x)\right)^{-1}f(x)f'(x)\right]\\
&-\frac{1-k}{\sigma}\left[f'(x)\left(g(x)\right)^{-1}-f(x)\left(g(x)\right)^{-2}g'(x)\right]\\
=&(1-h)\left(F(x)\right)^{-1}f(x)\left[2f'(x)-\left(F(x)\right)^{-1}\left(f(x)\right)^{2}\right]\\
&-\frac{1-k}{\sigma}\left(g(x)\right)^{-1}\left[f'(x)+\frac{k}{\sigma}f(x)\left(g(x)\right)^{-1}\right]\\
=&(1-h)\left(F(x)\right)^{-1}\frac{1}{\sigma}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-1}\\
&\times\left[2f'(x)-\left(F(x)\right)^{-1}\left(\frac{1}{\sigma}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-1}\right)^{2}\right]\\
&-\frac{1-k}{\sigma}\left(g(x)\right)^{-1}\left[f'(x)+\frac{k}{\sigma}\frac{1}{\sigma}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-1}\left(g(x)\right)^{-1}\right]\\
=&\frac{1-h}{\sigma}\left(F(x)\right)^{-h}\left(g(x)\right)^{1/k-1}\\
&\times\left[2f'(x)-\frac{1}{\sigma^{2}}\left(F(x)\right)^{1-2h}\left(g(x)\right)^{2/k-2}\right]\\
&-\frac{1-k}{\sigma}\left(g(x)\right)^{-1}\left[f'(x)+\frac{k}{\sigma^{2}}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-2}\right]\\
=&\left(g(x)\right)^{-1}\left\{\frac{1-h}{\sigma}\left(F(x)\right)^{-h}\left(g(x)\right)^{1/k}\right.\\
&\left[2f'(x)-\frac{1}{\sigma^{2}}\left(F(x)\right)^{1-2h}\left(g(x)\right)^{2/k-2}\right]\\
&\left.-\frac{1-k}{\sigma}\left[f'(x)+\frac{k}{\sigma^{2}}\left(F(x)\right)^{1-h}\left(g(x)\right)^{1/k-2}\right]\right\}.
\end{aligned}$$

Since this final expression cannot be simplified further, and since its roots cannot be found analytically, numerical solutions will be employed.

If the value for c given in (5.3.1), is substituted into (5.2.3), it follows that

$$\begin{aligned}
g(c) =&\left(\frac{1-k}{1-hk}\right)^{k}\\
F(c) =&\left(\frac{1-h}{1-hk}\right)^{1/h}\\
f(c) =&\frac{1}{\sigma}\left(\frac{1-h}{1-hk}\right)^{1/h-1}\left(\frac{1-k}{1-hk}\right)^{1-k}
\end{aligned}$$

© University of Pretoria

$$\begin{aligned}
\therefore \frac{F(c)}{f(c)} =& \sigma \left(\frac{1-h}{1-hk}\right)^{1/h} \left(\frac{1-h}{1-hk}\right)^{1-1/h} \left(\frac{1-k}{1-hk}\right)^{k-1} \\
\frac{F(c)}{f(c)} =& \sigma \frac{1-h}{1-k} \left(\frac{1-k}{1-hk}\right)^{k} \\
&\therefore c - \frac{F(c)}{f(c)} \\
&= \mu + \frac{\sigma}{k} - \sigma \left(\frac{1-k}{1-hk}\right)^{k} \left(\frac{1}{k} + \frac{1-h}{1-k}\right) \\
&= \mu + \frac{\sigma}{k} - \sigma \left(\frac{1-k}{1-hk}\right)^{k} \frac{1}{k} \left(\frac{1-k}{1-hk}\right)^{-1} \\
&= \mu + \frac{\sigma}{k} \left[1 - \left(\frac{1-k}{1-hk}\right)^{k-1}\right]. \qquad (5.3.2)
\end{aligned}$$

All of the derivatives of $F(\cdot)$ may be written in terms of $g(\cdot)$ (using (5.3.1)), and since the substitution of (5.3.2) into $g(\cdot)$ will yield a result that is independent of the choice of μ or σ, (h, k) combinations can be calculated numerically, such that

$$f''\left(c - \frac{F(c)}{f(c)}\right) = 0.$$

In order for the four-parameter kappa distribution to be a suitable solution to the problem, the "alR" package in "R" by Loots (2017) [54] is used for obtaining (h, k) combinations that yield suitable shape parameters. The Nelder-Mead simplex method, by Nelder and Mead (1965) [67], is used in order to minimise

$$\left| f''\left(c - \frac{F(c)}{f(c)}\right) \right|. \qquad (5.3.3)$$

Since a distribution is sought which has a global maximum of the PDF at c, it is noted from Hosking (1994) [45] that $f(x)$ has a single maximum when

$$h \in (-\infty, 0) \text{ and } k \in \left(\frac{1}{h}, 1\right), \text{ or}$$
$$h \in [0, 1) \text{ and } k \in (-\infty, 1).$$

Thus, $h \in (-\infty, 1)$, and $k \in (-\infty, 1)$. Furthermore, since the variable of interest is time, the distribution should have positive support, and should be bounded below. Following from the optimisation procedure, the following configurations are possible:

© University of Pretoria

Bound	Support	Condition
1	$x \in \left[\mu + \sigma\frac{1-h^{-k}}{k}, \mu + \frac{\sigma}{k}\right]$	$h > 0, k > 0$
2	$x \in [\mu + \sigma \ln h, \infty)$	$h > 0, k = 0$
3	$x \in \left[\mu + \sigma\frac{1-h^{-k}}{k}, \infty\right)$	$h > 0, k < 0$
4	$x \in \left(-\infty, \mu + \frac{\sigma}{k}\right]$	$h \leq 0, k > 0$
5	$x \in (-\infty, \infty)$	$h \leq 0, k = 0$
6	$x \in \left[\mu + \frac{\sigma}{k}, \infty\right)$	$h \leq 0, k < 0$

Table 5.1: (h, k) combinations for the four-parameter kappa distribution.

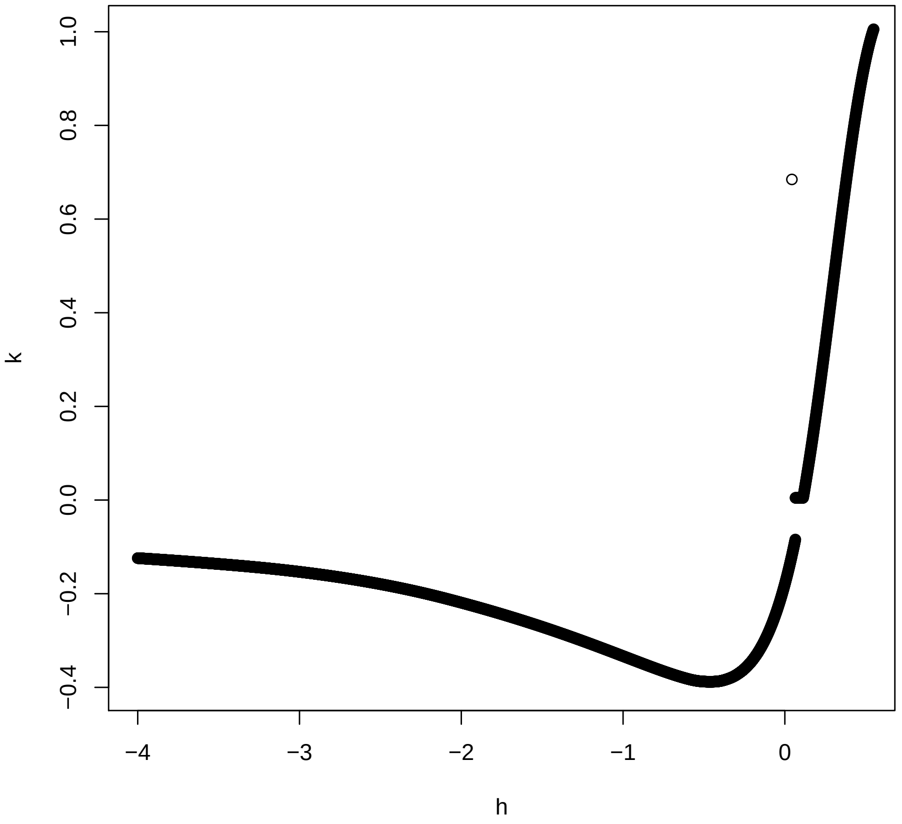


Figure 5.2: (h, k) combinations for the four-parameter kappa distribution that minimise (5.3.3).

© University of Pretoria

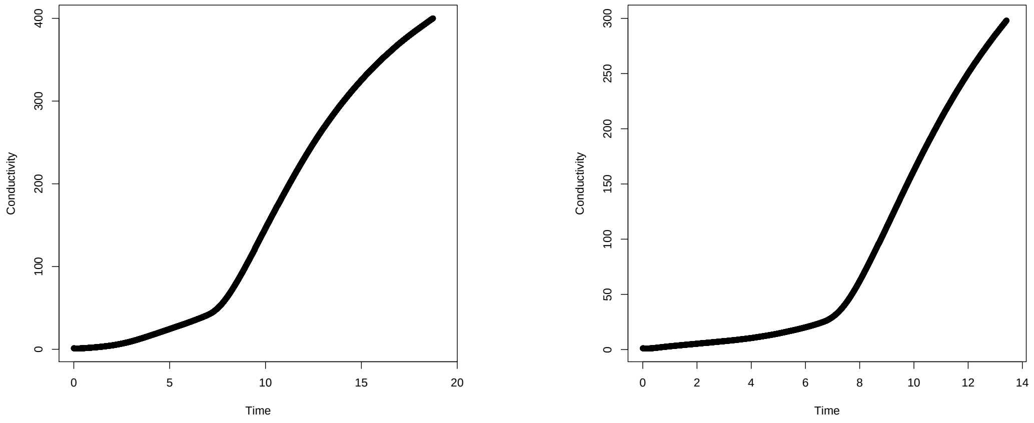
From Figure 5.2, this is a decreasing function of k for $h < -0.05$, and increasing when $h \geq -0.05$. There is a discontinuity at $h = 0$, which corresponds to a generalised extreme value distribution (GEV). $h < 0$ and $k < 0$ yield a Burr type III distribution. A special case of this is when $h = -1$ which renders a generalised logistic distribution. Bound 6 from Table 5.1 corresponds to this configuration, i.e. distributions bounded below, but not from above. $h > 0$ and $k > 0$ refer to distributions between the generalised extreme value, and generalised Pareto distributions. For this, Bound 1 from Table 5.1 applies, which is bounded above and below. For more information on these special cases see Hosking (1994) [45].

The relationship between h and k clearly indicates that none of the other special cases such as the exponential ($h = 1$ and $k = 0$), Gumbel ($h = 0$ and $k = 0$), logistic ($h = -1$ and $k = 0$), uniform ($h = 1$ and $k = 1$), or reverse exponential ($h = 0$ and $k = 1$) distributions are candidates for being a solution to the problem at hand. As noted above, from Table 5.1, Bound 2 – 4 seems unlikely to be observed, however, Bound 5, with infinite support, never occurs, since $k = 0$ is observed when $h > 0$.

5.4 Results

Two samples of sizes $n_1 = 5770$, and $n_2 = 4080$ are considered respectively. This correspond to two experiments at 0.33%, 0%, and 0.67% concentrations of antioxidants Orox PK, Naugard P and Anox 20 respectively. The sigmoidal (or time vs. conductivity) curves are shown in Figure 5.3.

© University of Pretoria



(a) Sample1.

(b) Sample2.

Figure 5.3: Time vs. conductivity curves.

In order to fit a CDF to this data, the data should be rescaled so that $\underline{y} \in [0,1]$, i.e. all $\underline{y}$ values have to be divided by

$$\tau(\underline{\beta}) = F(\max \underline{x}, \underline{\beta}).$$

This truncation may be considered as a dynamic process of scaling, since it depends on the parameter vector $\underline{\beta}$, which is to be estimated, and has no effect on the mode of the distribution, nor on the conditions for the method at hand (as derived in Section 5.3).

For fitting the CDF of the (right-truncated) four-parameter kappa distribution to data, two boundary specifications are considered separately, for $h \leq 0$ and $h > 0$, corresponding to Bound 1 and 6 from Table 5.1. This is done, in order to eliminate the estimation of μ. Since the data is bounded from below, say at x_0, μ is chosen such that:

$h < 0$

$$\begin{aligned} x_0 =& \mu + \frac{\sigma}{k} \\ \therefore \mu =& x_0 - \frac{\sigma}{k}, \end{aligned}$$

$h > 0$

$$\begin{aligned} x_0 =& \mu + \sigma \frac{1-h^{-k}}{k} \\ \therefore \mu =& x_0 - \sigma \frac{1-h^{-k}}{k} \end{aligned}$$

© University of Pretoria

The case where $h = 0$ and $k < 0$ will be considered if it seems that the estimate for h converges to zero.

5.4.1 Implementation Infrastructure

The “DEoptimR” package in “R” by Conceicao and Maechler (2015) [15] provides support for differential evolution with non-linear constraints, and was used to obtain NLS and NALR estimates, since it provides global solutions to the respective objective functions. These solutions were consequently used as starting values for the “constrOptim.nl” function from the “alabama” package by Varadhan (2015) [82]. This ensures that bootstrap solutions converge to a neighbourhood containing the global solution, and also increases the calculations involved.

5.4.2 Sample1

	mu	sigma	h	k
		NLS		
Estimate	12.4186	4.7450	-0.3632	-0.3821
StdErr	0.0295	0.0225	0.0053	0.0009
LCI	12.3776	4.7226	-0.3729	-0.3849
UCI	12.4855	4.8051	-0.3531	-0.3816
b.value	12.4305	4.7632	-0.3627	-0.3833
Bias	0.0119	0.0182	0.0005	-0.0012
p.value	0.0000	0.0000	0.0000	0.0000
		NALR		
Estimate	10.8447	4.1225	-0.7172	-0.3801
StdErr	2.3827	0.5250	0.2182	0.0268
LCI	6.7586	2.6503	-1.5606	-0.3923
UCI	17.8021	4.5204	-0.5622	-0.2712
b.value	10.5807	3.6665	-0.8978	-0.3556
Bias	-0.2640	-0.4560	-0.1806	0.0245
p.value	0.0000	0.0000	0.0330	0.0000

Table 5.2: Comparison of bootstrap results for NLS and NALR methods for sample1.

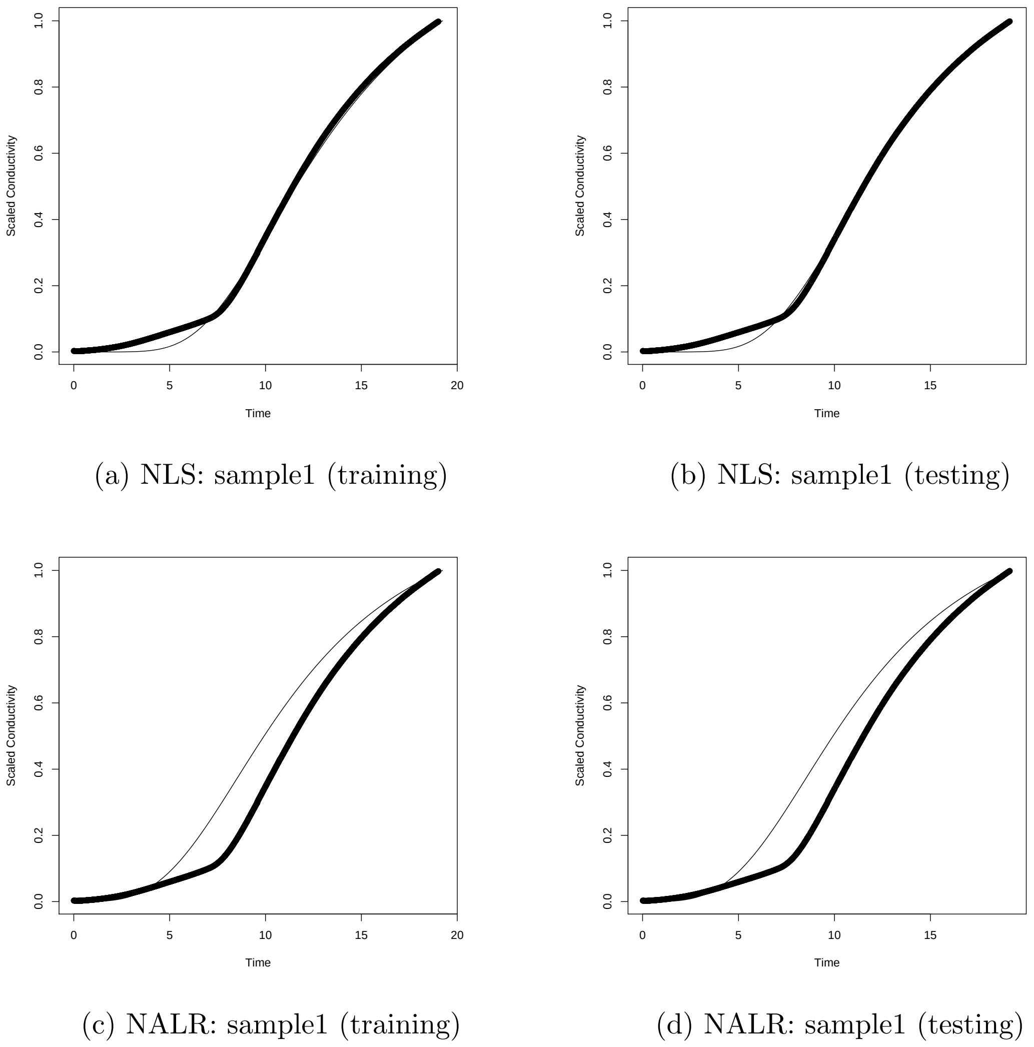
© University of Pretoria

	NLS	NALR
r2 (Train)	0.9977	0.9629
r2 (Test)	0.9980	0.9653
RSS	1.1054	5.8181
rMSE	0.0187	0.0984
Zero	1.0000E-05	-6.9690E-06
τ	0.7399	0.8142
IP (7.55)	6.4590	4.4802

Table 5.3: Comparison of performance results for NLS and AL methods for sample1.

Table 5.2 suggests that a four-parameter kappa distribution with parameters $\mu = 12.4185870463$, $\sigma = 4.744973998$, $h = -0.3632040805$, and $k = -0.3820864628$, implying a Burr type III distribution, fits the data the best, as obtained by NLS estimation. Although the p-values obtained from NALR, (with the arc length computed using the endpoints of the interval corresponding to the 0.7^{th} and 0.982^{nd} empirical quantiles of the data) are all significant at the $\alpha = 0.05$ level, the coefficient standard errors, biases, and confidence intervals for NLS estimation are smaller. Table 5.3 supports this fact by showing that the corresponding residual standard error (RSS) and rMSE values for NLS estimation are smaller than that of NALR. The r^2 values hold steady for both methods in the validation sample, and therefore provides evidence of a reasonably good fit. The "Zero" row indicates that the constrained optimisation is satisfied for both methods. The model suggested by the NLS method truncates the process at $\tau = 0.7399284597$, and NALR at $\tau = 0.8141514862$, i.e. further along. The IP value corresponding to the NLS method of 6.4590059957, is closer to the automated machine produced value of 7.55, than the NALR value of 4.4802030506. As a check of the validity of the results, the mode of the NLS model is attained at 10.3653615566, and that of the NALR model at 8.5010505495, which are both larger than the corresponding IP values, but smaller than the maximum value of 19.23243667.

© University of Pretoria



(a) NLS: sample1 (training)

(b) NLS: sample1 (testing)

(c) NALR: sample1 (training)

(d) NALR: sample1 (testing)

Figure 5.4: Fitted sigmoidal curves to sample1.

5.4.3 Sample2

Table 5.4 suggests that a four-parameter kappa distribution with parameters $\mu = 13.3572858936$, $\sigma = 5.0933422015$, $h = -0.3529822585$, and $k = -0.3813156536$ fits the data the best, as obtained by NALR. Here the endpoints of the interval over which the arc length was computed, corresponded to the 0.3^{th} and 0.63^{th} empirical quantiles of the data. Although the p-values obtained from both NLS and NALR are all significant, the coefficient standard errors, biases, and confidence intervals for NALR are smaller. Table 5.5 shows

© University of Pretoria

	mu	sigma	h	k
		NLS		
Estimate	12.4186	4.7450	-0.3632	-0.3821
StdErr	0.0295	0.0225	0.0053	0.0009
LCI	12.3776	4.7226	-0.3729	-0.3849
UCI	12.4855	4.8051	-0.3531	-0.3816
b.value	12.4305	4.7632	-0.3627	-0.3833
Bias	0.0119	0.0182	0.0005	-0.0012
p.value	0.0000	0.0000	0.0000	0.0000
		NALR		
Estimate	13.3573	5.0933	-0.3530	-0.3813
StdErr	0.0227	0.0000	0.0022	0.0006
LCI	13.3197	5.1258	-0.3581	-0.3852
UCI	13.4390	5.1259	-0.3533	-0.3816
b.value	13.4175	5.1259	-0.3554	-0.3821
Bias	0.0602	0.0325	-0.0024	-0.0008
p.value	0.0000	0.0000	0.0000	0.0000

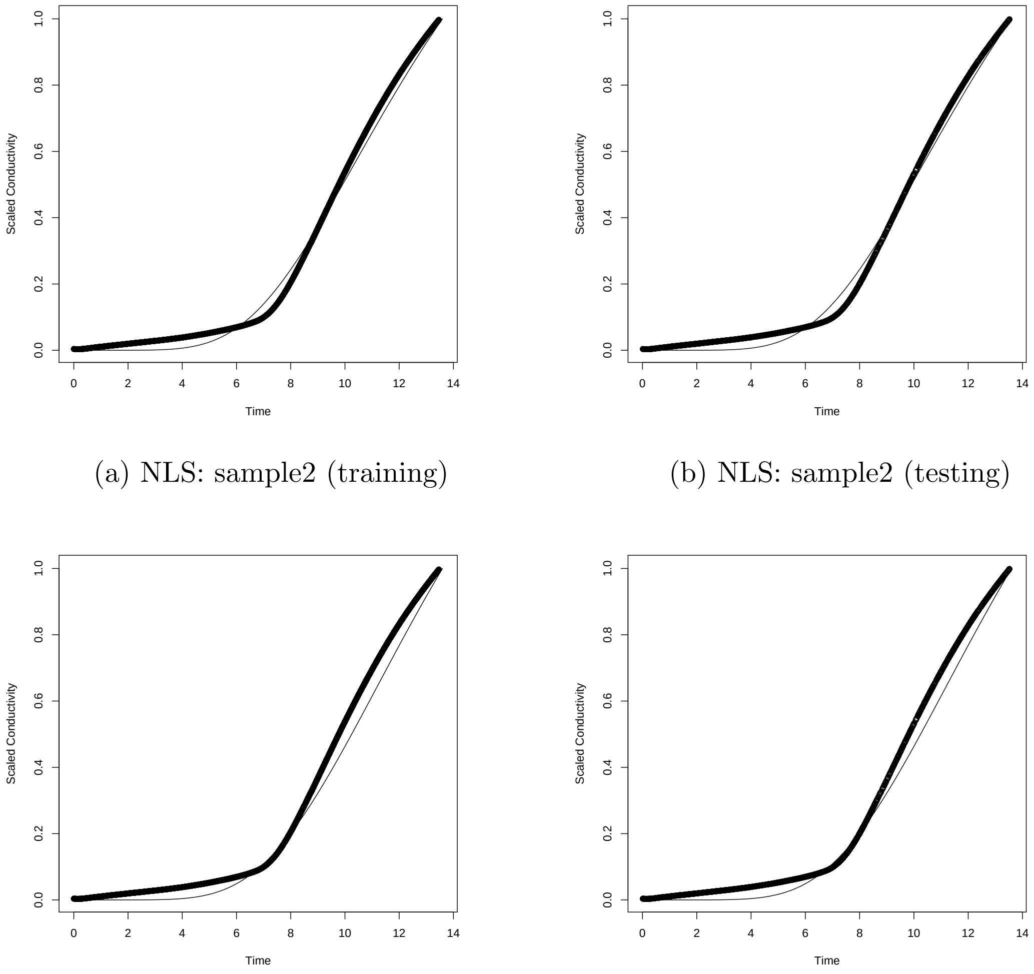
Table 5.4: Comparison of bootstrap results for NLS and NALR methods for sample2.

that although the corresponding RSS and rMSE values for NLS estimation are smaller than that of NALR, the r^2 values for NALR are more stable in a cross-validation setting than that of NLS estimation, and therefore provides evidence of over-fitting on the part of the latter. The "Zero" row indicates that the constrained optimisation is satisfied for both methods. The model suggested by the NLS method truncates the process at $\tau = 0.499876692$, and NALR at $\tau = 0.4394090106$, i.e. earlier in the process. The IP value corresponding to the NALR technique of 7.0017681635, is closer to the automated machine produced value of 7.20, than the NLS value of 6.4590059957. As a check of the validity of the results, the mode of the NLS model is attained at 10.3653615566, and that of the NALR model at 11.1759022681, which are both larger than the corresponding IP values, but smaller than the maximum value of 13.59872167.

© University of Pretoria

	NLS	NALR
r2 (Train)	0.9977	0.9953
r2 (Test)	0.9942	0.9961
RSS	1.1054	1.7258
rMSE	0.0187	0.0345
Zero	1.0000E-05	1.4158E-06
τ	0.4999	0.4394
IP (7.20)	6.4590	7.0018

Table 5.5: Comparison of performance results for NLS and AL methods for sample2.



(a) NLS: sample2 (training)

(b) NLS: sample2 (testing)

(c) NALR: sample2 (training)

(d) NALR: sample2 (testing)

Figure 5.5: Fitted sigmoidal curves to sample2.

© University of Pretoria

5.5 Summary

As an alternative to NLS, NALR was introduced in this chapter. The (truncated) four-parameter kappa CDF was considered as the non-linear function to be fitted, under a set of constraints. These constraints specifically resulted from biodiesel experiments where the induction period (IP) was of interest, obtained using the Rancimat method. From the results obtained, it was clear that circumstances exist under which NLS regression could be improved upon, and where NALR resulted in a plausible alternative. Both the NLS and NALR techniques recommended the Burr type III family of distributions as a solution to the problem posed here, although a much larger family was allowed for initially. The NALR technique provides for fitting models only on a subset of the data in a natural fashion, and is therefore ideal in situations where the data includes a natural drift, as was the case here.

In the following chapter, the mathematical premises for the estimation and regression techniques proposed in this thesis, are provided.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Chapter 6

A Theoretical Framework for Arc Length Based Statistics

In Chapters 3 – 5 arc length based measures have been proposed for parameter estimation (Section 3.3), GoF tests (Section 3.4), coefficient estimation in linear (Section 4.3) and non-linear (Chapter 5) regression. All these techniques make an implicit assumption of uniqueness of the arc length, in that it represents a characterisation of a random process.

This characterisation is presented in Section 6.1, and applied to the uniform distribution in Section 6.2, and exponential distribution in Section 6.3.

6.1 Introduction

Proschan and Pyke (1967) [70] proposed a test for monotone failure rate, based on spacings. This was further investigated by Bickel and Doksum (1969) [10] and Bickel (1969) [9]. It makes therefore sense to investigate the arc lengths of hazard functions, as the proposed sample arc length statistics are based on spacings (see Chapter 1).

Suppose that T is a continuous univariate random variable, with PDF $g(t)$, and CDF $G(t)$, with support $D \in \mathbb{R}$. The arc length of $g(t)$ on the interval $[a, b] \subseteq D$ is defined as

$$\mathcal{S}_g^{[a,b]} = \int_a^b \sqrt{1 + (g'(t))^2} dt$$

where $g'(t)$ denotes the first derivative of $g(t)$ with respect to t. This corresponds to Definition 2.1.1, but has been reproduced here for completeness.

© University of Pretoria

Proposition 6.1.1 (Arc Length Characterisation)**.** *Suppose that T is a continuous unimodal, univariate random variable with support on the interval $D = [t_0, t_1] \subset \mathbb{R}$, with PDF and CDF $g(t)$ and $G(t)$ respectively, and that $g'(t)$ exists. The arc length transformation*

$$h(t) = \sqrt{1 + (g'(t))^2} \tag{6.1.1}$$

defines a pseudo hazard function which characterises a class of distributions, having $f(t)$ and $F(t)$ respectively in the following way

$$\begin{aligned} F(t) =& \frac{1 - e^{-H(t)}}{1 - e^{-H(t_1)}}, \textit{ and} \\ f(t) =& \frac{h(t)e^{-H(t)}}{1 - e^{-H(t_1)}}. \end{aligned}$$

Proof. The result follows from the general solution to the first-order linear ordinary inhomogeneous differential equation with variable coefficients. Adopting the cumulative hazard function

$$H(t) = \int\limits_{t_0}^{t} h(x)dx$$

notation, with the understanding that $h(t)$ (see (6.1.1)) is strictly speaking not a hazard function, but will be referred to as a pseudo hazard function, since

$$\lim_{t \to \infty} H(t) < \infty.$$

© University of Pretoria

It follows that

$$\begin{aligned}
h(t) =& \frac{Af(t)}{1 - AF(t)} \\
=& \frac{AF'(t)}{1 - AF(t)} \\
\therefore AF'(t) =& h(t)(1 - AF(t)) \\
\therefore AF'(t) + h(t)AF(t) =& h(t) \\
\therefore F'(t) + h(t)F(t) =& \frac{1}{A}h(t) \\
\therefore F'(t)e^{\int\limits_{t_0}^{t} h(x)dx} + h(t)F(t)e^{\int\limits_{t_0}^{t} h(x)dx} =& \frac{1}{A}h(t)e^{\int\limits_{t_0}^{t} h(x)dx} \\
\therefore F'(t)e^{H(t)} + h(t)F(t)e^{H(t)} =& \frac{1}{A}h(t)e^{H(t)} \\
\therefore \frac{d}{dt}\left(F(t)e^{H(t)}\right) =& \frac{1}{A}h(t)e^{H(t)} \\
\therefore F(t) =& \frac{1}{A}e^{-H(t)}\left[\int\limits_{t_0}^{t} h(x)e^{H(x)}dx + C\right] \\
=& \frac{1}{A}e^{-H(t)}\left[e^{H(t)} - 1 + C\right], \qquad (6.1.2)
\end{aligned}$$

where A and C are constants. This is subject to the constraints that $F(t_0) = 0$, and $F(t_1) = 1$. It follows that

$$\begin{aligned}
0 =& \frac{1}{A}e^{-H(t_0)}\left[e^{H(t_0)} - 1 + C\right] \\
\therefore C =& 0 \\
1 =& \frac{1}{A}e^{-H(t_1)}\left[e^{H(t_1)} - 1 + C\right] \\
A =& e^{-H(t_1)}\left[e^{H(t_1)} - 1\right] \\
=& 1 - e^{-H(t_1)}.
\end{aligned}$$

Therefore, from (6.1.2)

$$\begin{aligned}
F(t) =& \frac{1}{1 - e^{-H(t_1)}}e^{-H(t)}\left(e^{H(t)} - 1\right) \\
=& \frac{1 - e^{-H(t)}}{1 - e^{-H(t_1)}} \\
\therefore f(t) =& \frac{h(t)e^{-H(t)}}{1 - e^{-H(t_1)}}.
\end{aligned}$$

□

Corollary 6.1.2. *Suppose that T is a continuous univariate random variable with semi-infinite support $D = [t_0, \infty) \subseteq \mathbb{R}$, with PDF $g(t)$, and CDF $G(t)$, and that $g'(t)$ exists. The arc length transformation*

$$h(t) = \sqrt{1 + (g'(t))^2}$$

defines a hazard function which characterises a class of distributions, having $f(t)$, and $F(t)$ as PDF and CDF respectively, in the following way

$$F(t) = 1 - e^{-H(t)}.$$

Proof. From standard definitions of hazard functions

$$\begin{aligned} h(t) =& \frac{f(t)}{1 - F(t)} \\ =& -\frac{d}{dt} \ln(1 - F(t)) \\ \therefore F(t) =& 1 - e^{-\int\limits_{t_0}^{t} h(x)dx} \\ =& 1 - e^{-H(t)}. \end{aligned}$$

note that

$$\lim_{t \to \infty} H(t) \to \infty$$

so that the hazard function $h(t)$ is well defined. □

Proposition 6.1.1 and Corollary 6.1.2 assert that T may be uniquely characterised by the full set of all possible arc length segments of $g(t)$ on its domain. The uniqueness sprouts from the association with $F(t)$, the CDF of a class of transformed variates, related to T through an arc length transformation, and by recognising

$$\mathcal{S}_g^{[t_0,t]} = H(t).$$

The uniform distribution is used in illustrating Proposition 6.1.1 as it is bounded, whereas the exponential distribution will serve as an example of a distribution with only a lower bound, in illustrating Corollary 6.1.2.

© University of Pretoria

6.2 Uniform Distribution

Let $T \sim \text{Uniform}(\alpha, \beta)$, $\alpha \leq t \leq \beta$, i.e.

$$g(t) = \frac{1}{\beta - \alpha} \; \alpha \leq t \leq \beta$$

so that the arc length on any arbitrary interval $[a, b] \in [\alpha, \beta]$ is given by

$$\begin{aligned} \mathcal{S}_g^{[a,b]} =& \int_a^b \sqrt{1 + (g'(x))^2} dx \\ =& \int_a^b dx \\ =& b - a, \end{aligned}$$

since the PDF of the continuous uniform distribution is just a horizontal line. Thus from Proposition 6.1.1, for $\alpha \leq t \leq \beta$,

$$\begin{aligned} F(t) =& \frac{1 - e^{-H(t)}}{1 - e^{-H(\beta)}} \\ =& \frac{1 - e^{-(t-\alpha)}}{1 - e^{-(\beta-\alpha)}}, \end{aligned}$$

and

$$f(t) = \frac{e^{-(t-\alpha)}}{1 - e^{-(\beta-\alpha)}},$$

which corresponds to a shifted exponential distribution, with scale parameter 1, right-truncated at β. This suggests that the arc length of the PDF of the Uniform(α, β) distribution, is uniquely characterised by the shifted, right-truncated exponential distribution (with scale parameter equal to 1), i.e. shifted with parameter α, and right-truncated with parameter β.

6.3 Exponential Distribution

Let $T \sim \text{Exp}(\theta)$, i.e.

$$g(t) = \theta e^{-\theta t}, \; t \geq 0$$

so that the arc length on any arbitrary interval $[a,b] \in \mathbb{R}^+$ is given by

$$\begin{aligned}\mathcal{S}_g^{[a,b]} &= \int\limits_a^b \sqrt{1+(g'(x))^2}dx \\ &= \int\limits_a^b \sqrt{1+\theta^2(g(x))^2}dx.\end{aligned}$$

Making the substitution

$$\begin{aligned}\theta e^{-\theta t} &= \sinh u \\ \therefore u &= arcsinh(\theta e^{-\theta t}) \\ &= \ln(\theta e^{-\theta t} + \sqrt{\theta e^{-2\theta t}+1}) \\ \therefore a^* &= \ln(\theta e^{-\theta a} + \sqrt{\theta e^{-2\theta a}+1}) \\ \therefore b^* &= \ln(\theta e^{-\theta b} + \sqrt{\theta e^{-2\theta b}+1}) \\ \therefore -\theta^2 e^{-\theta t} dt &= \cosh u du \\ \therefore dt &= -\frac{1}{\theta}\coth u du.\end{aligned}$$

so that

$$\begin{aligned}\mathcal{S}_g^{[a,b]} &= \int\limits_a^b \sqrt{1+\theta^2 e^{-2\theta t}}dt \\ &= -\frac{1}{\theta}\int\limits_{a^*}^{b^*} \sqrt{1+\sinh^2 u}\coth u du \\ &= -\frac{1}{\theta}\int\limits_{a^*}^{b^*} \frac{\cosh^2 u}{\sinh u}du \\ &= -\frac{1}{\theta}\cosh u]_{a^*}^{b^*} - \frac{1}{\theta}\int\limits_{a^*}^{b^*} csch\, u du \\ &= -\frac{1}{\theta}\left[\cosh u + \ln\left|\tanh\frac{u}{2}\right|\right]_{a^*}^{b^*}, \; u \neq 0.\end{aligned}$$

Although solvable exactly, a unrecognisable distribution is obtained, rendering any further analyses unuseable (see the preface to McElreath (2016) [58]).

© University of Pretoria

6.4 Summary

This chapter provided the mathematical premises for the arc length based measures proposed in this thesis. It was seen that the arc length transformation may be viewed as a type of generating function, which implies that a distribution is characterised by an infinite collection of arc lengths, akin to other generators, such as moments.

© University of Pretoria

Chapter 7

Conclusions

The conclusions of the work presented in this thesis, are summarised in this chapter, specifically, in Section 7.1. Future directions that could be taken by related research, are outlined in Section 7.2.

7.1 Summary of Conclusions

The arc length measure has provided the fuel for contemplation for some of the most basic notions encountered in Statistics. These included estimation of univariate distributions, linear regression, and non-linear regression. As statistical inference is the goal of these practises, this was highlighted throughout.

The method of arc lengths, for parameter estimation, was introduced, and it was shown that although it being consistent, generally leads to parameter estimates with greater bias and rMSE than ML, CvM, and AD methods. A discrete and continuous sample arc length statistic were proposed. In particular, the discrete arc length sample statistic, ${}_1\!\int^{[q_{0.01},q_{0.99}]}$ proved to be particularly effective for estimating σ, and offers results comparable to that obtained by the CvM method. These arc length sample statistics provide great flexibility through its tuning parameters.

The arc length (GoF) test was constructed using the continuous sample arc length statistic, for which a parametric bootstrap yielded a null distribution. A GoF test constructed using particular tuning parameters, with parameters estimated using the discrete sample arc length statistic, yielded the most powerful tests. It was shown that the arc length test is more powerful against positively skewed and platykurtic alternatives

© University of Pretoria

than both the CvM and AD tests.

Two new regression techniques were proposed based on equating properties of KDE's, based on a set of dependent and independent observations. Both these methods didn't tamper with the linear model configuration, but instead proposed new objective functions. Standard errors, confidence intervals and p-values were obtained using the bootstrap, and the results obtained compared to that obtained from OLS regression.

It was seen that although OLS regression resulted in the smallest residual standard errors, ALR could predict a larger percentage explained variation as measured by r^2, but that the MMM, had the smallest divergence between the observed, and predicted value distributions.

As an alternative to NLS, and as a generalisation of ALR, NALR was introduced. The (truncated) four-parameter kappa CDF was considered as the non-linear function to be fitted, under a set of constraints. From the results obtained, it was clear that circumstances exist under which NLS regression could be improved upon, and where NALR resulted in a plausible alternative.

In summary: Computationally intensive alternatives for central techniques used in Statistics, have been developed based on the arc length of particular statistical functions. In some cases, traditional methodologies could be improved upon, using predefined optimality criterion. All the above clearly indicate that Statistics has indeed a role to play in the development of computationally intensive procedures, and that such methods may very well outperform existing techniques.

7.2 Future Work

For the techniques developed in Chapters 3 and 4, the Gaussian KDE, along with Silverman's (rule of thumb) bandwidth selection method, were used. A natural extension is to change any of these two components, and to compare the results with established techniques in the literature, or to improve on the work presented here.

The method of arc lengths for parameter estimation and the arc length test for GoF in Chapter 3 were developed on the PDF, but could be extended to the CDF, or QF. The latter being important for distributions not possessing a PDF or CDF in closed form. However, the PDF naturally leads to multivariate generalisations of the method of arc lengths to the "method of surface areas", and arc length test as the "surface area

© University of Pretoria

test", as the theory for multivariate KDE's are well developed, and since the probability integral transform is not relied upon, as many of the spacings based methods, reviewed in Chapter 1, do. The normal distribution served as the vehicle for illustration of the proposed techniques, but could be applied to any other univariate probability distribution for which the derivative of the PDF can be calculated.

Although the techniques developed in this thesis were made with the implicit assumption of unlimited computational capacity, any effort made in speeding up calculations, would be advantageous. In this regard, the work done by Floater [30], and Floter and Rasmussen [31] could for instance be used, as the calculation of the arc length has been a long standing problem in numerical mathematics. By recognising the systems of equations required to be solved in Chapter 4 for the moment matching and arc length regression techniques, as systems of Fredholm equations, the work done by Twomey (1963) [81] could also yield faster, or approximations to explicit solutions.

The NALR technique, applied to the four-parameter kappa distribution in Chapter 5, could be applied to any non-linear regression problem, with or without constraints.

Alas, the resounding conclusion is that a great many work on the subjects touched upon in this thesis, will remain unconquered, and that Statistics has yet a role to play in the current age of computation.

© University of Pretoria

Bibliography

[1] R. J. Adcock. A problem in least squares. *The Analyst*, 5(2):53–54, 1878.

[2] D. Adler. *vioplot: Violin plot*, 2005. R package version 0.2.

[3] S. J. Ahn. Geometric fitting of parametric curves and surfaces. *JIPS*, 4(4):153–158, 2008.

[4] T. W. Anderson and D. A. Darling. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[5] A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178, 1985.

[6] A. Azzalini. *The R package* `sn`*: The Skew-Normal and Skew-t distributions (version 1.4-0)*. Università di Padova, Italia, 2016.

[7] A. Azzalini and B. Scarpa. *Data analysis and data mining: an Introduction*. Oxford University Press, New York, 2012.

[8] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the Indian Journal of Statistics*, 7(4):401–406, 1946.

[9] P. J. Bickel. Tests for monotone failure rate II. *The Annals of Mathematical Statistics*, 40(4):1250–1260, 1969.

[10] P. J. Bickel and K. A. Doksum. Tests for monotone failure rate based on normalized spacings. *The Annals of Mathematical Statistics*, 40(4):1216–1235, 1969.

[11] R. W. Butler. *Saddlepoint approximations with applications*. Cambridge University Press, New York, 2007.

© University of Pretoria

[12] W.-C. Chen, G. Ostrouchov, D. Schmidt, P. Patel, and H. Yu. pbdMPI: Programming with big data – interface to MPI, 2012. R Package, URL http://cran.r-project.org/package=pbdMPI.

[13] Y. L. Chen. A test for two-sample problem based on sample spacings. *Tamsui Oxford Journal of Mathematical Sciences*, 20(2):267–278, 2004.

[14] N. Chernov. *Fitting geometric curves to observed data.* Citeseer, United States, 2011.

[15] E. L. T. Conceicao and M. Maechler. *DEoptimR: Differential Evolution Optimization in Pure R*, 2015. R package version 1.0-4.

[16] J. L. Coolidge. Two geometrical applications of the method of least squares. *The American Mathematical Monthly*, 20(6):187–190, 1913.

[17] N. Cressie. On the logarithms of high-order spacings. *Biometrika*, 63(2):343–355, 1976.

[18] H. E. Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 25(4):631–650, 1954.

[19] M. L. Delignette-Muller and C. Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015.

[20] W. E. Deming. *Statistical adjustment of data.* Wiley, New York, 1943.

[21] D. Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer, New York, 2013. ISBN 978-1-4614-6867-7.

[22] D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[23] D. Eddelbuettel and C. Sanderson. Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063, 2014.

[24] B. Efron. *The jack knife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

© University of Pretoria

[25] B. Efron and T. Hastie. *Computer Age Statistical Inference*, volume 5. Cambridge University Press, New York, 2016.

[26] M. Ekström. Alternatives to maximum likelihood estimation based on spacings and the Kullback–Leibler divergence. *Journal of Statistical Planning and Inference*, 138(6):1778–1791, 2008.

[27] European Standard EN 14112. Fat and oil derivatives-fatty acid methyl esters (fame)-determination of oxidation stability (accelerated oxidation test). 2003.

[28] European Standard EN 14214. Automotive fuels-fatty acid methyl ester (fame) for diesel engines-requirements and test methods. 2008.

[29] J. Faraway, G. Marsaglia, J. Marsaglia, and A. Baddeley. *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*, 2015. R package version 1.0-3.

[30] M. S. Floater. Arc length estimation and the convergence of polynomial curve interpolation. *BIT Numerical Mathematics*, 45(4):679–694, 2005.

[31] M. S. Floater and A. F. Rasmussen. Point-based methods for estimating the length of a parametric curve. *Journal of Computational and Applied Mathematics*, 196(2):512–522, 2006.

[32] W. W. Focke, I. van der Westhuizen, and X. Oosthuizen. Biodiesel oxidative stability from rancimat data. *Thermochimica Acta*, 633:116–121, 2016.

[33] J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011.

[34] R. Gatto and J. S. Rao. A conditional saddlepoint approximation for testing problems. *Journal of the American Statistical Association*, 94(446):533–541, 1999.

[35] K. Ghosh and J. S. Rao. Small sample approximations for spacing statistics. *Journal of Statistical Planning and Inference*, 69(2):245–261, 1998.

[36] K. Ghosh and S. R. Jammalamadaka. A general estimation method using spacings. *Journal of Statistical Planning and Inference*, 93(1):71–82, 2001.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[37] G. H. Golub and C. F. Van Loan. An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, 17(6):883–893, 1980.

[38] M. Greenwood. The statistical study of infectious diseases. *Journal of the Royal Statistical Society*, 109(2):85–110, 1946.

[39] H. Z. K. Hadorn. Zur bestimmung der oxydationsstabilitt von len und fetten. *Deutsche Lebensmittel-Rundschau*, 70:57–65, 1974.

[40] J. L. Hintze and R. D. Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[41] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[42] L. Holst. Asymptotic normality of sum-functions of spacings. *The Annals of Probability*, 7(6):1066–1072, 1979.

[43] L. Holst and J. S. Rao. Asymptotic theory for some families of two-sample nonparametric statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, 42(1/2):19–52, 1980.

[44] L. Holst and J. S. Rao. Asymptotic spacings theory with applications to the two-sample problem. *Canadian Journal of Statistics*, 9(1):79–89, 1981.

[45] J. R. M. Hosking. The four-parameter kappa distribution. *IBM Journal of Research and Development*, 38(3):251 – 258, 1994.

[46] S. G. Hussey, M. T. Loots, K. van der Merwe, E. Mizrachi, and A. A. Myburg. Integrated analysis and transcript abundance modelling of h3k4me3 and h3k27me3 in developing secondary xylem. *Scientific Reports*, 2017.

[47] S. R. Jammalamadaka and A. Sengupta. *Topics in circular statistics*, volume 5. World Scientific, London, 2001.

[48] M. Kac, J. Kiefer, and J. Wolfowitz. On tests of normality and other tests of goodness of fit based on distance methods. *The Annals of Mathematical Statistics*, 26(2):189–211, 1955.

© University of Pretoria

[49] J. T. Kent, K. V. Mardia, and J. S. Rao. A characterization of the uniform distribution on the circle. *The Annals of Statistics*, 7(4):882–889, 1979.

[50] C. H. Kummell. Reduction of observation equations which contain more than one observed quantity. *The Analyst*, 6(4):97–105, 1879.

[51] M. W. Läubli and P. A. Bruttel. Determination of the oxidative stability of fats and oils: Comparison between the active oxygen method (aocs cd 12-57) and the rancimat method. *Journal of the American Oil Chemists' Society*, 63:792–795, 1986.

[52] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[53] Y. Liu and W. Wang. A revisit to least squares orthogonal distance fitting of parametric curves and surfaces. In *International Conference on Geometric Modeling and Processing*, pages 384–397. Springer, 2008.

[54] M. T. Loots. *alR: Arc lengths of statistical functions.* https://github.com/mtloots/alR, 2017.

[55] A. Luceño. Fitting the generalized pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics and Data Analysis*, 51(2):904–917, 2006.

[56] C. Ma and J. Robinson. Saddlepoint approximations for the difference of order statistics and studentized sample quantiles. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):563–577, 1999.

[57] K. V. Mardia and P. E. Jupp. *Directional statistics.* Wiley, New York, 2009.

[58] R. McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan.* CRC Press, New York, 2016.

[59] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015. R package version 1.6-7.

[60] P. W. Mielke Jr. Another family of distributions for describing and analyzing precipitation data. *Journal of Applied Meteorology*, 12(2):275 – 280, 1973.

© University of Pretoria

[61] D. Minda and S. Phelps. Triangles, ellipses, and cubic polynomials. *American Mathematical Monthly*, 115(8):679–689, 2008.

[62] A. M. Mineo. *normalp: Routines for Exponential Power Distribution*, 2014. R package version 0.7.0.

[63] A. M. Mineo, M. Ruggieri, et al. A software tool for the exponential power distribution: The normalp package. *Journal of Statistical Software*, 12(4):1–24, 2005.

[64] S. M. Mirakhmedov and S. R. Jammalamadaka. Higher-order expansions and efficiencies of tests based on spacings. *Journal of Nonparametric Statistics*, 25(2):339–359, 2013.

[65] P. A. P. Moran. The random division of an interval. *Supplement to the Journal of the Royal Statistical Society*, 9(1):92–98, 1947.

[66] E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

[67] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[68] K. Pearson. Note on Francis Galton's problem. *Biometrika*, 1:390–399, 1902.

[69] D. Pollard. The minimum distance method of testing. *Metrika*, 27(1):43–70, 1980.

[70] F. Proschan and R. Pyke. Tests for monotone failure rate. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 293–312. Univ of California Press, 1967.

[71] R. Pyke. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, 7:395–449, 1965.

[72] R. Pyke. Spacings revisited. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. The Regents of the University of California, 1972.

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

[74] J. S. Rao. Some tests based on arc-lengths for the circle. *Sankhyā: The Indian Journal of Statistics, Series B*, 38(4):329–338, 1976.

[75] G. S. Russell and D. J. Levitin. An expanded table of probability values for Rao's spacing test. *Communications in Statistics-Simulation and Computation*, 24(4):879–888, 1995.

[76] H. L. Seal. Studies in the history of probability and statistics. xv the historical development of the Gauss linear model. *Biometrika*, 54(1-2):1–24, 1967.

[77] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, London, 1986.

[78] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th international conference on Machine learning*, pages 992–999. ACM, 2008.

[79] D. D. Tung and J. S. Rao. U-statistics based on spacings. *Journal of Statistical Planning and Inference*, 142(3):673–684, 2012.

[80] D. D. Tung and J. S. Rao. On the Gini mean difference test for circular data. *Communications in Statistics-Theory and Methods*, 42(11):1998–2008, 2013.

[81] S. Twomey. On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature. *Journal of the ACM (JACM)*, 10(1):97–101, 1963.

[82] R. Varadhan. *alabama: Constrained Nonlinear Optimization*, 2015. R package version 2015.3-1.

[83] G. R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables. *gplots: Various R Programming Tools for Plotting Data*, 2016. R package version 3.0.1.

[84] L. Wasserman. Data science: The end of statistics?, 2013.

[85] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

© University of Pretoria

[86] J. Wolfowitz. Estimation by the minimum distance method. *Annals of the Institute of Statistical Mathematics*, 5(1):9–23, 1953.

© University of Pretoria

Appendix A

The alR Package

Highlights of the "alR" package [54] developed for the application of the methods proposed in this thesis, are presented in this appendix.

A.1 Estimation

A.1.1 Method of Arc Lengths

`alE`	*Arc length estimation.*

Description

A framework for arc length estimation.

Usage

```
alE(x, q1, q2, dc, type)

alEfitdist(x, q1, q2, dc, type, bootstraps)

alEfit(X, q1, q2, dc, type, bootstraps, ...)

## Default S3 method:
```

© University of Pretoria

```
alEfit(X, q1, q2, dc, type, bootstraps, ...)

## S3 method for class 'alEfit'
print(x, ...)
```

Arguments

`x`	An alEfit object.
`q1, q2`	Vectors specifying the quantiles over which arc length segments are to be computed.
`dc`	TRUE/FALSE: Should the discrete or continuous sample statistic be used.
`type`	The type of bandwidth estimator for the underlying KDE; see `bw`.
`bootstraps`	An integer specifying the size of the parametric bootstrap.
`X`	A vector of sample values.
`...`	Additional arguments passed to `alEfit` (not currently used).

Details

- Estimate distributional parameters using the method of arc lengths.
- Simulate bootstrap distributions for parameter estimates, resulting from sample arc length statistics.

This method is currently only implemented for the normal distribution. The underlying C code for the Nelder-Mead method of the optim function is used for optimising the objective function. The tolerance level is set at 1e-15, and a maximum number of 1000 iterations is allowed. The maximum likelihood estimates are used as initial values for the Nelder-Mead algorithm.

Value

alE: A list with the following components (see `optim`):

- par: The estimated parameters.
- abstol: The absolute tolerance level (default 1e-15).

© University of Pretoria

- fail: An integer code indicating convergence.
- fncount: Number of function evaluations.

alEfitdist: A matrix of parameter estimates resulting from the estimated arc lengths over the specified interval(s), i.e. the bootstrap distribution for the estimated parameters resulting from the chosen sample arc length statistic.

alEfit: A generic S3 object with class alEfit.

alEfit.default: A list with all components from `alE`, as well as :

- dc: TRUE/FALSE Was the discrete or continuous sample arc length statistic used?
- q1, q2: The segments over which the arc length(s) were calculated.
- bw: The bandwidth for the kernel density estimator.
- dist: A numeric matrix whose columns represent a bootstrap distribution for the corresponding parameter estimate.
- se: A numeric vector with standard errors, obtained by a parametric bootstrap.
- bootstraps: Number of bootstrap samples.

Methods (by class)

- `default`: default method for alEfit.
- `alEfit`: print method for alEfit.

Examples

```
x <- rnorm(1000)
alE(x,0.025, 0.975, TRUE, -1)
alE(x,c(0.025, 0.5), c(0.5, 0.975), TRUE, -1)
alE(x,0.025, 0.975, FALSE, -1)
alE(x,c(0.025, 0.5), c(0.5, 0.975), FALSE, -1)

## Not run:
alEfitdist(x, 0.025, 0.975, TRUE, -1, 100)
alEfitdist(x, 0.025, 0.975, FALSE, -1, 100)
```

© University of Pretoria

```
## End(Not run)
alEfit(x, q1=0.025, q2=0.975, dc=TRUE, type=-1, bootstraps=50)
alEfit(x, q1=0.025, q2=0.975, dc=FALSE, type=-1, bootstraps=50)
```

A.1.2 Arc Length Test

`alEdist` *Arc length estimation.*

Description

Goodness-of-fit using arc lengths.

Usage

```
alEdist(n, bootstraps, mu, sigma, q1, q2, quantile, dc, type)

alEtest(X, mu, sigma, q1, q2, type, bootstraps, ...)

## Default S3 method:
alEtest(X, mu, sigma, q1, q2, type, bootstraps, ...)

## S3 method for class 'alEtest'
print(x, ...)
```

Arguments

`n`	An integer specifying the sample size.
`bootstraps`	An integer specifying the size of the parametric bootstrap.
`mu`	A real value specifying the mean of the normal distribution.
`sigma`	A positive real number specifying the scale parameter of the normal distribution.

© University of Pretoria

`q1, q2` Vectors specifying the quantiles (or points if quantile=FALSE) over which arc length segments are to be computed.

`quantile` TRUE/FALSE whether q1 and q2 are quantiles, or elements of the domain of `x`.

`dc` TRUE/FALSE: Should the discrete or continuous sample statistic be used.

`type` The type of bandwidth estimator for the underlying KDE; see `bw`.

`X` A vector of sample values.

`...` Additional arguments passed to `alEtest` (not currently used).

`x` An alEtest object.

Details

First the distributional parameters of a sample is estimated using the continuous arc length sample statistic (see `alE`). The calculated sample arc length statistic is then compared to the distribution of that particular sample statistic, obtained by a parametric bootstrap, using the estimated parameters (see `alEdist`). This finally leads to the calculation of a p-value for a goodness-of-fit test, based on the simulated distribution.

This method is currently only implemented for the normal distribution, and for a single arc length segment.

Value

alEdist: A vector (matrix) of arc lengths over the specified interval(s), i.e. the simulated distribution for the chosen sample arc length statistic.

alEtest: A generic S3 object with class alEtest.

alEtest.default: A list with the following components:

- q1, q2: The segment over which the arc length was calculated.
- mu: A real value specifying the mean of the normal distribution.
- sigma: A positive real number specifying the scale parameter of the normal distribution.

© University of Pretoria

- bw: The bandwidth for the kernel density estimator.
- dist: A numeric matrix whose columns represent a bootstrap distribution for the corresponding sample arc length statistic.
- statistic: The value of the observed sample statistic.
- pvalue: The p-value for the test based on a parametric bootstrap sample.
- bootstraps: Number of bootstrap samples.

Methods (by class)

- `default`: default method for alEtest.
- `alEtest`: print method for alEtest.

Examples

```
## Not run:
alEdist(50, 100, 2, 3.5, 0.025, 0.975, TRUE, TRUE, -1)
alEdist(50, 100, 2, 3.5, c(0.025,0.5), c(0.5,0.975), TRUE, TRUE, -1)
alEdist(50, 100, 2, 3.5, 0.025, 0.975, TRUE, FALSE, -1)
alEdist(50, 100, 2, 3.5, c(0.025,0.5), c(0.5,0.975), TRUE, FALSE, -1)
alEdist(50, 100, 2, 3.5, qnorm(0.025,2,3.5),
qnorm(0.975, 2, 3.5), FALSE, FALSE, -1)
alEdist(50, 100, 2, 3.5, c(qnorm(0.025, 2, 3.5),2),
c(2,qnorm(0.975, 2, 3.5)), FALSE, FALSE, -1)

## End(Not run)
## Not run:
x <- rnorm(1000)
s1 <- alE(x, 0.025, 0.975, TRUE, -1)
alEtest(x, mu=s1$par[1], sigma=s1$par[2], q1=0.025, q2=0.975,
type=-1, bootstraps=50)
s2 <- alE(x, 0.025, 0.975, FALSE, -1)
alEtest(x, mu=s2$par[1], sigma=s2$par[2], q1=0.025, q2=0.975,
type=-1, bootstraps=50)

## End(Not run)
```

© University of Pretoria

A.2 Arc Length Regression

A.2.1 Moment Matching

mmKDEboot	*Moment matching for kernel density estimators.*

Description

Bootstrap estimates, along with standard errors and confidence intervals, of a linear model, resulting from moment matching of kernel density estimates.

Usage

```
mmKDEboot(formula, data = list(), xin, type, bootstraps, bootName, ...)

## Default S3 method:
mmKDEboot(formula, data = list(), xin, type, bootstraps,
  bootName, ...)

## S3 method for class 'mmKDEboot'
print(x, ...)

## S3 method for class 'mmKDEboot'
summary(object, ...)

## S3 method for class 'summary.mmKDEboot'
print(x, ...)

## S3 method for class 'formula'
mmKDEboot(formula, data = list(), xin, type, bootstraps,
  bootName, ...)
```

© University of Pretoria

```
## S3 method for class 'mmKDEboot'
predict(object, newdata = NULL, ...)
```

Arguments

`formula` An LHS ~ RHS formula, specifying the linear model to be estimated.

`data` A data.frame which contains the variables in `formula`.

`xin` Numeric vector of length equal to the number of independent variables, of initial values, for the parameters to be estimated.

`type` An integer specifying the bandwidth selection method used, see `bw`.

`bootstraps` An integer giving the number of bootstrap samples.

`bootName` The name of the .rds file to store the mmKDEboot object. May include a path.

`...` Arguments to be passed on to the control argument of the `optim` function.

`x` An mmKDEboot object.

`object` An mmKDEboot object.

`newdata` The data on which the estimated model is to be fitted.

Value

A generic S3 object with class mmKDEboot.

mmKDEboot.default: A list object (saved using `saveRDS` in the specified location) with the following components:

- intercept: Did the model contain an intercept TRUE/FALSE?
- coefficients: A vector of estimated coefficients.
- coefDist The bootstrap parameter distribution.
- bcoefficients: A vector of bootstrap coefficients, resulting from bootstrap estimation.
- df: Degrees of freedom of the model.

© University of Pretoria

- se: The standard errors for the estimates resulting from bootstrap estimation.
- error: The value of the objective function.
- errorList: A vector of values of the objective function for each bootstrap sample.
- fitted.values: A vector of estimated values.
- residuals: The residuals resulting from the fitted model.
- call: The call to the function.
- h_y: The KDE bandwidth estimator for the dependent variable.
- h_X: The KDE bandwidth estimator for the independent variables, i.e. $\mathbf{X}\underline{\hat{\beta}}$.
- MOMy: The first n non central moments of the dependent variable, where n is the number of columns in the design matrix.
- MOMX: The first n non central moments of the independent variables $\mathbf{X}\underline{\hat{\beta}}$, where n is the number of columns in the design matrix.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.

summary.mmKDEboot: A list of class summary.mmKDEboot with the following components:

- call: Original call to `mmKDEboot` function.
- coefficients: A matrix with estimates, estimated errors, and 95% parameter confidence intervals (based on the inverse empirical distribution function).
- moments: A matrix of the first n moments of the dependent and independent variables that were matched. The final row corresponds to the estimated bandwidth parameters for each, i.e. `h_y` and `h_X`, respectively.
- r.squared: The r^2 coefficient.
- adj.r.squared: The adjusted r^2 coefficient.
- sigma: The residual standard error.
- df: Degrees of freedom for the model.
- error: Value of the objective function.

© University of Pretoria

- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.
- residSum: Summary statistics for the distribution of the residuals.
- errorSum: Summary statistics for the distribution of the value of the objective function.

print.summary.mmKDEboot: The object passed to the function is returned invisibly.

predict.mmKDEboot: A vector of predicted values resulting from the estimated model.

Methods (by class)

- `default`: default method for mmKDEboot.
- `mmKDEboot`: print method for mmKDEboot.
- `mmKDEboot`: summary method for mmKDEboot.
- `summary.mmKDEboot`: print method for summary.mmKDEboot.
- `formula`: formula method for mmKDEboot.
- `mmKDEboot`: predict method for mmKDEboot.

A.2.2 Arc Length Matching

`alKDEboot` *Arc length matching for kernel density estimators.*

Description

Bootstrap estimates, along with standard errors and confidence intervals, of a linear model, resulting from arc length matching of kernel density estimates.

© University of Pretoria

Usage

```
alKDEboot(formula, data = list(), xin, q1, q2, type, bootstraps, bootName,
  ...)

## Default S3 method:
alKDEboot(formula, data = list(), xin, q1, q2, type,
  bootstraps, bootName, ...)

## S3 method for class 'alKDEboot'
print(x, ...)

## S3 method for class 'alKDEboot'
summary(object, ...)

## S3 method for class 'summary.alKDEboot'
print(x, ...)

## S3 method for class 'formula'
alKDEboot(formula, data = list(), xin, q1, q2, type,
  bootstraps, bootName, ...)

## S3 method for class 'alKDEboot'
predict(object, newdata = NULL, ...)
```

Arguments

`formula`	An LHS ~ RHS formula, specifying the linear model to be estimated.
`data`	A data.frame which contains the variables in `formula`.
`xin`	Numeric vector of length equal to the number of independent variables, of initial values, for the parameters to be estimated.
`q1, q2`	Numeric vectors, for the lower and upper bounds of the intervals over which arc lengths are to be computed.
`type`	An integer specifying the bandwidth selection method used, see `bw`.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

`bootstraps` An integer giving the number of bootstrap samples.

`bootName` The name of the .rds file to store the alKDEboot object. May include a path.

`...` Arguments to be passed on to the control argument of the `optim` function.

`x` An alKDEboot object.

`object` An alKDEboot object.

`newdata` The data on which the estimated model is to be fitted.

Value

A generic S3 object with class alKDEboot.

alKDEboot.default: A list object (saved using `saveRDS` in the specified location) with the following components:

- intercept: Did the model contain an intercept TRUE/FALSE?
- coefficients: A vector of estimated coefficients.
- coefDist The bootstrap parameter distribution.
- bcoefficients: A vector of bootstrap coefficients, resulting from bootstrap estimation.
- df: Degrees of freedom of the model.
- se: The standard errors for the estimates resulting from bootstrap estimation.
- error: The value of the objective function.
- errorList: A vector of values of the objective function for each bootstrap sample.
- fitted.values: A vector of estimated values.
- residuals: The residuals resulting from the fitted model.
- call: The call to the function.
- h_y: The KDE bandwidth estimator for the dependent variable.
- h_X: The KDE bandwidth estimator for the independent variables, i.e. $\mathbf{X}\underline{\hat{\beta}}$.
- ALy: Arc length segments of the KDE cast over the dependent variable.

© University of Pretoria

- ALX: Arc length segments of the KDE cast over the independent variables $\mathbf{X}\underline{\hat{\beta}}$. p1: The vector of quantiles in the domain of y corresponding to `q1`. p2: The vector of quantiles in the domain of y corresponding to `q2`.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.

summary.alKDEboot: A list of class summary.alKDEboot with the following components:

- call: Original call to the `alKDEboot` function.
- coefficients: A matrix with estimates, estimated errors, and 95% parameter confidence intervals (based on the inverse empirical distribution function).
- arclengths: A matrix of the arc length segments that were matched, for the dependent and independent variables. The final row corresponds to the estimated bandwidth parameters for each, i.e. `h_y` and `h_X`, respectively.
- r.squared: The r^2 coefficient.
- adj.r.squared: The adjusted r^2 coefficient.
- sigma: The residual standard error.
- df: Degrees of freedom for the model.
- error: Value of the objective function.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.
- residSum: Summary statistics for the distribution of the residuals.
- errorSum: Summary statistics for the distribution of the value of the objective function.

print.summary.alKDEboot: The object passed to the function is returned invisibly.

predict.alKDEboot: A vector of predicted values resulting from the estimated model.

Methods (by class)

- `default`: default method for alKDEboot.

© University of Pretoria

- `alKDEboot`: print method for alKDEboot.
- `alKDEboot`: summary method for alKDEboot.
- `summary.alKDEboot`: print method for summary.alKDEboot.
- `formula`: formula method for alKDEboot.
- `alKDEboot`: predict method for alKDEboot.

A.2.3 Bhattacharyya Divergence Test

`bhatt.test` *Bhattacharryya test for comparing two samples.*

Description

Use the multinomial distribution to test the hypothesis that two samples come from the same distribution.

Usage

```
bhatt.test(y, x, k)
```

Arguments

`y, x`	Samples to be compared.
`k`	Number of proportions.

Details

It is assumed that the two samples come from the same kernel density distribution. The support of the KDE of the first sample is divided into k equally spaced quantiles, and then compared to the corresponding proportions of the second.

© University of Pretoria

Value

bat.test: A list with the following components:

- df=2*k: where k is the number of proportions used.
- y.prop, x.prop: Vectors of proportions.
- D2: Measure of divergence between samples.
- test.stat: The test statistic for the Bhattacharrayya test.
- p.value: The p-value of the test.

Examples

```
y <- rnorm(1000)
x <- rnorm(1000)
bhatt.test(y,x,10)
```

A.3 Non-Linear Arc Length Regression

A.3.1 Non-linear Least Squares

`kappa4nlsBoot` *Sigmoidal curve fitting.*

Description

Bootstrap estimates, along with standard errors and confidence intervals, of a nonlinear model, resulting from nonlinear least squares fitting of the four-parameter kappa sigmoidal function.

Usage

```
kappa4nlsBoot(formula, data = list(), xin, lower, upper, tol, maxiter,
  bootstraps, bootName, ...)
```

© University of Pretoria

```
## Default S3 method:
kappa4nlsBoot(formula, data = list(), xin, lower = c(0,
  -5, -5), upper = c(10, 1, 1), tol = 1e-15, maxiter = 50000, bootstraps,
  bootName, ...)

## S3 method for class 'kappa4nlsBoot'
print(x, ...)

## S3 method for class 'kappa4nlsBoot'
summary(object, ...)

## S3 method for class 'summary.kappa4nlsBoot'
print(x, ...)

## S3 method for class 'formula'
kappa4nlsBoot(formula, data = list(), xin, lower, upper,
  tol, maxiter, bootstraps, bootName, ...)

## S3 method for class 'kappa4nlsBoot'
predict(object, newdata = NULL, ...)
```

Arguments

`formula` An LHS ˜ RHS formula, specifying the linear model to be estimated.

`data` A data.frame which contains the variables in `formula`.

`xin` Numeric vector of length 3 containing initial values, for σ, h, and k.

`lower` A vector of lower constraints for the parameters to be estimated; defaults to c(0, -5, -5).

`upper` A vector of upper constraints for the parameters to be estimated; defaults to c(10, 1, 1).

`tol` Error tolerance level; defaults to 1e-15.

`maxiter` The maximum number of iterations allowed; defaults to 50000.

© University of Pretoria

`bootstraps` An integer giving the number of bootstrap samples.

`bootName` The name of the .rds file to store the kappa4nlsBoot object. May include a path.

`...` Arguments to be passed on to the differential evolution function `JDEoptim`.

`x` A kappa4nlsBoot object.

`object` A kappa4nlsBoot object.

`newdata` The data on which the estimated model is to be fitted.

Value

A generic S3 object with class kappa4nlsBoot.

kappa4nlsBoot.default: A list object (saved using `saveRDS` in the specified location) with the following components:

- intercept: Did the model contain an intercept TRUE/FALSE?
- coefficients: A vector of estimated coefficients.
- bcoefficients: A vector of bootstrap coefficients, resulting from bootstrap estimation.
- se: The standard errors for the estimates resulting from bootstrap estimation.
- error: The value of the objective function.
- errorList: A vector of values of the objective function for each bootstrap sample.
- fitted.values: A vector of estimated values.
- residuals: The residuals resulting from the fitted model.
- call: The call to the function.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.

summary.kappa4nlsBoot: A list of class summary.kappa4nlsBoot with the following components:

- call: Original call to the `kappa4nlsBoot` function.

© University of Pretoria

- coefficients: A matrix with estimates, estimated errors, and 95% parameter confidence intervals (based on the inverse empirical distribution function).
- r.squared: The r^2 coefficient.
- sigma: The residual standard error.
- error: Value of the objective function.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.
- residSum: Summary statistics for the distribution of the residuals.
- errorSum: Summary statistics for the distribution of the value of the objective function.

print.summary.kappa4nlsBoot: The object passed to the function is returned invisibly.

predict.kappa4nlsBoot: A vector of predicted values resulting from the estimated model.

Methods (by class)

- `default`: default method for kappa4nlsBoot.
- `kappa4nlsBoot`: print method for kappa4nlsBoot.
- `kappa4nlsBoot`: summary method for kappa4nlsBoot.
- `summary.kappa4nlsBoot`: print method for summary.kappa4nlsBoot.
- `formula`: formula method for kappa4nlsBoot.
- `kappa4nlsBoot`: predict method for kappa4nlsBoot.

A.3.2 Non-Linear Arc Length Regression

© University of Pretoria

`kappa4alBoot` *Sigmoidal curve fitting.*

Description

Bootstrap estimates, along with standard errors and confidence intervals, of a nonlinear model, resulting from arc length fitting of the four-parameter kappa sigmoidal function.

Usage

```
kappa4alBoot(formula, data = list(), xin, lower, upper, q1, q2, tol, maxiter,
  bootstraps, bootName, ...)

## Default S3 method:
kappa4alBoot(formula, data = list(), xin, lower = c(0, -5,
  -5), upper = c(10, 1, 1), q1, q2, tol = 1e-15, maxiter = 50000,
  bootstraps, bootName, ...)

## S3 method for class 'kappa4alBoot'
print(x, ...)

## S3 method for class 'kappa4alBoot'
summary(object, ...)

## S3 method for class 'summary.kappa4alBoot'
print(x, ...)

## S3 method for class 'formula'
kappa4alBoot(formula, data = list(), xin, lower, upper, q1,
  q2, tol, maxiter, bootstraps, bootName, ...)

## S3 method for class 'kappa4alBoot'
predict(object, newdata = NULL, ...)
```

© University of Pretoria

Arguments

`formula` An LHS ~ RHS formula, specifying the linear model to be estimated.

`data` A data.frame which contains the variables in `formula`.

`xin` Numeric vector of length 3 containing initial values, for σ, h, and k.

`lower` A vector of lower constraints for the parameters to be estimated; defaults to c(0, -5, -5).

`upper` A vector of upper constraints for the parameters to be estimated; defaults to c(10, 1, 1).

`q1, q2` Numeric vectors, for the lower and upper bounds of the intervals over which arc lengths are to be computed.

`tol` Error tolerance level; defaults to 1e-15.

`maxiter` The maximum number of iterations allowed; defaults to 50000.

`bootstraps` An integer giving the number of bootstrap samples.

`bootName` The name of the .rds file to store the kappa4alBoot object. May include a path.

`...` Arguments to be passed on to the differential evolution function `JDEoptim`.

`x` A kappa4alBoot object.

`object` A kappa4alBoot object.

`newdata` The data on which the estimated model is to be fitted.

Value

A generic S3 object with class kappa4alBoot.

kappa4alBoot.default: A list object (saved using `saveRDS` in the specified location) with the following components:

- intercept: Did the model contain an intercept TRUE/FALSE?
- coefficients: A vector of estimated coefficients.
- bcoefficients: A vector of bootstrap coefficients, resulting from bootstrap estimation.

© University of Pretoria

- se: The standard errors for the estimates resulting from bootstrap estimation.
- error: The value of the objective function.
- errorList: A vector of values of the objective function for each bootstrap sample.
- fitted.values: A vector of estimated values.
- residuals: The residuals resulting from the fitted model.
- call: The call to the function.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.

summary.kappa4alBoot: A list of class summary.kappa4alBoot with the following components:

- call: Original call to the `kappa4alBoot` function.
- coefficients: A matrix with estimates, estimated errors, and 95% parameter confidence intervals (based on the inverse empirical distribution function).
- arclengths: A matrix of the arc length segments that were matched, for the dependent and independent variables.
- r.squared: The r^2 coefficient.
- sigma: The residual standard error.
- error: Value of the objective function.
- time: Min, mean and max time incurred by the computation, as obtained from `comm.timer`.
- residSum: Summary statistics for the distribution of the residuals.
- errorSum: Summary statistics for the distribution of the value of the objective function.

print.summary.kappa4alBoot: The object passed to the function is returned invisibly.

predict.kappa4alBoot: A vector of predicted values resulting from the estimated model.

© University of Pretoria

Methods (by class)

- `default`: default method for kappa4alBoot.
- `kappa4alBoot`: print method for kappa4alBoot.
- `kappa4alBoot`: summary method for kappa4alBoot.
- `summary.kappa4alBoot`: print method for summary.kappa4alBoot.
- `formula`: formula method for kappa4alBoot.
- `kappa4alBoot`: predict method for kappa4alBoot.

A.4 Summary

Key functions and methods from the alR package were presented in this appendix. The source code and complete package documentation are available at `https://github.com/mtloots/alR`.

© University of Pretoria

Appendix B

Simulation Results for the Method of Arc Lengths

In this Appendix, simulation results for the method of arc lengths, applied to the normal distribution, are presented. Here, the results presented in Chapter 3, Section 3.3 are extended to include a larger number of parameter choices for the normal distribution. For each of the resulting parameter choices (μ and σ) of the normal distribution, results for the estimated bias and rMSE values (in ascending order of rMSE) for the median values, of the parameters estimated from 1,000 samples of size 1,000, are given. Performance results are also provided for all these (μ, σ) combinations, by varying the sample size. The estimation techniques to be compared correspond to those outlined in Section 3.3, but is reproduced in Section B.1 for completeness. The remaining sections correspond to a particular (μ, σ) combination.

B.1 Introduction

The method of arc lengths (resulting from the two proposed sample arc length statistics given in (3.2.1) and (3.2.2)), ML, CvM, and AD estimation techniques are compared. For each of the proposed sample arc length statistics, five variants were included depending on the interval $[a, b]$, where the two end points are specified in terms of the sample quantiles:

(a) $[0.01, 0.99]$,

© University of Pretoria

(b) $[0.025, 0.975]$,

(c) $[0.05, 0.95]$,

(d) $[0.075, 0.925]$, and

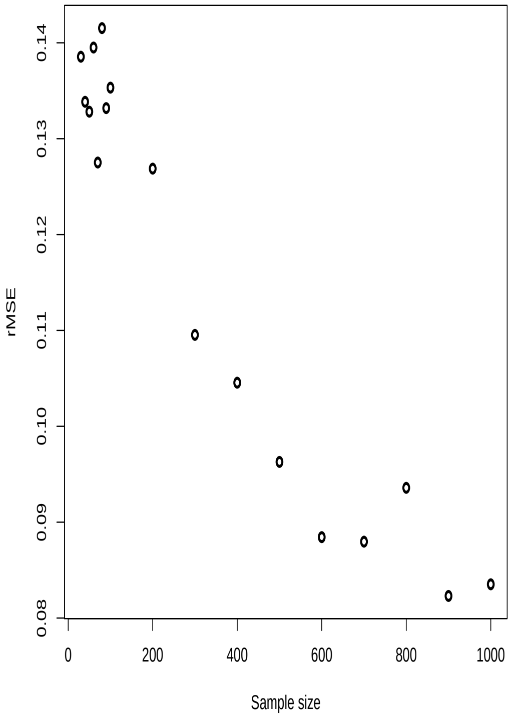
(e) $[0.1, 0.9]$.

B.2 Simulation Results for $N(-2, 0.5)$

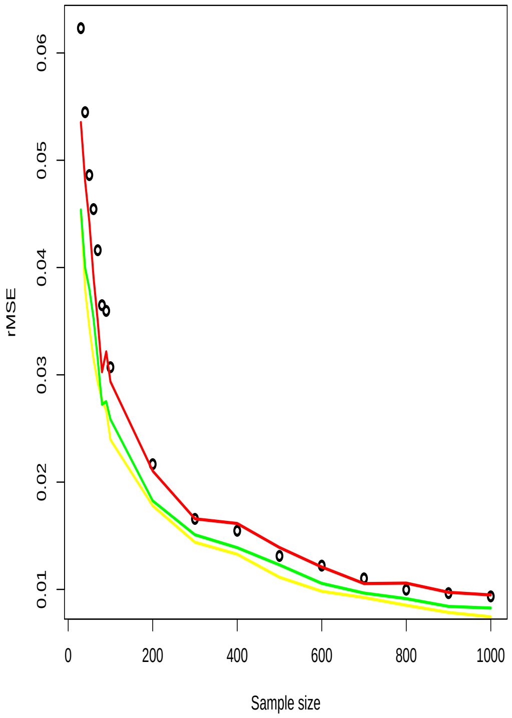
	bias	rMSE
MLE	-0.00006	0.01072
AD	-0.00025	0.01087
CvM	0.00066	0.01114
s1b	0.01763	0.03340
s1d	0.03984	0.04221
s1c	0.02891	0.04503
s1a	0.05921	0.07764
s1e	0.01506	0.09161
s2c	0.11299	0.11757
s2d	0.11563	0.12153
s2b	0.12032	0.12417
s2a	0.13285	0.13442
s2e	0.12910	0.13542

Table B.1: Estimated bias and rMSE for μ=-2 and σ=0.5.

© University of Pretoria



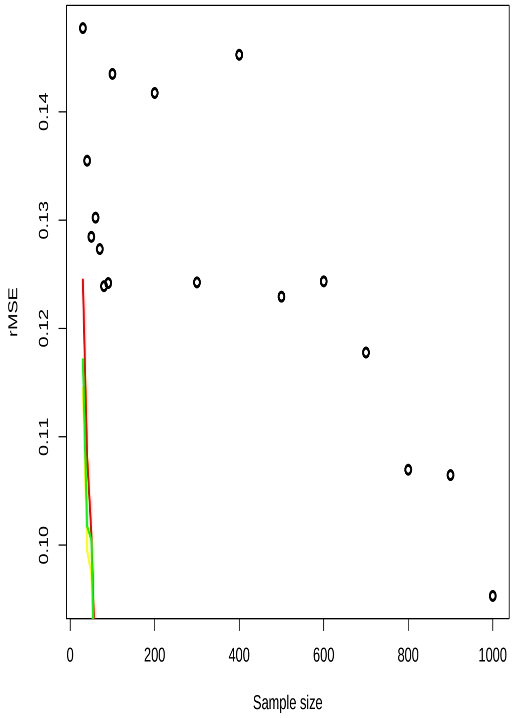
(a) Estimated rMSE for $\mu = -2$.



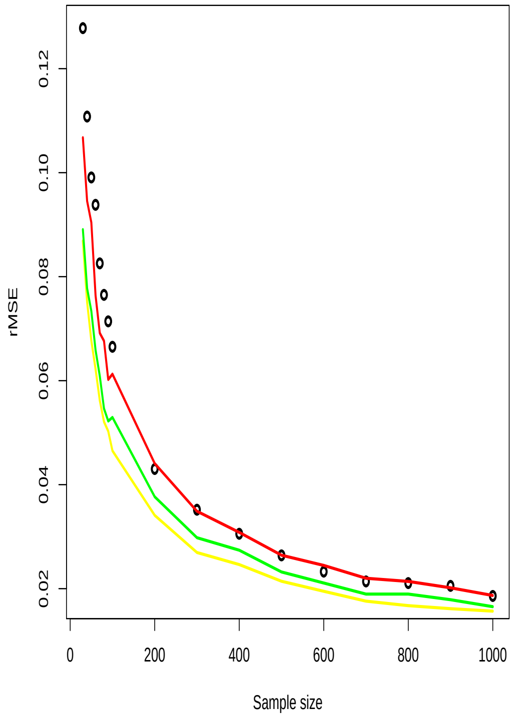
(b) Estimated rMSE for σ=0.5.

Figure B.1: Estimated rMSE for $\mu = -2$ and $\sigma = 0.5$ for varying sample sizes.

© University of Pretoria



(a) Estimated rMSE for $\mu = -2$.

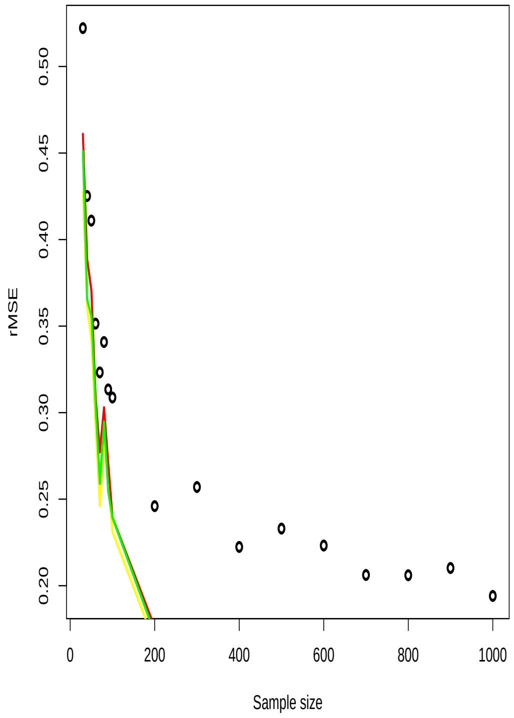


(b) Estimated rMSE for σ=1.

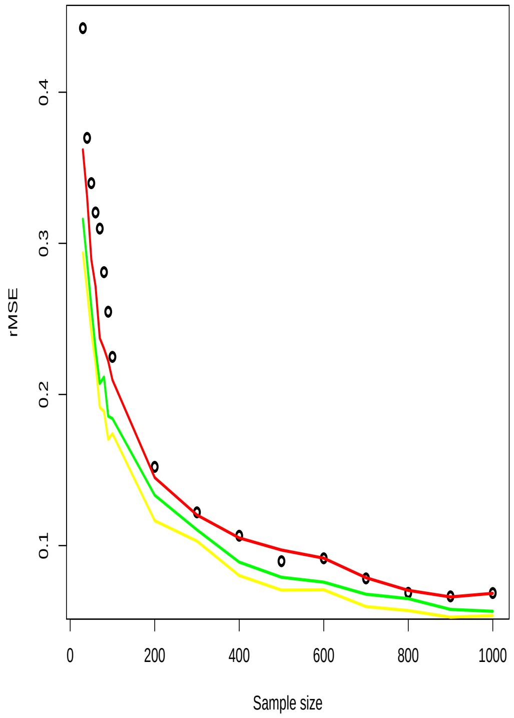
Figure B.2: Estimated rMSE for $\mu = -2$ and $\sigma = 1$ for varying sample sizes.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA



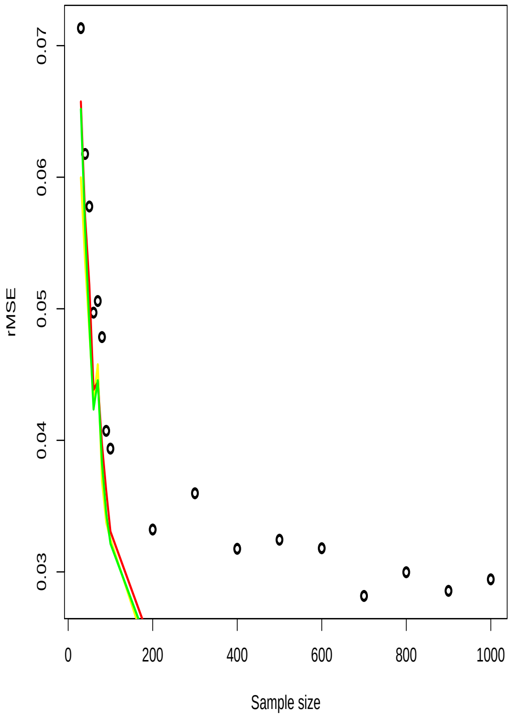
(a) Estimated rMSE for $\mu = -2$.



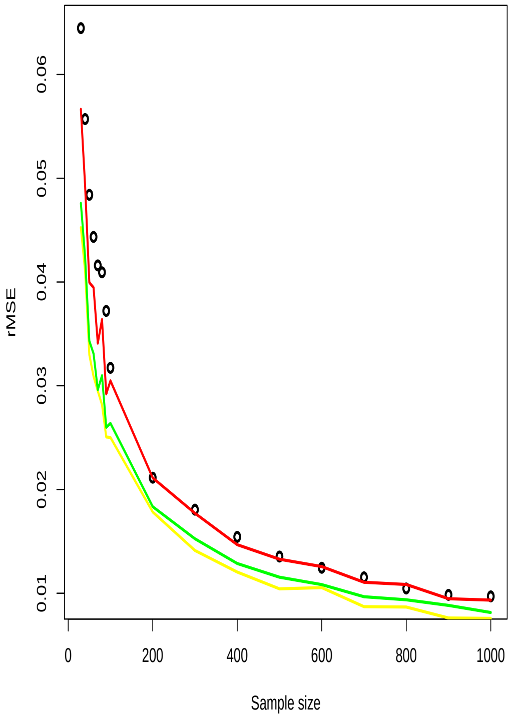
(b) Estimated rMSE for σ=3.5.

Figure B.3: Estimated rMSE for $\mu = -2$ and $\sigma = 3.5$ for varying sample sizes.

© University of Pretoria



(a) Estimated rMSE for $\mu = 0$.



(b) Estimated rMSE for σ=0.5.

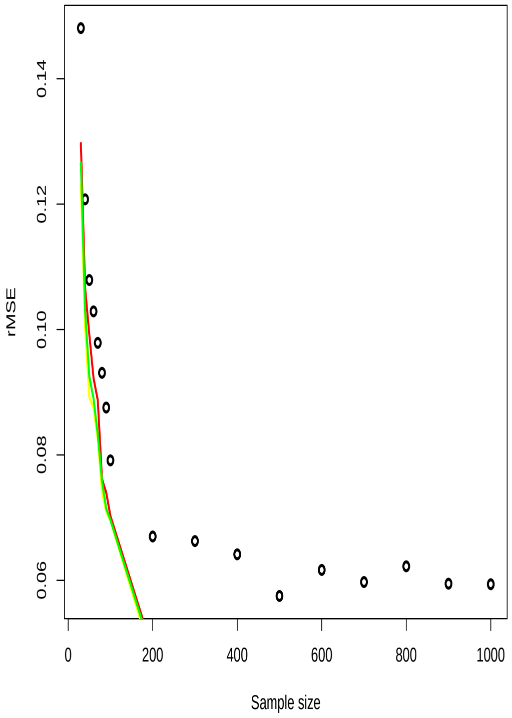
Figure B.4: Estimated rMSE for $\mu = 0$ and $\sigma = 0.5$ for varying sample sizes.

© University of Pretoria

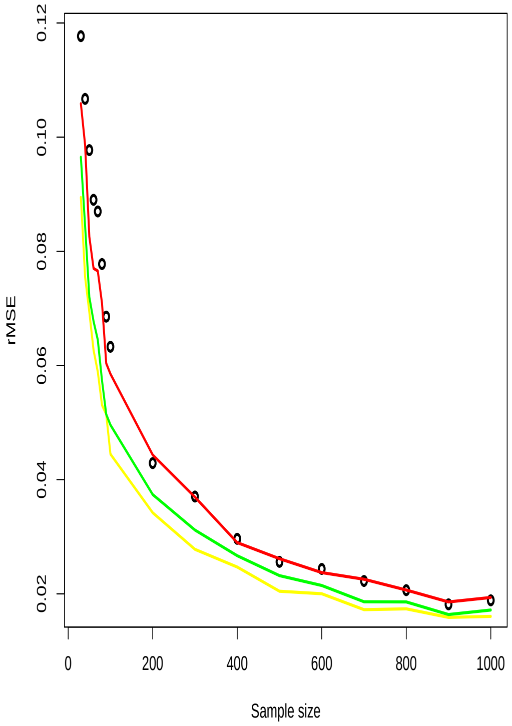
	bias	rMSE
MLE	-0.00104	0.00770
AD	-0.00080	0.00825
s1a	-0.00246	0.00911
CvM	-0.00155	0.00922
s2b	0.00607	0.01029
s2a	0.00714	0.01092
s2c	0.00522	0.01107
s2d	0.00551	0.01181
s2e	0.00594	0.01406
s1b	-0.02060	0.02061
s1c	-0.05053	0.05053
s1d	-0.07608	0.07608
s1e	-0.10350	0.10350

Table B.2: Estimated bias and rMSE for μ=-2 and σ=0.5.

© University of Pretoria



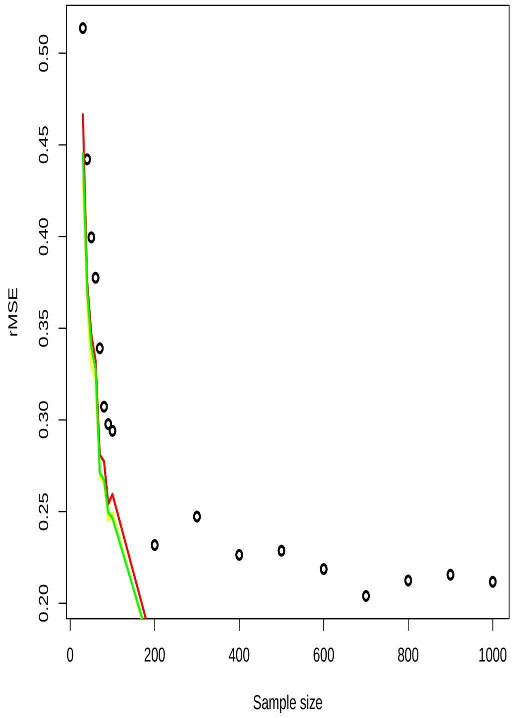
(a) Estimated rMSE for $\mu = 0$.



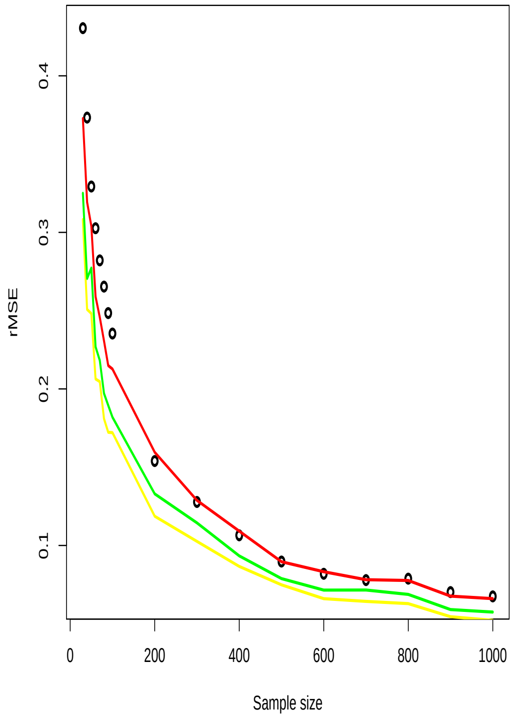
(b) Estimated rMSE for σ=1.

Figure B.5: Estimated rMSE for $\mu = 0$ and $\sigma = 1$ for varying sample sizes.

© University of Pretoria



(a) Estimated rMSE for $\mu = 0$.

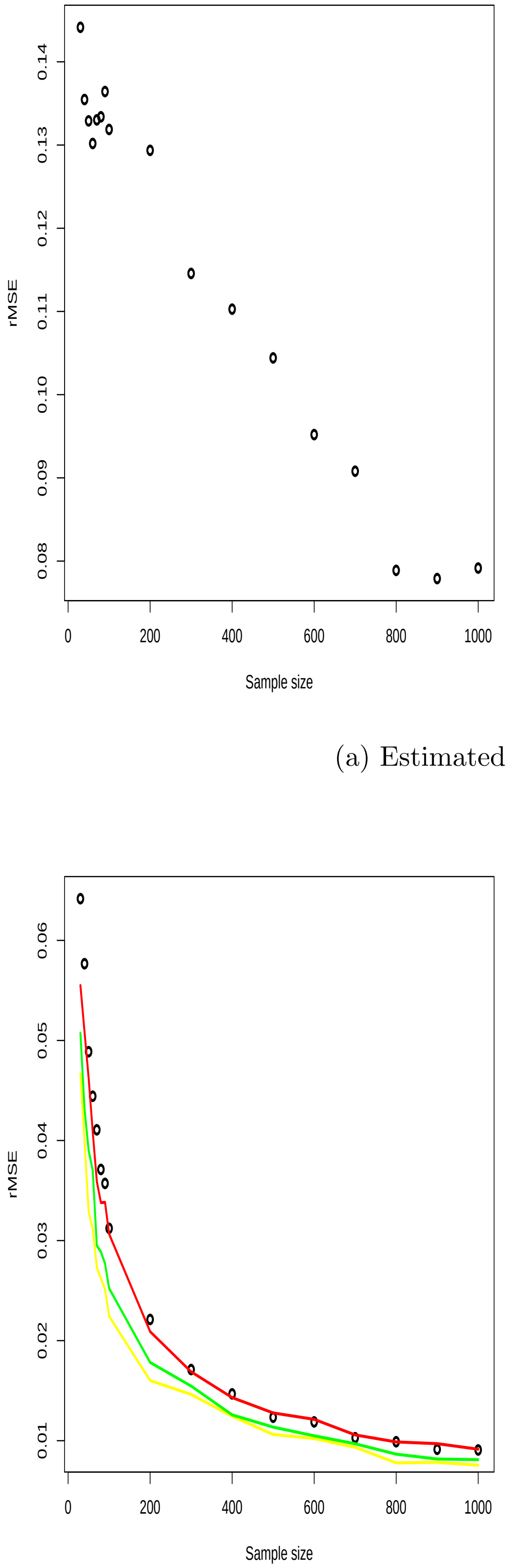


(b) Estimated rMSE for σ=3.5.

Figure B.6: Estimated rMSE for $\mu = 0$ and $\sigma = 3.5$ for varying sample sizes.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

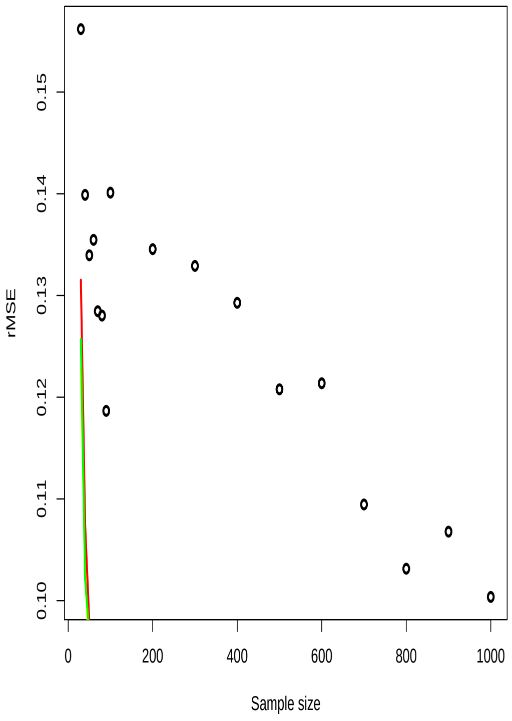


(a) Estimated rMSE for $\mu = 2$.

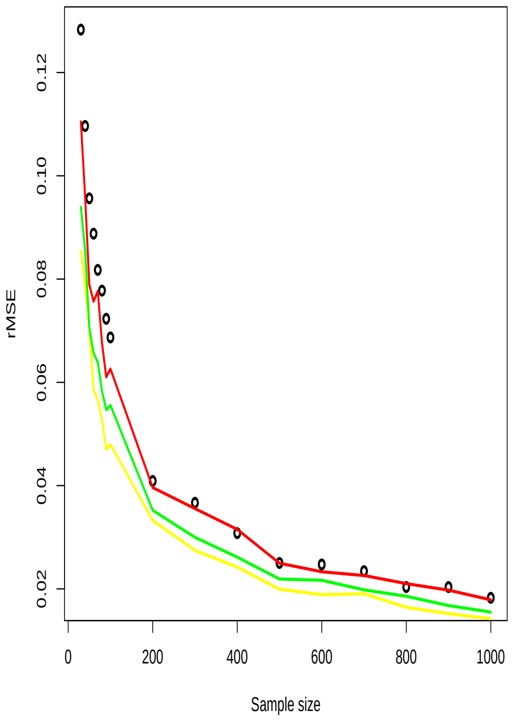
(b) Estimated rMSE for σ=0.5.

Figure B.7: Estimated rMSE for $\mu = 2$ and $\sigma = 0.5$ for varying sample sizes.

© University of Pretoria



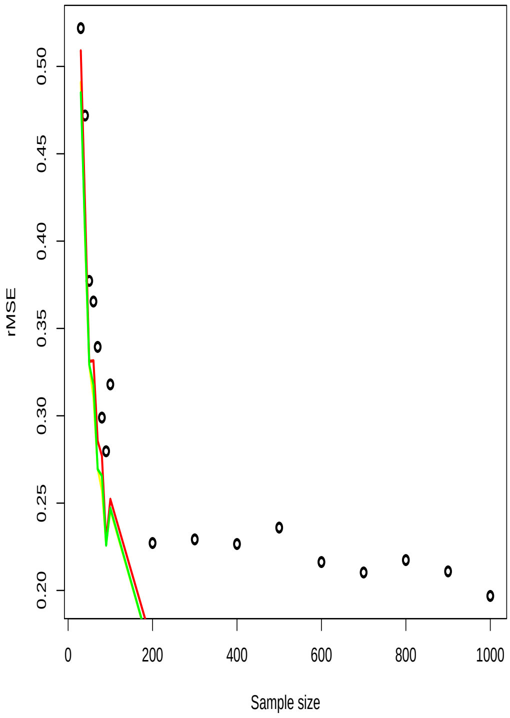
(a) Estimated rMSE for $\mu = 2$.



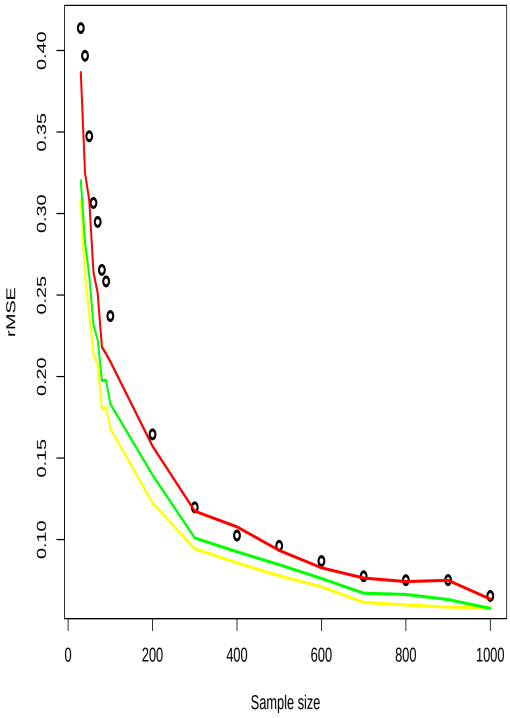
(b) Estimated rMSE for σ=1.

Figure B.8: Estimated rMSE for $\mu = 2$ and $\sigma = 1$ for varying sample sizes.

© University of Pretoria



(a) Estimated rMSE for $\mu = 2$.



(b) Estimated rMSE for σ=3.5.

Figure B.9: Estimated rMSE for $\mu = 2$ and $\sigma = 3.5$ for varying sample sizes.

© University of Pretoria

	bias	rMSE
AD	0.00159	0.02247
MLE	0.00204	0.02254
CvM	0.00126	0.02267
s1b	0.02074	0.04976
s1e	0.04274	0.06001
s1c	0.02995	0.08081
s1d	0.09801	0.09832
s1a	0.08571	0.10128
s2d	0.14946	0.15240
s2e	0.15296	0.15386
s2c	0.15346	0.15669
s2b	0.16216	0.16242
s2a	0.16603	0.16615

Table B.3: Estimated bias and rMSE for μ=-2 and σ=1.

B.3 Simulation Results for $N(-2, 1)$

B.4 Simulation Results for $N(-2, 3.5)$

B.5 Simulation Results for $N(0, 0.5)$

B.6 Simulation Results for $N(0, 1)$

B.7 Simulation Results for $N(0, 3.5)$

B.8 Simulation Results for $N(2, 0.5)$

B.9 Simulation Results for $N(2, 1)$

B.10 Simulation Results for $N(2, 3.5)$

B.11 Summary

Using the method of arc lengths with ${}_1f^{[q_{0.01}, q_{0.99}]}$ yielded the smallest bias and rMSE values for simultaneously estimating μ and σ. All sample arc length statistics performed

© University of Pretoria

	bias	rMSE
MLE	0.00091	0.01584
AD	0.00060	0.01660
s1a	-0.00093	0.01832
CvM	-0.00083	0.01874
s2a	0.01896	0.02392
s2b	0.01891	0.02525
s2c	0.01899	0.02609
s2d	0.01970	0.02691
s2e	0.01960	0.02849
s1b	-0.03849	0.03860
s1c	-0.10036	0.10036
s1d	-0.16021	0.16021
s1e	-0.21289	0.21289

Table B.4: Estimated bias and rMSE for μ=-2 and σ=1.

worse than the MLE, AD and CvM methods for estimating μ. The same holds true for the estimation of σ, accept for ${}_1\int^{[q_{0.01},q_{0.99}]}$ and CvM, for which the ordering are interchanged in certain cases.

The rMSE values for the method of arc lengths with ${}_1\int^{[q_{0.01},q_{0.99}]}$ decrease for increasing sample sizes. For estimating σ, this corresponds well to the values obtained for the CvM method, but are not as strong as that obtained for the MLE and AD methods.

© University of Pretoria

	bias	rMSE
MLE	0.01087	0.07410
AD	0.01130	0.07786
CvM	0.01288	0.07813
s1b	0.10162	0.14077
s1c	0.14525	0.18590
s1a	0.19016	0.20837
s2e	0.23234	0.24271
s2d	0.24803	0.25292
s2a	0.25027	0.25466
s2c	0.25603	0.25810
s2b	0.26368	0.26918
s1e	0.28161	0.28161
s1d	0.28937	0.29609

Table B.5: Estimated bias and rMSE for μ=-2 and σ=3.5.

	bias	rMSE
MLE	-0.00283	0.05254
AD	-0.00155	0.05966
CvM	0.00061	0.06765
s1a	-0.00776	0.06851
s2a	0.06200	0.08184
s2b	0.06885	0.08602
s2c	0.07539	0.09110
s2d	0.08150	0.10362
s2e	0.08922	0.11406
s1b	-0.13474	0.13521
s1c	-0.35352	0.35352
s1d	-0.56281	0.56281
s1e	-0.74744	0.74744

Table B.6: Estimated bias and rMSE for μ=-2 and σ=3.5.

© University of Pretoria

	bias	rMSE
AD	-0.00007	0.00948
MLE	0.00015	0.00974
CvM	-0.00029	0.00977
s1b	0.01305	0.01927
s1c	0.02132	0.02557
s1a	0.02436	0.02745
s1d	0.02261	0.03087
s2a	0.03344	0.03417
s2e	0.03273	0.03442
s2b	0.03426	0.03503
s2d	0.03443	0.03520
s2c	0.03507	0.03556
s1e	0.05028	0.05028

Table B.7: Estimated bias and rMSE for μ=0 and σ=0.5.

	bias	rMSE
MLE	-0.00016	0.00792
AD	-0.00046	0.00866
s1a	-0.00022	0.00943
CvM	-0.00101	0.00982
s2a	0.01059	0.01229
s2b	0.01096	0.01278
s2c	0.01155	0.01361
s2d	0.01213	0.01460
s2e	0.01210	0.01578
s1b	-0.01907	0.01912
s1c	-0.04828	0.04828
s1d	-0.07452	0.07452
s1e	-0.09949	0.09949

Table B.8: Estimated bias and rMSE for μ=0 and σ=0.5.

© University of Pretoria

	bias	rMSE
MLE	0.00097	0.02182
AD	0.00127	0.02188
CvM	0.00207	0.02202
s1b	0.02540	0.03881
s1c	0.04050	0.05153
s1a	0.05126	0.05509
s2e	0.06404	0.06618
s2d	0.06669	0.06785
s2a	0.06650	0.06872
s2b	0.06943	0.06992
s2c	0.07017	0.07123
s1e	0.08112	0.08136
s1d	0.07916	0.08266

Table B.9: Estimated bias and rMSE for μ=0 and σ=1.

	bias	rMSE
MLE	-0.00216	0.01549
AD	-0.00169	0.01637
s1a	-0.00345	0.01839
CvM	-0.00204	0.01868
s2a	0.01757	0.02489
s2b	0.01858	0.02571
s2c	0.01938	0.02638
s2d	0.01899	0.02794
s2e	0.02033	0.02976
s1b	-0.03980	0.03989
s1c	-0.10192	0.10192
s1d	-0.16084	0.16084
s1e	-0.21481	0.21481

Table B.10: Estimated bias and rMSE for μ=0 and σ=1.

© University of Pretoria

	bias	rMSE
MLE	0.00561	0.08044
AD	0.00578	0.08117
CvM	0.00799	0.08486
s1b	0.10402	0.14333
s1c	0.14193	0.17495
s1a	0.18542	0.20705
s2e	0.23322	0.24246
s2d	0.24376	0.24918
s2a	0.24670	0.25151
s2b	0.24762	0.25291
s2c	0.25634	0.26252
s1e	0.28892	0.28892
s1d	0.29646	0.30524

Table B.11: Estimated bias and rMSE for μ=0 and σ=3.5.

	bias	rMSE
MLE	-0.00568	0.05577
AD	-0.00108	0.06063
CvM	-0.00237	0.06906
s1a	-0.00731	0.07027
s2a	0.06397	0.08520
s2b	0.06810	0.08653
s2c	0.07346	0.09464
s2d	0.07793	0.10631
s2e	0.08362	0.11173
s1b	-0.13743	0.13824
s1c	-0.35376	0.35376
s1d	-0.56209	0.56209
s1e	-0.75169	0.75169

Table B.12: Estimated bias and rMSE for μ=0 and σ=3.5.

© University of Pretoria

	bias	rMSE
AD	-0.00027	0.01093
MLE	-0.00013	0.01107
CvM	-0.00059	0.01150
s1b	0.01704	0.03667
s1d	0.04165	0.04305
s1c	0.03247	0.04825
s1a	0.06260	0.08087
s1e	0.01917	0.08915
s2c	0.11502	0.11815
s2d	0.12322	0.12526
s2a	0.12209	0.12601
s2b	0.12340	0.12645
s2e	0.13347	0.13498

Table B.13: Estimated bias and rMSE for μ=2 and σ=0.5.

	bias	rMSE
MLE	-0.00097	0.00731
AD	-0.00064	0.00805
s1a	-0.00267	0.00932
CvM	-0.00133	0.00933
s2b	0.00563	0.00997
s2c	0.00569	0.01064
s2a	0.00694	0.01077
s2d	0.00579	0.01097
s2e	0.00651	0.01305
s1b	-0.02075	0.02098
s1c	-0.05072	0.05072
s1d	-0.07583	0.07583
s1e	-0.10301	0.10301

Table B.14: Estimated bias and rMSE for μ=2 and σ=0.5.

© University of Pretoria

	bias	rMSE
MLE	0.00320	0.02049
CvM	0.00359	0.02060
AD	0.00295	0.02087
s1b	0.02300	0.04957
s1e	0.03746	0.05971
s1c	0.04644	0.08729
s1a	0.08355	0.09824
s1d	0.10210	0.10254
s2d	0.14307	0.14375
s2e	0.14556	0.14890
s2c	0.15196	0.15399
s2b	0.15756	0.15923
s2a	0.16576	0.16606

Table B.15: Estimated bias and rMSE for μ=2 and σ=1.

	bias	rMSE
MLE	-0.00141	0.01498
AD	-0.00037	0.01605
CvM	-0.00124	0.01811
s1a	-0.00305	0.01891
s2a	0.01665	0.02246
s2b	0.01699	0.02267
s2c	0.01763	0.02393
s2d	0.01899	0.02618
s2e	0.01851	0.02921
s1b	-0.04018	0.04037
s1c	-0.10202	0.10202
s1d	-0.16047	0.16047
s1e	-0.21216	0.21216

Table B.16: Estimated bias and rMSE for μ=2 and σ=1.

© University of Pretoria

	bias	rMSE
MLE	-0.00103	0.07285
AD	-0.00058	0.07686
CvM	-0.00212	0.08028
s1b	0.09716	0.13748
s1c	0.14438	0.18765
s1a	0.17579	0.19975
s2e	0.22476	0.23433
s2b	0.23539	0.24263
s2d	0.23676	0.24575
s2a	0.24221	0.24808
s2c	0.24751	0.25267
s1e	0.27871	0.27911
s1d	0.28460	0.29591

Table B.17: Estimated bias and rMSE for μ=2 and σ=3.5.

	bias	rMSE
MLE	-0.00122	0.05204
AD	0.00210	0.05767
CvM	-0.00148	0.06477
s1a	-0.00415	0.06553
s2a	0.06506	0.08510
s2b	0.06905	0.08712
s2c	0.07328	0.09199
s2d	0.07606	0.09958
s2e	0.07942	0.10760
s1b	-0.13569	0.13599
s1c	-0.35565	0.35565
s1d	-0.55801	0.55801
s1e	-0.74934	0.74934

Table B.18: Estimated bias and rMSE for μ=2 and σ=3.5.

© University of Pretoria

Appendix C

Machine Learning

In this Appendix, two popular "machine learning" techniques are used for analysing the data used in Chapter 4. This is done as benchmark techniques for that, newly proposed in this thesis. The random forest (RF) is applied in Section C.1, while support vector regression (SVR) follows in Section C.2.

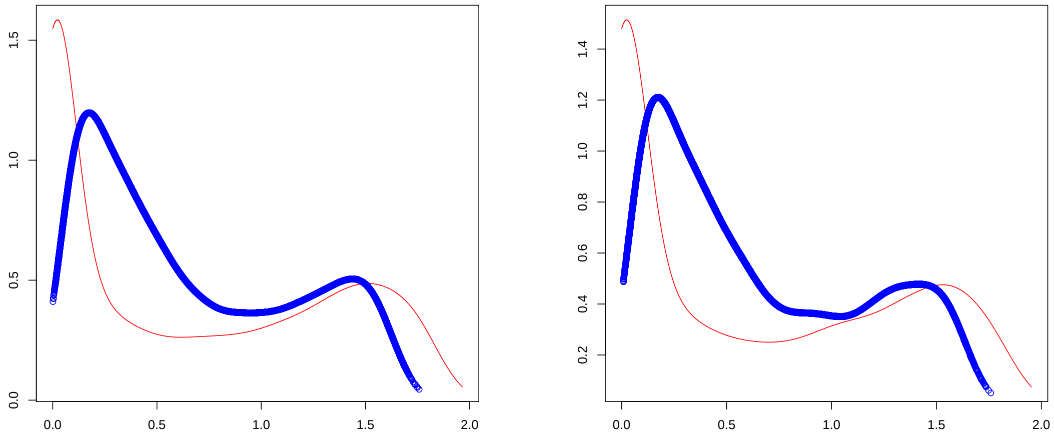
C.1 Random Forest

The "randomForest" function from the similarly named R package by Liaw and Wiener (2002) [52] was used for fitting the RF in this section. The output is shown below:

```

Call:
 randomForest(formula = FPKM_lambda ~ H3K4me3_signal_bin25 +
    H3K27me3_signal_bin21,       data = training, ntree = 500, xtest =
    data.frame(testing$H3K4me3_signal_bin25,
    testing$H3K27me3_signal_bin21), ytest = testing$FPKM_lambda)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 0.22259
                    % Var explained: 46.9
                       Test set MSE: 0.22
                    % Var explained: 47.18
```

© University of Pretoria

Figure C.1 compares the KDE's for the dependent variable, with the predictions resulting from the RF fit. In each case, the smoothing parameter is also reported.



(a) KDE's for transformed FPKM signal and the random forest fit thereof, with bandwidths 0.09333 and 0.07121 respectively (training set).

(b) KDE's for transformed FPKM signal and the random forest fit thereof, with bandwidths 0.10109 and 0.07724 respectively (testing set).

Figure C.1: Plots for transformed expression data resulting from a random forest.

C.2 Support Vector Regression

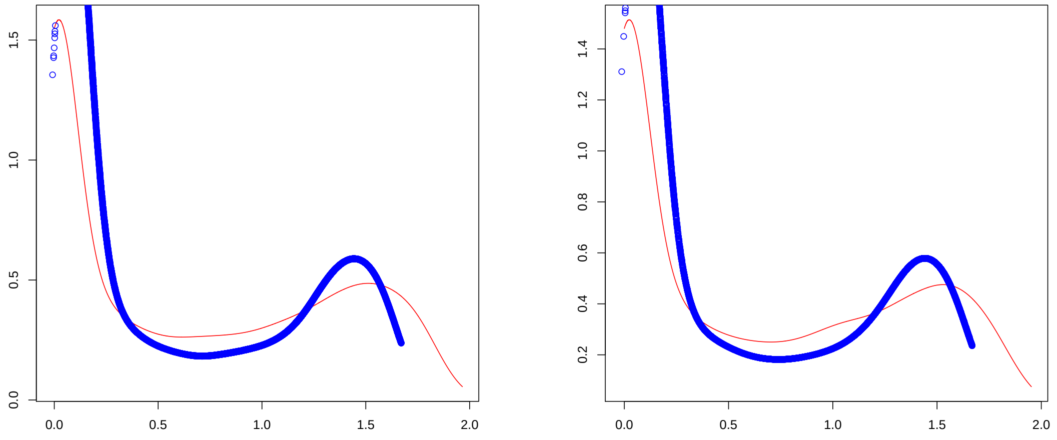
The "svm" function from the "e1071" by Meyer et al (2015) [59] was used for performing SVR. All tuning parameters were left at their default values, and the output provided below:

```

Call:
svm(formula = FPKM_lambda ~ H3K4me3_signal_bin25 +
    H3K27me3_signal_bin21,
     data = training)



Parameters:
```

© University of Pretoria

```
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.5
    epsilon:  0.1


Number of Support Vectors:  16836
```

Figure C.2 compares the KDE's for the dependent variable, with the predictions resulting from SVR. In each case, the smoothing parameter is also reported.



(a) KDE's for transformed FPKM signal and the support vector regression fit thereof, with bandwidths 0.09333 and 0.08369 respectively (training set).

(b) KDE's for transformed FPKM signal and the support vector regression fit thereof, with bandwidths 0.10109 and 0.09052 respectively (testing set).

Figure C.2: Plots for transformed expression data resulting from support vector regression.

C.3 Summary

Comparison of Table C.1 with Table 4.2 clearly indicates that neither the RF, nor SVR yields superior results, without serious parameter tuning. Note that the number of proportions used in the Bhattacharyya test, were reduced to 5, and may therefore produce

© University of Pretoria

	RF	SVR
r2 (Train)	0.4690	0.4616
r2 (Test)	0.4718	0.4593
D2 (Train)	0.2010	0.2507
D2 (Test)	0.2069	0.2518

Table C.1: Comparison of performance results for random forest and support vector regression methods.

an artificially small divergence measure; for larger values, this measure couldn't be computed because of empty cells.

© University of Pretoria

Appendix D

Acronyms

The acronyms used in this thesis are listed in this appendix. The list is arranged alphabetically and the respective acronyms typeset in bold, with its corresponding meaning alongside.

AD	Anderson-Darling
ALR	arc length regression
CDF	cumulative distribution function
CvM	Cramér-von Mises
GoF	goodness-of-fit
IID	independent and identically distributed
KDE	kernel density estimate
ML	maximum-likelihood
MLE	maximum-likelihood estimation
MMM	moment matching
NALR	non-linear arc length regression
NLS	non-linear least squares
OLS	ordinary least squares

© University of Pretoria

PDF	probability density function
QF	quantile function
RF	random forest
rMSE	root-mean-square error
RSS	residual standard error
SVR	support vector regression
TLS	total least squares

© University of Pretoria

Appendix E

Symbols

In this appendix, the symbols used throughout the thesis are defined, under the chapter in which they first appear.

E.1 Chapter 2: Preliminaries

$\mathcal{S}_f^{[a,b]}$	Theoretical arc length	[Eq. (2.1.1), pg. 7]

E.2 Chapter 3: Estimation

${}_1\!\int^{[a,b]}$	Discrete sample arc length statistic	[Eq. (3.2.1), pg. 11]
${}_2\!\int^{[a,b]}$	Continuous sample arc length statistic	[Eq. (3.2.2), pg. 12]

E.3 Chapter 5: Non-Linear Arc Length Regression

$\int(a,b)$	Sample arc length statistic	[Eq. (5.1.1), pg. 42]

© University of Pretoria

Appendix F

Derived Publications

The following is a list of the publications derived from, and that are associated with, this thesis.

- Walter W. Focke, Isbe van der Westhuizen, Ndeke Musee and Mattheüs T. Loots. Kinetic interpretation of log-logistic dose-time response curves. *Scientific Reports* (Accepted for publication).
- Steven G. Hussey, Mattheüs T. Loots, Karen van der Merwe, Eshchar Mizrachi and Alexander A. Myburg. H3K4me3 and H3K27me3 in developing secondary xylem. *Scientific Reports* (Accepted for publication).
- M. Theodor Loots and Andriëtte Bekker. A method of arc lengths with application to the normal probability density function. *Submitted.*
- M. Theodor Loots and Andriëtte Bekker. Arc length regression. *Submitted.*
- M. Theodor Loots and Andriëtte Bekker. Fitting the four-parameter kappa sigmoidal function using non-linear arc length regression. *Submitted.*

© University of Pretoria

Index

© University of Pretoria

© University of Pretoria