
A New Method for Transforming Data to Normality with Application to Density Estimation

Gerhard Koekemoer, M.Sc.

Thesis submitted for the degree Philosophiae Doctor in
Statistics at the North-West University

Promoter: Prof. J.W.H. Swanepoel

November 2004
Potchefstroom

Summary

One of the main objectives of this dissertation is to derive efficient nonparametric estimators for an unknown density f . It is well known that the ordinary kernel density estimator has, despite of several good properties, some drawbacks. For example, it suffers from boundary bias and it also exhibits spurious bumps in the tails. Various solutions to overcome these defects are presented in this study, which include the application of a transformation kernel density estimator. The latter estimator (if implemented correctly) is pursued as a simultaneous solution for both boundary bias and spurious bumps in the tails. The estimator also has, among others, the ability to detect and estimate density modes more effectively.

To apply the transformation kernel density estimator an effective transformation of the data is required. To achieve this objective, an extensive discussion of parametric transformations introduced and studied in the literature is presented firstly, emphasizing the practical feasibility of these transformations. Secondly, known methods of estimating the parameters associated with these transformations are discussed (e.g. profile maximum likelihood), and two new estimation techniques, referred to as the minimum residual and minimum distance methods, are introduced. Furthermore, new procedures are developed to select a parametric transformation that is suitable for application to a given set of data. Finally, utilizing the above techniques, the desired optimal transformation to any target distribution (e.g. the normal distribution) is introduced, which has the property that it can also be iterated. A polynomial approximation of the optimal transformation function is presented. *It is shown that the performance of this transformation exceeds that of any transformation available in the literature.*

In the context of transformation kernel density estimation, we present a comprehensive literature study of current methods available and then introduce the new semi-parametric

transformation estimation procedure based on the optimal transformation of data to normality. However, application of the optimal transformation in this context requires special attention. In order to create a density estimator that addresses both boundary bias and spurious bumps in the tails simultaneously in an automatic way, a generalized bandwidth adaptation procedure is developed, which is applied in conjunction with a newly developed constant shift procedure.

Furthermore, the optimal transformation function is based on a kernel distribution function estimator. A new data-based smoothing parameter (bandwidth selector) is invented, and it is shown that this selector has better performance than a well established bandwidth selector proposed in the literature.

To evaluate the performance of the newly proposed semi-parametric transformation estimation procedure, a simulation study is presented based on densities that consist of a wide range of forms. Some of the main results derived in the Monte Carlo simulation study include that:

- *the proposed optimal transformation function can take on all the possible shapes of a parametric transformation as well as any combination of these shapes, which result in high p-values when testing normality of the transformed data.*
- *the new minimum residual and minimum distance techniques contribute to better transformations to normality, when a parametric transformation is applicable.*
- *the newly proposed semi-parametric transformation kernel density estimator performs well for unimodal, low and high kurtosis densities. Moreover, it estimates densities with much curvature (e.g. modes and valleys) more effectively than existing procedures in the literature.*
- *the new transformation density estimator does not exhibit spurious bumps in the tail regions.*
- *boundary bias is addressed automatically.*

In conclusion, practical examples based on real-life data are presented.

Opsomming

Een van die hoof mikpunte van hierdie proefskrif is om doeltreffende nie-parametriese beramers vir 'n onbekende digtheidsfunksie f af te lei. Dit is alombekend dat die gewone kerndigtheidsfunksie beramer, ten spyte van verskeie goeie eienskappe, ook sekere defekte besit. Voorbeelde hiervan is grenssydigheid asook die voorkoms van kunsmatige bulte in die stertgebiede. Verskeie oplossings om hierdie tekortkominge aan te spreek, word in hierdie studie gegee, wat die toepassing van 'n transformasie kerndigtheidsfunksie beramer insluit. Laasgenoemde beramer (indien korrek toegepas) word voorgestel as 'n gelyktydige oplossing vir beide grenssydigheid en die voorkoms van kunsmatige bulte in die sterte. Die beramer besit ook, onder andere, die vermoë om modusse meer effektief waar te neem en te beraam.

'n Effektiewe datatransformasie word benodig om die transformasie kerndigtheidsfunksie beramer te kan implementeer. Ten einde hierdie mikpunt te verwesenlik, word daar eerstens 'n uitgebreide bespreking van bestaande parametriese transformasies in die literatuur gegee, en die praktiese toepasbaarheid van die transformasies word bespreek. Tweedens, word bekende metodes van die beraming van parameters wat geassosieer word met hierdie transformasies, bespreek (bv. profiel maksimumaanneemlikheid). Verder word twee nuwe beramingsmetodes, nl. die minimum residu metode en die minimum afstand metode, voorgestel. Nuwe prosedures word ook ontwikkel vir die seleksie van 'n parametriese transformasie wat geskik is om toegepas te word op 'n gegewe dataset. Laastens, word die optimale transformasie na enige teikenverdeling (bv. die normaalverdeling) m.b.v. bogenoemde tegnieke bekendgestel. 'n Polinoombenadering van die optimale transformasiefunksie word gegee. *Dit word aangetoon dat die gedrag van hierdie transformasie beter vaar as enige transformasie in die literatuur.*

'n Omvattende literatuurstudie van bestaande transformasie kerndigtheidsfunksie be-

ramers word gegee. Hierna word 'n nuwe semi-parametriese transformasie beramingsprosedure, wat gebaseer is op die optimale transformasie van data na normaliteit, bekendgestel. Vir die korrekte toepassing van laasgenoemde prosedure, word 'n algemene bandwydte aanpassingsprosedure ontwikkel, wat in samewerking met 'n nuwe konstante skuifparameter toegepas word.

Die optimale transformasiefunksie is gebaseer op 'n kerndistribusiefunksie beramer. 'n Nuwe data-gebaseerde gladstrykparameter word ontwikkel, en dit word aangetoon dat hierdie data-gebaseerde gladstrykparameter beter vertoon as voorgestelde metodes in die literatuur.

Ten einde nuutvoorgestelde prosedures te evalueer, word 'n omvattende Monte Carlo studie uitgevoer. Die hoofresultate wat verkry is uit hierdie studie bestaan daaruit dat:

- *die voorgestelde optimale transformasiefunksie alle vorms van 'n parametriese transformasie, en enige kombinasie van hierdie vorms, kan aanneem. Dit lei tot hoë p-waardes wanneer die getransformeerde data vir normaliteit getoets word.*
- *die nuwe minimum residu tegniek en minimum afstand tegniek dra by tot beter transformasies na normaliteit, indien 'n parametriese transformasie van toepassing is.*
- *die nuwe semi-parametriese transformasie kerndigtheidsfunksie beramer is effektief om unimodale, lae en hoë kurtose digtheidsfunksies, asook digtheidsfunksies met baie kurwes te beraam.*
- *die nuwe transformasie digtheidsfunksie beramer besit nie kunsmatige bulte in die stertgebiede nie.*
- *grenssydigheid word outomaties aangespreek.*

Ten slotte, word die nuutvoorgestelde prosedures op werklike data toegepas.

Bedankings

Die skrywer wil hiermee graag die volgende bedankings doen:

- Prof. J.W.H. Swanepoel, vir sy leiding, insig, entoesiasme en voortgesette ondersteuning wat noodsaaklik was vir die voltooiing van hierdie studie.
- Prof. F.C. van Graan, vir waardevolle samesprekings.
- My ouers, vir liefde, opvoeding en bystand.
- My skoonouers, vir volgehoue belangstelling.
- My pragtige vrou, Salomie, en my pasgebore dogtertjie Kayla, vir liefde, geduld en onderskraging.

Vir geleenthede en die voorreg om hierdie taak te voltooi, bring ek dank aan God.

Contents

1	Introduction	1
1.1	Overview	1
1.2	Mathematical notation and some known facts	4
2	Two nonparametric estimation methods	9
2.1	Kernel density estimation	9
2.1.1	An appropriate discrepancy measure	11
2.1.2	Efficiency measure for the kernel density estimator	14
2.1.3	The choice of an appropriate kernel function	17
2.1.4	The choice of a smoothing parameter	20
2.1.5	Boundary bias	31
2.1.6	Spurious bumps in the tails	38
2.2	Kernel distribution function estimation	42
2.2.1	An appropriate discrepancy measure	44
2.2.2	The choice of an appropriate kernel function	48
2.2.3	The choice of a smoothing parameter	51
3	Transformation of data	59
3.1	QQ-plots: key to transformations	59
3.2	A new transformation to any distribution	62
3.2.1	The transformation	62
3.2.2	Polynomial approximation of the optimal transformation function	66
3.3	Parametric transformations	68
3.3.1	Overview	68
3.3.2	Transformation curvature	71
3.3.3	Parameter estimation and transformation selection	80

3.4	A new optimal semi-parametric transformation to normality	90
3.5	Application of the optimal transformation to simulated data	93
4	Transformation kernel density estimation	99
4.1	The transformation kernel density estimator	100
4.2	The new optimal semi-parametric TKDE	114
5	Empirical studies	127
5.1	Simulation study	127
5.1.1	Normal	135
5.1.2	Uniform	137
5.1.3	Bimodal	139
5.1.4	Trimodal	141
5.1.5	Claw	143
5.1.6	Skewed bimodal	145
5.1.7	Skewed unimodal	147
5.1.8	Weibull	149
5.1.9	Lognormal	151
5.1.10	Exponential	153
5.1.11	Strict-Pareto	155
5.1.12	Kurtotic unimodal	157
5.1.13	Separated bimodal	159
5.1.14	Conclusions	161
5.2	Applications to real data	164
5.2.1	Example 1: British income data	164
5.2.2	Example 2: Astrophysical data	167
5.2.3	Example 3: Buffalo snowfall data	169

1

Introduction

1.1 Overview

The *probability density function* is a fundamental concept in statistics. Consider any random variable X that has probability density function f . Specifying the function f gives a natural description of the *distribution* of X , and allows probabilities associated with X to be found from the relation

$$P(a \leq X \leq b) = \int_a^b f(x)dx \text{ for all } a < b.$$

Suppose now that X_1, X_2, \dots, X_n are independent and identically distributed (i.i.d.) continuous random variables having a density f . *Density estimation*, as discussed in this dissertation, is the construction of an estimate of f from the observed data X_1, X_2, \dots, X_n . The *parametric* approach to estimation of f involves assuming that f belongs to a parametric family of distributions, such as the normal or gamma family, and then estimating the unknown parameters using, for example, maximum likelihood estimation. On the other hand, a *nonparametric density estimator* assumes no pre-specified functional form of f . Nonparametric density estimation is an important data analytic tool which provides a very effective way of showing structure in a set of data at the beginning of its analysis.

The oldest and most widely used nonparametric density estimator is the *histogram*. This is usually formed by dividing the real line into equally sized intervals, often called *bins*.

The histogram is then a step function with heights being the proportion of the sample contained in each bin divided by the width of the bin. Two choices have to be made when constructing a histogram: the binwidth and the positioning of the bin edges. Each of these choices can have a significant effect on the resulting histogram. The binwidth is usually called a *smoothing parameter* since it controls the amount of “smoothing” being applied to the data. All nonparametric curve estimates have an associated smoothing parameter. We will see in the following chapters that, for kernel density estimators introduced in Chapter 2, the scale of the kernel plays a role analogous to that of the binwidth. The sensitivity of the histogram to the placement of the bin edges is a problem not shared by other density estimators such as the kernel density estimator. The bin edge problem is one of the histogram’s main disadvantages.

The histogram has several other problems not shared by kernel density estimators. Most densities are not step functions, yet the histogram has the unattractive feature of estimating all densities by a step function. A further problem is the extension of the histogram to the multivariate setting, especially the graphical display of a multivariate histogram. Finally, the histogram can be shown not to use the data as effectively as the kernel estimator. Despite these drawbacks, the simplicity of histograms ensures their continuing popularity.

A large class of nonparametric density estimators has appeared in the statistical literature as alternatives to the histogram, of which the kernel approach (mentioned above) is a popular and conceptually simple one. Kernel estimators have been around since the seminal papers of Rosenblatt (1956) and Parzen (1962). These estimators have the advantage of being very intuitive and relatively easy to analyze mathematically.

It is well known that the *ordinary* kernel density estimator has, despite of several good properties, some drawbacks (a comprehensive discussion of kernel density and distribution function estimation is given in Chapter 2). For example, it suffers from boundary bias and it also exhibits spurious bumps in the tails. Various solutions to overcome these defects are presented in this study, which include the application of a transformation kernel density estimator. The latter estimator (if implemented correctly) is pursued as a simultaneous solution for both boundary bias and spurious bumps in the tails. The

estimator also has, among others, the ability to detect and estimate density modes more effectively.

To apply the transformation kernel density estimator an effective transformation of the data is required. To achieve this objective, an extensive discussion of parametric transformations introduced and studied in the literature is presented in Chapter 3 firstly, emphasizing the practical feasibility of these transformations. Secondly, known methods of estimating the parameters associated with these transformations are discussed (e.g. profile maximum likelihood), and two new estimation techniques, referred to as the minimum residual and minimum distance methods, are introduced. Furthermore, new procedures are developed to select a parametric transformation that is suitable for application to a given set of data. Finally, utilizing the above techniques, the desired optimal transformation to any target distribution (e.g. the normal distribution) is introduced, which has the property that it can also be iterated. A polynomial approximation of the optimal transformation function is presented. *It is shown that the performance of this transformation exceeds that of any transformation available in the literature.*

In the context of transformation kernel density estimation, we present in Chapter 4 a comprehensive literature study of current methods available and then introduce the new semi-parametric transformation estimation procedure based on the optimal transformation of data to normality. However, application of the optimal transformation in this context requires special attention. In order to create a density estimator that addresses both boundary bias and spurious bumps in the tails simultaneously in an automatic way, a generalized bandwidth adaptation procedure is developed, which is applied in conjunction with a newly developed constant shift procedure.

Furthermore, the optimal transformation function is based on a kernel distribution function estimator. A new data-based smoothing parameter (bandwidth selector) is invented in Chapter 2, and it is shown that this selector has better performance than a well established bandwidth selector proposed in the literature.

To evaluate the performance of the newly proposed semi-parametric transformation estimation procedure, a simulation study is presented in Chapter 5 based on densities that

consist of a wide range of forms. Some of the main results derived in the Monte Carlo simulation study include that:

- *the proposed optimal transformation function can take on all the possible shapes of a parametric transformation as well as any combination of these shapes, which result in high p -values when testing normality of the transformed data.*
- *the newly formulated minimum residual and minimum distance techniques contribute to better transformations to normality, when a parametric transformation is applicable.*
- *the newly proposed semi-parametric transformation kernel density estimator performs well for unimodal, low and high kurtosis densities. Moreover, it estimates densities with much curvature (e.g. modes and valleys) more effectively than existing procedures in the literature.*
- *the new transformation density estimator does not exhibit spurious bumps in the tail regions.*
- *boundary bias is addressed automatically.*

In conclusion, practical examples based on real-life data are presented.

1.2 Mathematical notation and some known facts

In this section a summary of the most prominent mathematical notation and some mathematical calculations will be presented. This section serves as a quick reference and promote readability in the rest of this dissertation. The informed reader may proceed to Chapter 2. In this section an unqualified integral sign \int will be taken to mean integration over the entire real line, \mathbf{R} .

1. General notation

- (a) The j^{th} moment: $\mu_j(k) = \int x^j k(x) dx$, for some density function k , with the assumption that $\int |x|^j k(x) dx < \infty$, $\forall j > 0$.
- (b) k is a r^{th} – order kernel if
 - $\mu_0(k) = 1$,

- $\mu_j(k) = 0, \quad j = 1, \dots, r-1,$
- $\mu_r(k) \neq 0.$

(c) The convolution of f and $g : (f * g)(x) = \int f(x-y)g(y)dy.$

(d) Real-valued O and o notation: Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers then

- $a_n = O(b_n)$, if and only if $\limsup_{n \rightarrow \infty} \left| \frac{a_n}{b_n} \right| < \infty$ and consequently $a_n = O(1)$ is equivalent to a_n being bounded. We will say " a_n is of order b_n " if $a_n = O(b_n)$.
- $a_n = o(b_n)$, if and only if $\lim_{n \rightarrow \infty} \left| \frac{a_n}{b_n} \right| = 0$ and consequently $a_n = o(1)$ is equivalent to $a_n \rightarrow 0$ as $n \rightarrow \infty$.

(e) Asymptotic notation: a_n is asymptotically equivalent to b_n thus $a_n \sim b_n$ if and only if $\lim_{n \rightarrow \infty} \left(\frac{a_n}{b_n} \right) = 1.$

(f) Derivatives:

- $k^{(m)}(x) = \frac{d^m}{dx^m} k(x).$
- If $k(x)$ is a symmetric function then $k^{(m)}(x)$ is also a symmetric function for m being even, hence

$$k^{(m)}(x) = (-1)^m k^{(m)}(-x) = \begin{cases} k^{(m)}(-x) & \text{if } m \text{ is even.} \\ -k^{(m)}(-x) & \text{if } m \text{ is odd.} \end{cases}$$

(g) The kernel estimate of $f^{(m)}(x)$ is given by

$$\hat{f}^{(m)}(x; h) = \frac{1}{nh^{m+1}} \sum_{i=1}^n k^{(m)} \left(\frac{x - X_i}{h} \right).$$

(h) Taylor's theorem: Assume that f has m continuous derivatives in an interval $(x - \delta, x + \delta)$ for some $\delta > 0$. Then for any sequence α_n converging to zero

$$f(x + \alpha_n) = \sum_{i=0}^m \left(\frac{\alpha_n^i}{i!} \right) f^{(i)}(x) + o(\alpha_n^m).$$

(i) Define: $R(k) = \int k(x)^2 dx.$

(j) For $k(\cdot)$ and $K(\cdot)$ the symmetric around zero kernel density and distribution functions respectively, we have:

$$\text{i. } \int_{-\infty}^{+\infty} k(x)K(x)dx = 1/2.$$

$$\text{ii. } \int_0^{+\infty} k(x)K(x)dx = 3/8.$$

$$\text{iii. } \int_0^{+\infty} [K(z)]^2 k(z)dz = 7/24.$$

$$\text{iv. } \int_0^{+\infty} k(z) [2K(z) - 1]^2 dz = 1/6.$$

$$\text{v. } \int_0^{+\infty} \{zk(z)^{1/2}\} \{k(z)^{1/2} [2K(z) - 1]\} dz \leq \frac{\mu_2(k)^{1/2}}{2\sqrt{3}}.$$

(k) Let $F(x)$ be a distribution function with associated density function $f(x)$ and let $g(x)$ be any real valued function assuming values between 0 and 1. If $y = F^{-1}[g(x)]$ then

$$\frac{dy}{dx} = \frac{dg(x)/dx}{f[F^{-1}[g(x)]]}.$$

2. Properties of the normal distribution

(a) The standard normal probability density function: $\phi(x) = 1/\sqrt{2\pi} e^{-x^2/2}$.

(b) The standard normal probability distribution function: $\Phi(x) = \int_{-\infty}^x \phi(t)dt$.

(c) Rescaling: The $N(\mu, \sigma^2)$ normal density is defined as

$$\phi_\sigma(x - \mu) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2}.$$

(d) Define the odd factorial for $m = 0, 1, \dots$ as

$$OF(m) = \frac{m!}{2^{\frac{m}{2}} \left(\frac{m}{2}\right)!}.$$

(e) Define the m^{th} Hermite polynomial as

$$H_m(x) = xH_{m-1}(x) - (m-1)H_{m-2}(x).$$

Table 1.1: The first 10 Hermite polynomials

$H_0(x) = 1$
$H_1(x) = x$
$H_2(x) = x^2 - 1$
$H_3(x) = x^3 - 3x$
$H_4(x) = x^4 - 6x^2 + 3$
$H_5(x) = x^5 - 10x^3 + 15x$
$H_6(x) = x^6 - 15x^4 + 45x^2 - 15$
$H_7(x) = x^7 - 21x^5 + 105x^3 - 105x$
$H_8(x) = x^8 - 28x^6 + 210x^4 - 420x^2 + 105$
$H_9(x) = x^9 - 36x^7 + 378x^5 - 1260x^3 + 945x$
$H_{10}(x) = x^{10} - 45x^8 + 630x^6 - 3150x^4 + 4725x^2 - 945$

(f) The Hermite polynomial and odd factorial will be used to calculate the derivatives of the normal distribution, using

i. $\phi^{(m)}(x) = (-1)^m H_m(x) \phi(x)$.

ii. $\phi_\sigma^{(m)}(0) = \begin{cases} (-1)^{m/2} \frac{1}{\sqrt{2\pi}} OF(m) \sigma^{-m-1} & \text{if } m \text{ is even.} \\ 0 & \text{if } m \text{ is odd.} \end{cases}$

(g) $\int \phi_\sigma^{(m)}(x - \mu) \phi_{\sigma'}^{(m')}(x - \mu') dx = (-1)^m \phi_{\sqrt{\sigma^2 + \sigma'^2}}^{(m+m')}(\mu - \mu')$.

(h) For $\sigma > 0, m = 0, 1, 2, \dots$ and $X \sim N(\mu, \sigma^2)$,

$$E(X^m) = \sum_{j=0}^{\lfloor m/2 \rfloor} OF(2j) \binom{m}{2j} \mu^{m-2j} \sigma^{2j},$$

where $\lfloor x \rfloor =$ greatest integer less than or equal to x .

(i) For $X \sim N(0, \sigma^2)$,

$$E(X^m) = \begin{cases} \sigma^m OF(m) & \text{if } m \text{ is even.} \\ 0 & \text{if } m \text{ is odd.} \end{cases}$$

(j) For $\sigma_1, \sigma_2 > 0$

$$\phi_{\sigma_1}(x - \mu_1) \phi_{\sigma_2}(x - \mu_2) = \phi_{\sqrt{\sigma_1^2 + \sigma_2^2}}(\mu_1 - \mu_2) \phi_{\sigma_1 \sigma_2 / \sqrt{\sigma_1^2 + \sigma_2^2}}(x - \mu^*),$$

$$\text{where } \mu^* = \frac{\sigma_2^2 \mu_1 + \sigma_1^2 \mu_2}{\sigma_1^2 + \sigma_2^2}.$$

(k)

$$\phi(x)^m = \frac{1}{m^{1/2}} (2\pi)^{(1-m)/2} \phi_{m^{-1/2}}(x).$$

(l) Using the properties above it is a simple matter to verify that

- i. $\mu_2(\phi) = 1.$
- ii. $\phi^{(2)}(0) = \frac{-1}{\sqrt{2\pi}}.$
- iii. $\phi^{(4)}(0) = \frac{3}{\sqrt{2\pi}}.$
- iv. $\phi^{(6)}(0) = \frac{-15}{\sqrt{2\pi}}.$
- v. $R(\phi_\sigma(x - \mu)) = \frac{1}{2\sqrt{\pi}\sigma}.$
- vi. $R(\phi'_\sigma(x - \mu)) = \frac{1}{4\sqrt{\pi}\sigma^3}.$
- vii. $R(\phi''_\sigma(x - \mu)) = \frac{3}{8\sqrt{\pi}\sigma^5}.$
- viii. $\int \phi_\sigma^{(1)}(x - \mu) \phi_\sigma^{(1)}(x - \mu) \phi_\sigma(x - \mu) dx = \frac{1}{3\sqrt{3}} (2\pi)^{-1} \sigma^{-4}.$
- ix. $\int \phi_\sigma^{(1)}(x - \mu) \phi_\sigma^{(3)}(x - \mu) \phi_\sigma(x - \mu) dx = \frac{-1}{3\sqrt{3}\pi} \sigma^{-6}.$
- x. $\iiint \phi'(s) \phi'(t) \phi'(u) \phi'(t + u - s) ds dt du = \frac{3}{2^5} (2\pi)^{-1/2}.$
- xi. $\int [\phi_\sigma(x - \mu)]^4 dx = \frac{1}{2} (2\pi)^{-3/2} \sigma^{-3}.$
- xii. $\int x \phi(x) \Phi(x) dx = \frac{1}{2\sqrt{\pi}}.$

2

Two nonparametric estimation methods

In this section the kernel density estimator and kernel distribution function estimator are discussed in detail. In the context of kernel density estimation we will discuss an appropriate discrepancy measure, difficulty of estimation, the choice of an appropriate kernel function, the choice of the smoothing parameter, boundary bias and spurious bumps in the tails. In the context of kernel distribution estimation we will discuss an appropriate discrepancy measure, the choice of an appropriate kernel function and the choice of the smoothing parameter. For the choice of the smoothing parameter, a slight alteration to an existing plug-in selector will be introduced.

2.1 Kernel density estimation

Let X_1, \dots, X_n be i.i.d. continuous random variables from the probability law F_X , having a continuous univariate density f_X . Using the compact notation $k_h(u) = \frac{1}{h}k\left(\frac{u}{h}\right)$, the kernel density estimator is then given by

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n k_h(x - X_i), \quad (2.1)$$

where k is the so-called kernel (or weight) function and h is the smoothing parameter or bandwidth. In this and subsequent chapters the kernel estimator will be referred to as $\hat{f}(x; h)$, $\hat{f}_h(x)$, $f_{n,h}(x)$ or $f_n(x)$. We assume the kernel function has the following properties

- $k(u) \geq 0, \quad \forall u \in \mathbf{R}$.
- $\int k(u)du = \mu_0(k) = 1$, hence k is a density function.
- $k(-u) = k(u)$, hence k is a symmetric function. This implies that
- $\int uk(u)du = \mu_1(k) = 0$.
- $\int u^2k(u)du = \mu_2(k) = a^2 < +\infty$.

Requiring that k must be a density function, ensures that the kernel estimate is also a density function. Using the standard normal density function as kernel, one can think of the kernel density estimator (2.1) at a specific point, say x , as the average of n normal density functions with means $X_i, i = 1, \dots, n$, and standard deviation h . This is explained graphically in Figure 2.1, where a sample of 10 data points from the standard normal distribution is used for illustration. From Figure 2.1 it should be clear that data points

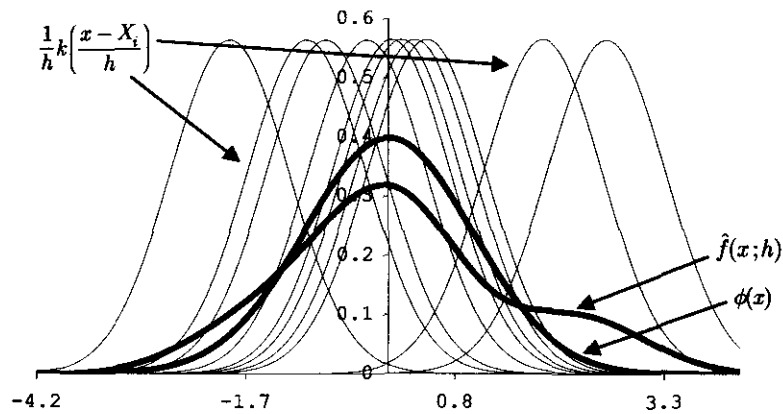


Figure 2.1: *Kernel density estimation*

in the region of x contribute more to the estimation of the density in that point. The visualization given in Figure 2.1 is useful when explaining concepts such as boundary bias and spurious bumps in the tails. These concepts will be explained in greater detail in Section 2.1.5 and Section 2.1.6.

2.1.1 An appropriate discrepancy measure

In order to assess the performance of the kernel estimator given in (2.1), one needs to define a discrepancy measure between the estimator and the target density. In existing literature, the most popular discrepancy measures are the mean squared error (MSE), the mean integrated squared error (MISE) and the asymptotic mean integrated squared error (AMISE). Wand and Jones (1995) pointed out that there are good reasons for working with other discrepancy measures such as the mean integrated absolute error defined as $MIAE\{\hat{f}(\cdot; h)\} = E \int |\hat{f}(x; h) - f(x)| dx$. The interested reader is referred to Devroye and Györfi (1985), Loots (1995) p.43, and Jones, Marron and Sheather (1996) for further discussion of other discrepancy measures as well as references to other papers. Henceforth, an unqualified integral sign \int will be taken to mean integration over the entire real line, \mathbf{R} . For verification and mathematical derivation of the results presented in this section, the reader is referred to Wand and Jones (1995) and Koekemoer (1999).

The mean squared error

The mean squared error of the kernel estimator $\hat{f}(x; h)$ at some point $x \in \mathbf{R}$ is given by

$$MSE [\hat{f}(x; h)] = E [\hat{f}(x; h) - f(x)]^2.$$

This expression can be written in an alternative, easier to interpret way namely

$$MSE [\hat{f}(x; h)] = Var [\hat{f}(x; h)] + \{Bias [\hat{f}(x; h)]\}^2. \quad (2.2)$$

Using notation from Section 1.2 we can write the bias term in (2.2) as

$$\begin{aligned} Bias [\hat{f}(x; h)] &= E\hat{f}(x; h) - f(x) \\ &= \int k_h(x - y)f(y)dy - f(x) \\ &= (k_h * f)(x) - f(x). \end{aligned} \quad (2.3)$$

Using the same notation we can write the variance term as

$$\begin{aligned} Var [\hat{f}(x; h)] &= E\hat{f}(x; h)^2 - [E\hat{f}(x; h)]^2 \\ &= \frac{1}{n} (k_h^2 * f)(x) - \frac{1}{n} (k_h * f)^2(x). \end{aligned} \quad (2.4)$$

Substitution of (2.3) and (2.4) into (2.2) lead to an expression for the discrepancy measure MSE at a single point x . This is given by

$$MSE [\hat{f}(x; h)] = \frac{1}{n} (k_h^2 * f)(x) - \frac{1}{n} (k_h * f)^2(x) + \{(k_h * f)(x) - f(x)\}^2. \quad (2.5)$$

The mean integrated squared error

The mean squared error can be used as a discrepancy measure at a point x . This measure is, therefore, a local measure of discrepancy. Evaluating (2.5) at each x point and then integrating with respect to x gives rise to the mean integrated squared error, which is consequently a global measure of discrepancy. A successful kernel density estimator in all points $x \in \mathbf{R}$ will result in a small MISE. The MISE is defined as $MISE[\hat{f}(\cdot; h)] = \int MSE[\hat{f}(x; h)] dx$. Using (2.5) we can write the MISE in a more manageable form:

$$\begin{aligned} & MISE[\hat{f}(\cdot; h)] \\ &= \frac{1}{n} \int (k_h^2 * f)(x) dx + \left(1 - \frac{1}{n}\right) \int (k_h * f)^2(x) dx - 2 \int (k_h * f)(x) f(x) dx + \int f(x)^2 dx \end{aligned} \quad (2.6)$$

where

$$\begin{aligned} \frac{1}{n} \int (k_h^2 * f)(x) dx &= \frac{1}{n} \iint k_h^2(x-y) f(y) dy dx \\ &= \frac{1}{nh} \int k^2(z) dz. \end{aligned} \quad (2.7)$$

Substituting (2.7) into (2.6) lead to the following MISE expression

$$\begin{aligned} & MISE[\hat{f}(\cdot; h)] \\ &= \frac{1}{nh} \int k^2(x) dx + \left(1 - \frac{1}{n}\right) \int (k_h * f)^2(x) dx - 2 \int (k_h * f)(x) f(x) dx + \int f(x)^2 dx \end{aligned} \quad (2.8)$$

The MISE given in (2.8) can be used to find the optimal smoothing parameter, for which this discrepancy measure will be small. The MISE expression depends, however, on h in a complicated manner. For this reason, the asymptotic mean integrated squared error is developed. This expression depends on h in a simple manner and gives rise to the asymptotic optimal bandwidth.

The asymptotic mean integrated squared error

In this section we will derive large sample approximations for the leading variance and bias terms in (2.8), and then study the dependence on h of the resulting expression. In order to derive these approximations we need to make some assumptions. These are

1. The density f has a continuous, square integrable and ultimately monotone second derivative f'' . An ultimately monotone function is one that is monotone over both $(-\infty, -M)$ and $(M, +\infty)$ for some $M > 0$.
2. The bandwidth h is a non-random sequence of positive numbers. Also assume that h satisfies

$$\lim_{n \rightarrow \infty} h = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} nh = \infty.$$

This is equivalent to saying that h approaches zero slower than n goes to infinity.

3. The kernel function k is a bounded probability density function with finite fourth moment, and is symmetric about the origin.

Assumption (2) is made mainly to ensure that the asymptotic variance term converges to zero, see expression (2.13) below for more detail. Understanding this assumption is important since it places a restriction on the order of h . For example, one can take

$$h = cn^{-t} \quad \text{where} \quad 0 < t < 1, \quad (2.9)$$

and c is a finite positive constant. It is worthwhile to note that larger values of t imply faster convergence rates of h to zero as $n \rightarrow \infty$, thus smaller bandwidths.

We will now proceed with first finding the asymptotic mean squared error (AMSE), and then the asymptotic mean integrated squared error (MISE). The bias and variance terms are treated separately. From (2.3) it follows that, using the notation from Section 1.2, (1a), the bias term is given by:

$$\text{Bias} [\hat{f}(x; h)] = \frac{1}{2}h^2\mu_2(k)f''(x) + o(h^2). \quad (2.10)$$

Note that the leading term in (2.10) is $O(h^2)$ and therefore, using assumption (2), it follows that $\hat{f}(x; h)$ is asymptotically an unbiased estimator for the target density f . Next, we will find an asymptotic expression for the variance term. From (2.4) we find, using the notation from Section 1.2, (1i), that

$$\text{Var} [\hat{f}(x; h)] = \frac{1}{nh}R(k)f(x) + o\left(\frac{1}{nh}\right). \quad (2.11)$$

Note that the leading term in (2.11) is $O(\frac{1}{nh})$ and therefore, using assumption (2), it follows that $\text{Var} [\hat{f}(x; h)]$ converges to zero. Using (2.2), (2.10) and (2.11) we define the AMSE to be

$$\text{AMSE} [\hat{f}(x; h)] = \frac{1}{nh}R(k)f(x) + \frac{1}{4}h^4\mu_2(k)^2f''(x)^2. \quad (2.12)$$

We will now proceed with the calculation of the AMISE. Using (2.12) we find that

$$\begin{aligned} AMISE[\hat{f}(\cdot; h)] &= \int AMSE[\hat{f}(x; h)] dx \\ &= \frac{1}{nh}R(k) + \frac{1}{4}h^4\mu_2(k)^2R(f''). \end{aligned} \quad (2.13)$$

From (2.13) it is important to note that the asymptotic integrated squared bias is proportional to h^4 , and hence we need to choose h as small as possible. Contrary to this, the asymptotic variance is proportional to $\frac{1}{nh}$, hence small values of h will increase the variance term. This is known as the variance-bias trade-off. The consequence of this phenomenon is that for small h , we will get a density estimate that is spiky (under smoothed), and for large h , we will get a density estimate that is smooth, with larger bias (over smoothed). It is clear that we must find a balance between the $O(h^4)$ squared bias term and the $O(\frac{1}{nh})$ variance term. It is easy to show that this balance is given by the following choice of h

$$h_{AMISE} = \left[\frac{R(k)}{\mu_2(k)^2 R(f'')} \right]^{\frac{1}{5}} n^{-1/5}. \quad (2.14)$$

To implement (2.14) in practice an estimate of $R(f'')$ is needed, this is discussed in Section 2.1.4. By substituting (2.14) into (2.13) we find that

$$\inf_{h>0} AMISE[\hat{f}(\cdot; h)] = \frac{5}{4}C(k)R(f'')^{1/5}n^{-4/5}, \quad (2.15)$$

where $C(k) = \mu_2(k)^{2/5}R(k)^{4/5}$ is a constant only depending on the kernel function k . Expression (2.15) is the smallest possible AMISE that can be attained using h_{AMISE} and the kernel function k .

2.1.2 Efficiency measure for the kernel density estimator

In this section we will derive a formula that measures how well a particular density can be estimated using the kernel density estimator. This section is extremely important in the context of the transformation kernel density estimator (which will be defined in Chapter 4), since the result obtained here is instrumental in finding an optimal distribution for the transformed data. Using the asymptotic optimal bandwidth (2.14), the global discrepancy measure AMISE given in (2.15) should be smaller for a density that is easy to estimate when compared to a target density that is difficult to estimate. On closer inspection of (2.15) it is clear that this expression only depends on the unknown

target density f via the functional $R(f'')$. We can, therefore, conclude that the functional $R(f'') = \int f''(x)^2 dx$ gives us an indication of how well f can be estimated even when h is chosen optimally. For target densities, f , with “sharp” features such as high skewness or several modes $|f''(x)|$ will take on relatively large values resulting in a large value of $R(f'')$. For densities without these features $R(f'')$ should be smaller, hence easier to estimate.

Ultimately, one would like to compare the estimation difficulty of different target densities. This, however, cannot be accomplished using $R(f'')$ since $R(f'')$ is not scale invariant, thus distributions with a larger scale measure, $\sigma_x > 0$, will result in larger values of $R(f'')$. Consider the random variable X with density f_X and set $Y = X/\sigma_x$, where σ_x is the population standard deviation of X . The random variable Y is scale invariant, hence using the density of Y , we can construct a scale invariant difficulty measure. Noting that the density of Y is given by $f_Y(y) = \sigma_x f_X(\sigma_x y)$ it is easily verified that

$$D(f_X) = R(f_Y'') = \sigma_x^5 R(f_X''), \quad (2.16)$$

is the scale invariant difficulty measure, henceforth referred to as $D(f)$. Small values of $D(f)$ entail that f is easy to estimate. Comparing the difficulty measure for several target densities requires a reference point. The beta(α, β) density function is defined as

$$f(x) = \frac{\Gamma(\alpha + \beta)}{b\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x-a}{b}\right)^{\alpha-1} \left(\frac{a+b-x}{b}\right)^{\beta-1}, \quad a \leq x \leq a+b, \quad (2.17)$$

where $\Gamma(\cdot)$ is the gamma-function. Choosing $a = -1$ and $b = 2$, Terrell (1990) showed that $R(f^{(r)})$ is minimized by the beta($r+2, r+2$) density function. Hence, $R(f'')$ is minimal for the beta(4,4) density defined as

$$f^*(x) = \frac{35}{32}(1-x^2)^3, \quad -1 \leq x \leq 1.$$

Note that any shift or rescaling of f^* will also minimize $D(f)$. One can therefore conclude that the beta(4,4) density is the easiest to estimate using kernel density estimation, and can be used as a reference point. The beta(4,4) density is shown in Figure 2.2. Using the beta(4,4) density as reference point, the efficiency measure of the kernel estimator is defined as

$$\text{Eff}(f) = \frac{D(f^*)}{D(f)}.$$

Table 2.1 summarizes the efficiency measure for several densities. For the definition and graphical inspection of these densities the reader is referred to Section 5.1. From Table 2.1

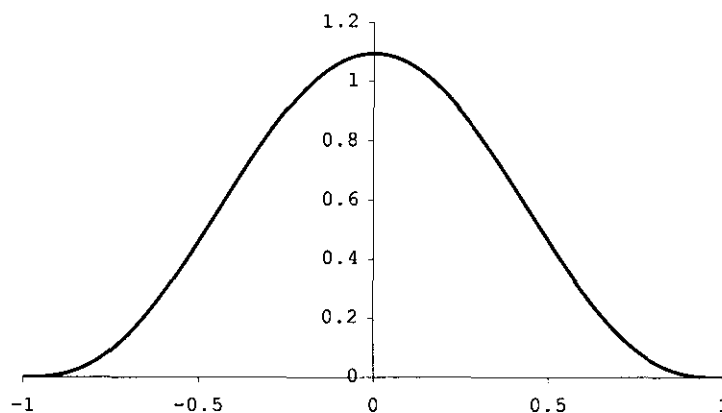
Figure 2.2: The $\text{beta}(4,4)$ density

Table 2.1: Efficiencies of the kernel estimator for several densities

Density	$\text{Eff}(f) = D(f^*)/D(f)$
Beta(4,4)	1
Normal	0.908
Extreme value	0.688
Skewed bimodal	0.568
Bimodal	0.536
Gamma(3)	0.327
Kurtotic unimodal	0.114
Lognormal	0.053

it is clear that although the $\text{beta}(4,4)$ density is the easiest to estimate, the normal density is almost as easy. This is useful information in the context of transformation kernel density estimation, since this enables us to transform data to normality and then estimate the density of the transformed data with a high efficiency. Hence, this serves as a motivation for a transformation to normality when applying the transformation kernel density estimator. This topic is explored in greater detail in Chapter 4. Chapter 3 is devoted to transforming data to normality.

2.1.3 The choice of an appropriate kernel function

In this section the choice of an appropriate kernel function is explored, after which a few possible kernel functions will be defined for utilization. The kernel function, k , is a r^{th} -order kernel if (using notation from Section 1.2, (1a))

- $\mu_0(k) = 1$,
- $\mu_j(k) = 0, \quad j = 1, \dots, r-1$,
- $\mu_r(k) \neq 0$.

It should be noted that for higher-order kernels, ($r > 2$), the restriction that k must be a density function is relaxed and consequently better rates of convergence of AMISE to zero can be obtained. This, however, is not advised since the density restriction on k ensures that the kernel estimate will be a density. For this reason only second order symmetric kernels will be considered in this dissertation. The interested reader is referred to Wand and Schucany (1990), Müller (1991), Jones and Foster (1993) and Wand and Jones (1995) for a discussion of these higher-order kernels.

Following the same logic from Section 2.1.2 we will find the optimal kernel function in the AMISE sense. Recall that from (2.15), using an optimal bandwidth (see (2.14)), the resulting AMISE is given by

$$\inf_{h>0} AMISE [\hat{f}(\cdot; h)] = \frac{5}{4} C(k) R(f'')^{1/5} n^{-4/5},$$

where $C(k) = \mu_2(k)^{2/5} R(k)^{4/5} = \{\mu_2(k)^{1/2} R(k)\}^{4/5}$ is a constant only depending on the kernel function k . Since this equation only depends on the kernel function via the constant $C(k)$, it should be clear that an optimal kernel will minimize this constant. $C(k)$ is however not scale invariant. Hodges and Lehmann (1956) showed that the quantity $C(k)$ is minimized for the kernel function

$$k^a(x) = \begin{cases} \frac{3}{4\sqrt{5}a} \left[1 - \frac{x^2}{5a^2} \right], & -\sqrt{5}a \leq x \leq +\sqrt{5}a, \\ 0, & \text{otherwise,} \end{cases}$$

where a is an arbitrary scale parameter. The simplest version of k^a corresponds to $a^2 = 1/5$, and is often called the Epanechnikov kernel. This kernel is given by

$$k^*(x) = \begin{cases} \frac{3}{4}(1-x^2), & -1 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases} \quad (2.18)$$

The Epanechnikov kernel is shown in Figure 2.3. Using this kernel as reference point,

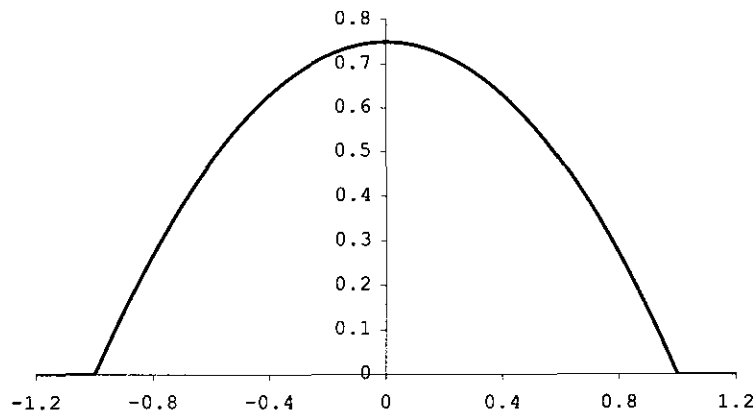


Figure 2.3: *The Epanechnikov kernel*

consider the kernel efficiency measure

$$\text{Eff}(k) = \left\{ \frac{C(k^*)}{C(k)} \right\}^{\frac{5}{4}}. \quad (2.19)$$

The kernel efficiency measure (2.19) can be used to compare the performance of other kernels to the optimal Epanechnikov kernel.

Next, two popular choices of the kernel functions will be discussed and subsequently compared to the Epanechnikov kernel using (2.19). These two choices are summarized in the following list

- The standard normal density. This is a kernel with unbounded support and is defined as

$$k(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < +\infty.$$

- The compactly supported “polynomial kernel”.

$$k(x) = \begin{cases} k_{rs} (1 - |x|^r)^s, & -1 \leq x \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (2.20)$$

where $k_{rs} = \frac{r}{2B(s+1, \frac{1}{r})}$, $r > 0, s \geq 0$ and $B(s, r)$ is the beta-function.

The compactly supported “polynomial kernel” gives rise to five popular kernels, for certain parameter choices, namely

- Rectangular or uniform kernel: $s = 0$.
- Epanechnikov kernel: $r = 2, s = 1$.
- Biweight kernel: $r = 2, s = 2$.
- Triweight kernel: $r = 2, s = 3$.
- Triangular kernel: $r = 1, s = 1$.

Note that by setting $a = -1$ and $b = 2$ in the definition of the beta(α, β) density given in (2.17) it also follows that the rectangular kernel is the beta(1,1), the Epanechnikov kernel is the beta(2,2), the biweight kernel is the beta(3,3) and the triweight kernel is the beta(4,4). Using (2.19) and (2.20) the formula for these kernel functions and their efficiencies are displayed in Table 2.2. The message from Table 2.2 is that AMISE is

Table 2.2: Kernel functions and their efficiency

Kernel Function	Definition	$\mu_2(k)$	Eff(k)
Epanechnikov	$\frac{3}{4}(1 - x^2)$	$\frac{1}{5}$	1
Biweight	$\frac{15}{16}(1 - x^2)^2$	$\frac{1}{7}$	0.994
Triweight	$\frac{35}{32}(1 - x^2)^3$	$\frac{1}{9}$	0.987
Triangular	$1 - x $	$\frac{1}{6}$	0.986
Standard Normal	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$	1	0.951
Rectangular	$\frac{1}{2}$	$\frac{1}{3}$	0.930

insensitive to the choice of the kernel function k . It should be noted that uniform kernels are not very popular in practice since the corresponding density estimate is piecewise constant, and even the Epanechnikov kernel gives an estimate having a discontinuous first derivative which can be unattractive because of its “kinks”. We conclude, therefore,

that k should be chosen based on other issues, such as ease of computation. For this reason the standard normal kernel is used in this dissertation.

2.1.4 The choice of a smoothing parameter

There exists an extensive literature on the selection of the optimal data-based smoothing parameter. In this section we will present a short summary of the existing methods after which the normal scaled rule of thumb and the high-tech plug-in procedure of Sheather and Jones (1991) will be discussed in some detail. The normal scaled rule of thumb plays an important role in understanding the high-tech procedure of Sheather and Jones (1991) and can be considered as a special case of this procedure. The authors, Sheather and Jones (1991), consider their selection procedure to be second to none in the existing literature. It should therefore be no surprise that we based all bandwidth selection required in this dissertation on this widely regarded procedure. Nevertheless, we will now proceed with a short literature study of the most prominent procedures.

Rudemo (1982) and Bowman (1984) proposed the least-squares cross-validation procedure which is based on the MISE expansion of the form

$$\text{MISE} [\hat{f}(\cdot; h)] - \int f(x)^2 dx = E \left[\int \hat{f}(x; h)^2 dx - 2 \int \hat{f}(x; h) f(x) dx \right].$$

The authors propose to minimize

$$\text{LSCV}(h) = \int \hat{f}(x; h)^2 dx - \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n k_h(X_i - X_j),$$

with respect to h . For the least-squares cross-validation procedure, the discrepancy measure used is the exact MISE. Scott and Terrell (1987) proposed using the asymptotic counterpart, i.e., AMISE presented in (2.13). The resulting selector is called the biased cross-validation method and minimizes

$$\text{BCV}(h) = \frac{1}{nh} R(k) + \frac{1}{4} h^4 \mu_2(k)^2 R(\widehat{f}''),$$

with respect to h , where

$$R(\widehat{f}'') = \frac{1}{n^2 h^6} \int \left[\sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n k'' \left(\frac{x - X_i}{h} \right) k'' \left(\frac{x - X_j}{h} \right) \right] dx.$$

Müller (1985), Staniswalis (1989) and Hall, Marron and Park (1992) proposed the smoothed cross-validation bandwidth selection procedure that is based on an approximate MISE discrepancy measure given by (see (2.8))

$$\text{MISE} [\hat{f}(\cdot; h)] \approx \frac{1}{nh} R(k) + \int (k_h * f - f)^2(x) dx.$$

The proposed procedure minimizes

$$\text{SCV}(h) = \frac{1}{nh} R(k) + \widehat{ISB}(h),$$

where

$$\widehat{ISB}(h) = \int (k_h * \hat{f}_l(\cdot; g) - \hat{f}_l(\cdot; g))^2(x) dx,$$

and

$$\hat{f}_l(\cdot; g) = \frac{1}{n} \sum_{i=1}^n l_g(x - X_i),$$

is a pilot kernel density estimator with a possibly different kernel l and bandwidth g . Chiu (1991a), Chiu (1991b) and Chiu (1992) rewrote the MISE expression [see (2.6)] in terms of the characteristic function, which he then minimizes utilizing cross-validation. Chiu also considered the MISE discrepancy measure obtained for the density estimator:

$$\hat{f}(x; \Lambda) = \frac{1}{2\pi} \int_{-\Lambda}^{+\Lambda} \tilde{\varphi}(t) e^{-itx} dt,$$

where $\tilde{\varphi}(t)$ is the sample characteristic function, to determine the cut-off frequency Λ required in his cross-validation procedure. Lastly, Chiu considered the AMISE optimal bandwidth presented in (2.14) and found an estimator for $R(f'')$ based on the characteristic function. For a more extensive discussion concerning the methods described above, the reader is referred to Wand and Jones (1995) and Koekemoer (1999). Simulation and comparative studies can be found in Park and Marron (1990), Park and Turlach (1992), Cao, Cuevas and González-Manteiga (1994), Loader (1995), Jones et al. (1996), Chiu (1996) and Koekemoer (1999). We will now proceed with a detailed discussion concerning the normal scaled rule of thumb and the high-tech procedure proposed by Sheather and Jones (1991).

Normal scaled rule of thumb

In the rest of this section we will assume that the kernel function is the standard normal

density, thus $k(\cdot) = \phi(\cdot)$. Recall that from the AMISE point of view we may write the asymptotic optimal bandwidth (2.14) as

$$h_{AMISE} = \left[\frac{R(k)}{\mu_2(k)^2 R(f'')} \right]^{\frac{1}{5}} n^{-1/5}.$$

From this expression it is clear that the only unknown value is $R(f'')$. A novel idea is to assume that the unknown density f is a normal density with mean μ and variance σ^2 . This can then be used to calculate $R(f'')$ and consequently the asymptotic optimal bandwidth. Using the properties of the normal distribution as discussed in Section 1.2, in specific (2d), (2f) and (2g), we find that

$$\begin{aligned} R(\phi_\sigma^{(2)}) &= \int \phi_\sigma^{(2)}(x - \mu) \phi_\sigma^{(2)}(x - \mu) dx = \phi_{\sqrt{2}\sigma}^{(4)}(0) = \frac{3}{8\sqrt{\pi}\sigma^5}, \\ R(k) &= R(\phi) = \int \phi(x) \phi(x) dx = \phi_{\sqrt{2}}(0) = \frac{1}{2\sqrt{\pi}}, \\ \mu_2(k) &= \mu_2(\phi) = \left[\int x^2 \phi(x) dx \right] = 1. \end{aligned}$$

Replacing the quantities calculated above into expression (2.14) the normal scaled rule of thumb is found to be

$$h_{NS} = \left[\frac{4}{3} \right]^{1/5} \sigma n^{-1/5} \approx 1.0592 \sigma n^{-1/5}. \quad (2.21)$$

To implement (2.21) it is necessary to estimate the scale parameter σ , which can be affected by outlier data points. Consequently, a larger bandwidth will be obtained, meaning that the density estimate will tend to oversmooth. Silverman (1986) p.47 suggested the use of the robust scale estimator

$$\hat{\sigma} = \min \left[s, \frac{\hat{q}_3 - \hat{q}_1}{\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)} \right], \quad (2.22)$$

where \hat{q}_1 and \hat{q}_3 are the first and third sample quartiles respectively, s is the usual sample standard deviation and $\Phi(\cdot)$ is the standard normal distribution function. Throughout the discussions below this scale estimator will be used when determining the bandwidth for any data, i.e., the original input data and any subsequent transformed data. For a discussion on more sophisticated scale estimates the reader is referred to Janssen, Marron, Veraverbeke and Sarle (1995). It is also important to note that in the context of data transformation, standardization is required, and the scale estimate is determined in a similar fashion as above.

Estimation of density functionals

In order to calculate the asymptotic optimal bandwidth given in (2.14), one needs to find an estimate of the unknown quantity $R(f'')$. Once this estimate is obtained, one can plug the estimate into expression (2.14) to find the asymptotic optimal bandwidth. This procedure is in essence the highly regarded Sheather and Jones (1991) plug-in method. It is therefore essential to find a good estimate for the unknown quantity. The quantity $R(f'')$ fulfil an important role in the context of density estimation, since this quantity is used to

- measure the difficulty of estimating f (see Section 2.1.2),
- calculate the well respected Sheather and Jones (1991) plug-in bandwidth,
- find the appropriate transformation parameters. (see Section 3.3.3 and Section 4.1).

It is therefore imperative that the reader should understand the estimation procedure of $R(f'')$. It should also be noted that $R(f''')$, $R(f^{iv})$, \dots will be required in the method of Sheather and Jones (1991). In addition, $R(f')$ plays an important role in the context of kernel distribution function estimation (see Section 2.2.1 and Section 2.2.3 for more detail). Hence, an attempt is made to find an estimate for the general functional $R(f^{(s)})$, $s = 0, 1, 2, 3, \dots$. The bandwidth used to estimate this quantity is denoted by g and the kernel function by w . For all practical purposes we will set $w(\cdot) = k(\cdot)$, where $k(\cdot)$ is the kernel function used for estimating the density f , when the estimate of $R(f^{(s)})$ is employed.

With the assumption of sufficient smoothness on f , we may write

$$R(f^{(s)}) = (-1)^s \int f^{(2s)}(x)f(x)dx = (-1)^s \psi_{2s} = (-1)^{m/2} \psi_m$$

(with $m = 2s$ and $s = 0, 1, 2, \dots$).

It is therefore appropriate to consider estimation of functionals of the form

$$R(f^{(m/2)}) = \psi_m = \int f^{(m)}(x)f(x)dx = E [f^{(m)}(X)], \quad (2.23)$$

where m will be an even integer. Hall and Marron (1987) and Sheather and Jones (1991) proposed the following estimator

$$\hat{\psi}_m(g) = \frac{1}{n} \sum_{i=1}^n \hat{f}^{(m)}(X_i; g) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n w_g^{(m)}(X_i - X_j). \quad (2.24)$$

Hall and Marron (1987) argued that the terms for which $i = j$ do not involve the data and can be thought of as bias terms, and so they proposed an estimator which explicitly excludes those terms. However, Sheather and Jones (1991) showed that the excluded terms can actually be used to improve the estimator by cancelling other bias terms. In order to find the bandwidth, g , an expression for the asymptotic mean squared error of $\hat{\psi}_m(g)$ is required.

Before proceeding with the derivation, consider the following assumptions

1. The kernel w is a symmetric kernel of order r , $r = 2, 4, \dots$, possessing m derivatives, such that

$$(-1)^{(m+r)/2+1} w^{(m)}(0) \mu_r(w) > 0.$$

2. The density f has p continuous derivatives that are each ultimately monotone, where $p > r$.
3. The bandwidth $g = g_n$ is a positive-valued sequence of bandwidths satisfying

$$\lim_{n \rightarrow \infty} g = 0 \text{ and } \lim_{n \rightarrow \infty} n g^{2m+1} = \infty.$$

Thus g^{2m+1} decays to zero at a slower rate than n increases to infinity.

Assumption (3) is made mainly to ensure that the asymptotic variance term converges to zero, see expression (2.41) for more detail. From this assumption it is clear that g can be restricted to the form

$$g = c n^{-t(2m+1)} \quad \text{for } 0 < t < \frac{1}{2m+1}, \quad (2.25)$$

where c is a positive finite constant. The realization of this restriction will come in handy when minimizing the asymptotic mean squared error. Furthermore, we may write the estimator (2.24) as

$$\hat{\psi}_m(g) = \frac{1}{n^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n w_g^{(m)}(X_i - X_j) + \frac{1}{n} w_g^{(m)}(0), \quad (2.26)$$

where $w_g^{(m)}(0)/n$ is a constant. By noting that

$$MSE [\hat{\psi}_m(g)] = Var [\hat{\psi}_m(g)] + \{Bias [\hat{\psi}_m(g)]\}^2, \quad (2.27)$$

we are now able to derive an expression for the asymptotic mean squared error of $\hat{\psi}_m(g)$. Using (2.26) and (2.27) we will calculate the asymptotic bias and variance in turn. For the mathematical derivation of the results presented below the reader is referred to Wand and Jones (1995) and Koekemoer (1999).

CALCULATION OF ASYMPTOTIC BIAS

The following lemma will be needed to derive an expression for the bias term.

Lemma 2.1: If f is sufficiently smooth then

$$\int w_g^{(m)}(x-y)f(y)dy = \int w_g(x-y)f^{(m)}(y)dy \quad \text{and} \quad (2.28)$$

$$\int f^{(m)}(y)f^{(r)}(y)dy = \int f^{(m+r)}(y)f(y)dy = \psi_{m+r}. \quad (2.29)$$

Using (2.26) it follows that

$$E[\hat{\psi}_m(g)] = \left(1 - \frac{1}{n}\right) E[w_g^{(m)}(X_1 - X_2)] + \frac{1}{n}w_g^{(m)}(0). \quad (2.30)$$

From (2.30) it is clear that an expression is needed for $E[w_g^{(m)}(X_1 - X_2)]$. Using (2.28) and (2.29) it can be shown that

$$E[w_g^{(m)}(X_1 - X_2)] = \psi_m + \frac{g^r}{r!}\mu_r(w)\psi_{m+r} + O(g^{r+1}). \quad (2.31)$$

Using (2.30) and (2.31) we can now calculate the asymptotic bias of $\hat{\psi}_m(g)$

$$\begin{aligned} \text{Bias}[\hat{\psi}_m(g)] &= E\hat{\psi}_m(g) - \psi_m \\ &= \frac{1}{n}w_g^{(m)}(0) + \frac{g^r}{r!}\mu_r(w)\psi_{m+r} + O(g^{r+1}). \end{aligned} \quad (2.32)$$

CALCULATION OF ASYMPTOTIC VARIANCE

The following lemma will be needed to derive an expression for the variance term.

Lemma 2.2:

1. Let X_1, X_2, \dots, X_n be a set of i.i.d. random variables and let

$$U = 2\frac{1}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n S(X_i - X_j), \quad \text{where the function } S \text{ is symmetric about zero.}$$

Then

$$\begin{aligned} \text{Var}(U) &= \frac{2(n-1)}{n^3} \text{Var}[S(X_1 - X_2)] \\ &+ \frac{4(n-1)(n-2)}{n^3} \text{Cov}[S(X_1 - X_2), S(X_2 - X_3)]. \end{aligned} \quad (2.33)$$

2. $w^{(m)}$ is a symmetric function for m even.
3. With the assumption of significant smoothness on f

$$\begin{aligned} \iiint w_g^{(m)}(x-y)w_g^{(m)}(y-z)f(x)f(y)f(z)dx dy dz = \\ \iiint w_g(x-y)w_g(y-z)f^{(m)}(x)f^{(m)}(y)f^{(m)}(z)dx dy dz. \end{aligned} \quad (2.34)$$

We will now proceed to derive the asymptotic variance term. Using (2.24), (2.33) and (2) from Lemma 2.2 it follows that

$$\begin{aligned} & \text{Var} [\hat{\psi}_m(g)] \\ &= \text{Var} \left[\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n w_g^{(m)}(X_i - X_j) \right] = \text{Var} \left[\frac{2}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n w_g^{(m)}(X_i - X_j) \right] \\ &= \frac{2(n-1)}{n^3} \text{Var} [w_g^{(m)}(X_1 - X_2)] \\ &+ \frac{4(n-1)(n-2)}{n^3} \text{Cov} [w_g^{(m)}(X_1 - X_2), w_g^{(m)}(X_2 - X_3)]. \end{aligned} \quad (2.35)$$

In order to calculate (2.35) it should be noted that we may write the variance term as

$$\text{Var} [w_g^{(m)}(X_1 - X_2)] = E [w_g^{(m)}(X_1 - X_2)]^2 - [Ew_g^{(m)}(X_1 - X_2)]^2, \quad (2.36)$$

and the covariance term as

$$\begin{aligned} & \text{Cov} [w_g^{(m)}(X_1 - X_2), w_g^{(m)}(X_2 - X_3)] \\ &= E [w_g^{(m)}(X_1 - X_2)w_g^{(m)}(X_2 - X_3)] - Ew_g^{(m)}(X_1 - X_2)Ew_g^{(m)}(X_2 - X_3). \end{aligned} \quad (2.37)$$

The calculation of $\text{Var} [\hat{\psi}_m(g)]$ will proceed as follows: first we will calculate the values $E [w_g^{(m)}(X_1 - X_2)]^2$, $Ew_g^{(m)}(X_1 - X_2)$ and $E [w_g^{(m)}(X_1 - X_2)w_g^{(m)}(X_2 - X_3)]$, we will then plug these values into (2.36) and (2.37), which will be used to calculate the asymptotic variance expression given in (2.35).

First we find

$$E [w_g^{(m)}(X_1 - X_2)]^2 = \frac{1}{g^{2m+1}} \psi_0 R(w^{(m)}) + o(g^{-2m-1}). \quad (2.38)$$

Secondly, using (2.28) it follows that:

$$Ew_g^{(m)}(X_1 - X_2) = \psi_m + o(1). \quad (2.39)$$

Lastly, using (2.34) we find

$$E \left[w_g^{(m)}(X_1 - X_2) w_g^{(m)}(X_2 - X_3) \right] = \int \{f^{(m)}(y)\}^2 f(y) dy + o(1). \quad (2.40)$$

The asymptotic variance $Var \left[\hat{\psi}_m(g) \right]$ can now be calculated by substituting (2.38), (2.39) and (2.40) into (2.36) and (2.37), then substitute the results into (2.35). The result of these substitutions are

$$\begin{aligned} & Var \left[\hat{\psi}_m(g) \right] \\ &= 2 \frac{1}{n^2} \left[\frac{1}{g^{2m+1}} \psi_0 R \left(w^{(m)} \right) \right] + 4 \frac{1}{n} \left[\int f^{(m)}(y)^2 f(y) dy - \psi_m^2 \right] + o \left(n^{-2} g^{-2m-1} + n^{-1} \right). \end{aligned} \quad (2.41)$$

CALCULATION OF THE RESULTING ASYMPTOTIC MEAN SQUARED ERROR

Using the expressions for the asymptotic bias (2.32) and the asymptotic variance (2.41), we can now proceed to calculate the asymptotic mean squared error of $\hat{\psi}_m(g)$ using (2.27). It follows that

$$\begin{aligned} AMSE \left[\hat{\psi}_m(g) \right] &= 2 \frac{1}{n^2} \left[\frac{1}{g^{2m+1}} \psi_0 R \left(w^{(m)} \right) \right] + 4 \frac{1}{n} \left[\int f^{(m)}(y)^2 f(y) dy - \psi_m^2 \right] \\ &\quad + \left\{ \frac{1}{n g^{m+1}} w^{(m)}(0) + \frac{g^r}{r!} \mu_r(w) \psi_{m+r} \right\}^2. \end{aligned} \quad (2.42)$$

THE OPTIMAL DATA-DRIVEN g

In this section will find the optimal data-driven bandwidth by minimizing the AMSE given in expression (2.42). At first sight this seems a daunting task. However, by utilizing the restricted form of g , given in (2.25), the minimization process can be simplified. One would hope that the asymptotic expression (2.42) will converge to zero as $n \rightarrow \infty$. This will happen if both the asymptotic variance and asymptotic squared bias terms converge to zero. By inspecting (2.42) it is clear that the required convergence will be obtained if both of the terms

$$\frac{1}{n^2 g^{2m+1}} \quad \text{and} \quad \frac{1}{n^2 g^{2m+2}}$$

converge to zero. The first term given above belongs to the asymptotic variance and the second term belongs to the asymptotic squared bias expression. Recall that the

restriction on g , given in (2.25), is given by

$$g = cn^{-t(2m+1)} \quad \text{for } 0 < t < \frac{1}{2m+1}.$$

Using this choice of g we find that

$$c^{-2m-1}n^{-2}n^{t(2m+1)(2m+1)} \quad \text{and} \quad c^{-2m-2}n^{-2}n^{t(2m+1)(2m+2)}$$

must both converge to zero. The convergence is obtained if

$$-2 + t(2m+1)(2m+1) < 0 \quad \text{and} \quad -2 + t(2m+1)(2m+2) < 0,$$

implying that

$$0 < t < \frac{2}{(2m+1)(2m+1)} \quad \text{and} \quad 0 < t < \frac{2}{(2m+1)(2m+2)}.$$

From the expressions above it is clear that if we choose the value of t according to the variance term, the squared bias might not converge to zero, but, if we choose t according to the bias term, both the squared bias and variance terms will converge to zero. For the reason outlined above, we will minimize (2.42) by allowing the bias term to vanish. Thus, by setting the bias term (see (2.32)) equal to zero we obtain the AMSE optimal bandwidth

$$g_{AMSE,m} = \left[\frac{-r!w^{(m)}(0)}{\mu_r(w)\psi_{m+r}} \right]^{1/(m+r+1)} n^{-1/(m+r+1)}. \quad (2.43)$$

Replacing the AMSE optimal bandwidth (2.43) into (2.42) yields that

$$\begin{aligned} AMSE \left[\hat{\psi}_m(g_{AMSE,m}) \right] &= O \left(n^{-(2r+1)/(m+r+1)} \right) \\ &= O \left(n^{-5/7} \right), \end{aligned}$$

if we choose $r = 2$ (second order kernel) and $m = 4$ for the estimation of $\psi_4 = R(f'')$.

The method of Sheather and Jones

In this section the well respected plug-in method of Sheather and Jones (1991) will be described. Before proceeding consider the following short summary of important previous results.

- From (2.14), the AMISE optimal bandwidth for estimation of f is given by

$$h_{AMISE} = \left[\frac{R(k)}{\mu_2(k)^2 R(f'')} \right]^{\frac{1}{5}} n^{-1/5} = \left[\frac{R(k)}{\mu_2(k)^2 \psi_4} \right]^{\frac{1}{5}} n^{-1/5}.$$

- From (2.24), the estimator

$$\hat{\psi}_m(g) = \frac{1}{n} \sum_{i=1}^n \hat{f}^{(m)}(X_i; g) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n w_g^{(m)}(X_i - X_j),$$

is proposed for the unknown parameter $\psi_m = R(f^{(m/2)})$.

- From (2.43), the AMSE optimal bandwidth for estimation of ψ_m is given by

$$g_{AMSE,m} = \left[\frac{-r! w^{(m)}(0)}{\mu_r(w) \psi_{m+r}} \right]^{1/(m+r+1)} n^{-1/(m+r+1)},$$

where r is the order of the kernel function w . Note that since the standard normal kernel is used in this dissertation we have $r = 2$.

From the summary given above it is clear that in order to find an AMISE optimal bandwidth an estimate of ψ_4 is required. The AMSE optimal bandwidth, g , needed to estimate ψ_4 requires an estimate of ψ_6 . Again using the kernel method to estimate ψ_6 , an estimate of ψ_8 is required for the optimal AMSE bandwidth. In general an estimate of ψ_{m+r} is required for the estimation of ψ_m . The procedure is therefore recursive. Sheather and Jones (1991) proposed to stop this recursive behavior after l stages by plugging in the normal reference for f in the l^{th} stage. Hence, the normal scaled rule of thumb (2.21) can be considered as a Sheather & Jones procedure with $l = 0$.

Using a normal reference for f and properties of the normal distribution from Section 1.2, in specific (2f) and (2g), it follows that

$$\psi_m = \int \phi_\sigma^{(m)}(x - \mu) \phi_\sigma(x - \mu) dx = \frac{(-1)^{m/2} m!}{(2\sigma)^{m+1} (m/2)! \sqrt{\pi}} \quad \text{for } m \text{ even.} \quad (2.44)$$

AN EXAMPLE:

SHEATHER AND JONES (1991) PROCEDURE WITH $l = 2$ and $w = \phi$

For comfortable reading we will use the notation $\hat{g}_{AMSE,m} = \hat{g}_m$ and $\hat{\psi}_m(\hat{g}_{AMSE,m}) = \hat{\psi}_m$ in the following illustration. We will also use the standard normal kernel function in all the estimation procedures, thus $k(\cdot) = w(\cdot) = \phi(\cdot)$.

Step 1 Estimate ψ_8 using the normal reference, thus

$$\hat{\psi}_8 = \frac{105}{32\sqrt{\pi}\hat{\sigma}^9}.$$

[using (2.44), and $\hat{\sigma}$ as defined in (2.22)]

Step 2 Use $\hat{\psi}_8$ to estimate ψ_6 , thus

$$\hat{\psi}_6 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \phi_{\hat{g}_6}^{(6)}(X_i - X_j), \text{ where}$$

$$\phi^{(6)}(x) = (x^6 - 15x^4 + 45x^2 - 15) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

[using from Section 1.2, (2e) and (2f)]

In the expression above \hat{g}_6 is obtained through direct application of (2.43), thus

$$\hat{g}_6 = \left[\frac{-2\phi^{(6)}(0)}{\mu_2(\phi)\hat{\psi}_8} \right]^{1/9} n^{-1/9}, \quad \text{where } \mu_2(\phi) = 1 \text{ and } \phi^{(6)}(0) = \frac{-15}{\sqrt{2\pi}}.$$

[using from Section 1.2, 2(1)i and 2(1)iv]

Step 3 Use $\hat{\psi}_6$ to estimate $\psi_4 = R(f'')$, thus

$$\hat{\psi}_4 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \phi_{\hat{g}_4}^{(4)}(X_i - X_j), \text{ where } \phi^{(4)}(x) = (x^4 - 6x^2 + 3) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

[using from Section 1.2, (2e) and (2f)]

In the expression above \hat{g}_4 is obtained through direct application of (2.43), thus

$$\hat{g}_4 = \left[\frac{-2\phi^{(4)}(0)}{\mu_2(\phi)\hat{\psi}_6} \right]^{1/7} n^{-1/7}, \quad \text{where } \mu_2(\phi) = 1 \text{ and } \phi^{(4)}(0) = \frac{3}{\sqrt{2\pi}}.$$

[using from Section 1.2, 2(1)i and 2(1)iii]

Step 4 Use $\hat{\psi}_4$ to calculate the AMISE optimal bandwidth, h , (direct plug-in) thus

$$\hat{h}_{DPI,2} = \left[\frac{R(\phi)}{\mu_2(\phi)^2 \hat{\psi}_4} \right]^{\frac{1}{5}} n^{-1/5}, \quad \text{where } R(\phi) = \frac{1}{2\sqrt{\pi}} \text{ and } \mu_2(\phi) = 1.$$

[using from Section 1.2, (2(1)v)]

In the example above the two-stage procedure was described. One can, however, speculate as to what a suitable value of l will be. Wand and Jones (1995) simulated 500 bandwidths ($\hat{h}_{DPI,l}$) using the direct plug-in rule with $l = 0, 1, 2, 3$ for samples of size 100 from the skewed bimodal density (see Section 5.1 for the definition and a graph of this density). Subsequently they calculated $\log_{10}(\hat{h}_{DPI,l}) - \log_{10}(h_{MISE})$ and estimated the densities from these samples. The reader is referred to Figure 3.4, p. 73, of Wand and Jones (1995) for inspection of the results, from which it is clear that as l increases the selected bandwidth becomes less-biased, however, the extra functional estimation steps for larger

l lead to increased variability in the selected bandwidth. Hence, a variance-bias trade-off. According to some theoretical considerations, Aldershof (1991) and Park and Marron (1992) suggest to choose l to be at least 2, with $l = 2$ being a common choice and consequently implemented in this dissertation. Sheather and Jones (1991) also suggested a so-called solve the equation method, which will not be considered in this dissertation.

2.1.5 Boundary bias

In this section we will explore the behavior of the kernel density estimator near the boundary domain of the random variable considered, provided that the random variable is naturally bounded from below, above or both. Consider the i.i.d. random variables X_1, X_2, \dots, X_n with support on $[a, b]$. Suppose that the associated density is estimated using the kernel density estimator (2.1), with the usual restrictions on the kernel function, i.e., k is a nonnegative symmetric kernel function so that $\int k(z)dz = 1$, $\int zk(z)dz = 0$ and $\int z^2k(z)dz < \infty$ and the usual assumptions for the asymptotic calculations to be valid with special reference to the restriction $\lim_{n \rightarrow \infty} h = 0$ and $\lim_{n \rightarrow \infty} nh = \infty$. Then from arguments leading to the bias expression (2.10) it is clear that from asymptotic considerations we may write

$$E\hat{f}(x; h) = f(x) \int_{\frac{x-b}{h}}^{\frac{x-a}{h}} k(z)dz - hf'(x) \int_{\frac{x-b}{h}}^{\frac{x-a}{h}} zk(z)dz + \frac{1}{2}h^2 f''(x) \int_{\frac{x-b}{h}}^{\frac{x-a}{h}} z^2k(z)dz + O(h^3). \quad (2.45)$$

In what is to follow it will become apparent that the kernel function k plays an important role in the occurrence of boundary bias. Hence, we will consider the behavior of the kernel estimates at the boundary using both the compactly supported “polynomial kernel” ($-1 \leq z \leq 1$) and the standard normal kernel function which enjoys unbounded support. The reader is referred to Table 2.2 for a summary of these kernels. For explanatory purposes, the standard exponential population density and the uniform population density on $[0, 1]$ will be considered. These densities are bounded with the support $(0, +\infty)$ and $[0, 1]$ respectively.

Boundary behavior when using compactly supported “polynomial kernels”

By positioning ourselves at the lower bound a , using (2.45) and the fact that $(a - b)/h \rightarrow -\infty$ and $(b - a)/h \rightarrow +\infty$ since $\lim_{n \rightarrow \infty} h = 0$, it should be clear that

$$E\hat{f}(a; h) = f(a) \int_{\frac{a-b}{h}}^0 k(z)dz + O(h) = f(a) \int_{-1}^0 k(z)dz + O(h) = \frac{1}{2}f(a) + O(h),$$

and similarly, if the support is bounded from above, then a position assumed at the bound b will result in

$$E\hat{f}(b; h) = f(b) \int_0^{\frac{b-a}{h}} k(z) dz + O(h) = f(b) \int_0^1 k(z) dz + O(h) = \frac{1}{2}f(b) + O(h).$$

From the two expressions above and (2.45) it is clear that

- Half of the actual density is returned on expectation at the boundaries.
- The order of the remaining bias terms will reduce from $O(h^3)$ (for densities with a unbounded support) to $O(h)$ in the region of the boundaries.
- In the interior (see later on for a definition of the interior) of the bounded support, no boundary effects are detectable.

One needs to explain this phenomenon mathematically in order to rectify it. The occurrence of this phenomenon is easily understood once we inspect the actual form of the kernel density estimator given in (2.1) ,i.e.,

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n k_h(x - X_i).$$

From the expression above it is clear that the kernel estimator is an average of n kernel functions evaluated in the points $(x - X_i)/h$ and scaled according to h . Consequently, although the density f might be compactly supported, the kernel weights will be significantly different from zero outside the compactly supported domain. This means that the support of $\hat{f}(x; h)$ will be wider than the support of $f(x)$. The support of the kernel estimator, $\hat{f}(x; h)$, depends on the support of the kernel function k . For the “polynomial kernels” considered here, the kernel weight contributions to $\hat{f}(x; h)$ will be zero for any x where

$$\frac{x - X_i}{h} \leq -1 \quad \text{or} \quad \frac{x - X_i}{h} \geq +1, \quad i = 1, 2, \dots, n.$$

Consider the order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, then, the inequalities above imply that the kernel estimator will be compactly supported on the interval

$$[X_{(1)} - h; X_{(n)} + h], \tag{2.46}$$

which, in addition, means that extrapolation beyond these bounds is impossible. This serves as another motivation for using the standard normal kernel, for which extrapolation

is possible further into the tails of densities with an unbounded support. Consider the positions $a + h$ and $b - h$, then using (2.45) we find

$$E\hat{f}(a + h; h) = f(a + h) + \frac{1}{2}h^2 f''(a + h) \int_{-1}^{+1} z^2 k(z) dz + O(h^3) \quad \text{and}$$

$$E\hat{f}(b - h; h) = f(b - h) + \frac{1}{2}h^2 f''(b - h) \int_{-1}^{+1} z^2 k(z) dz + O(h^3).$$

The expressions above is also true for any point, x , in the interval

$$[a + h; b - h],$$

which is defined as the interior of the bounded domain where no boundary bias is present. Hence, for the “polynomial kernels” boundary bias is restricted to the domains

$$[a; a + h) \quad \text{and} \quad (b - h; b].$$

Broadly speaking, one may think of boundary bias as the kernel estimator having to find a compromise between estimating the two distinct values of f on either side of the boundary. An example is shown in Figure 2.4 using the standard normal kernel. It was noted earlier that the effect of boundary bias is determined by the kernel used, thus, preceding some popular solutions, a short discussion now follows to illustrate this effect when the standard normal kernel is used.

Boundary behavior when using the unbounded standard normal kernel

For the standard normal kernel, $k(\cdot) = \phi(\cdot)$, defining the support of the kernel estimator is not as clear since the support of the kernel function is infinite. However, using the fact that a random variable following the standard normal distribution contains almost all of its possible values between the bounds $[-4; +4]$ we may argue that

$$\phi(z) \approx 0 \quad \text{where } z \leq -4 \quad \text{and} \quad z \geq +4.$$

The kernel weight contributions to $\hat{f}(x; h)$ will then be zero for any x where

$$\frac{x - X_i}{h} \leq -4 \quad \text{or} \quad \frac{x - X_i}{h} \geq +4, \quad i = 1, 2, \dots, n.$$

The inequalities above imply that the kernel estimator will for all practical purposes be compactly supported on the interval

$$[X_{(1)} - 4h; X_{(n)} + 4h], \tag{2.47}$$

where $X_{(1)}$ and $X_{(n)}$ are the minimum and maximum values respectively. Similar arguments as presented previously lead to the domain

$$[a; a + 4h) \text{ and } (b - 4h; b],$$

where boundary bias is observed and

$$[a + 4h; b - 4h],$$

where boundary bias is absent. A comparison of (2.46) with (2.47) reveals that the domain in which boundary bias is expected is larger when using a standard normal kernel as opposed to a “polynomial kernel”. Although the conclusion is correct, one should keep in mind that the AMISE optimal bandwidths utilized for each of these kernels are different. To explain this, consider the Epanechnikov kernel for which it is easy to verify that, using (2.14), the AMISE optimal bandwidth is given by

$$h_{AMISE} = \left(\frac{3}{5}\right)^{1/5} \left(\frac{1}{5}\right)^{-2/5} [R(f'')n]^{-1/5} = 1.718772 [R(f'')n]^{-1/5}.$$

In comparison, the AMISE optimal bandwidth for the standard normal kernel is given by

$$h_{AMISE} = \left(\frac{1}{2\sqrt{\pi}}\right)^{1/5} [R(f'')n]^{-1/5} = 0.776388 [R(f'')n]^{-1/5}.$$

Consequently, the AMISE optimal bandwidth obtained from the Epanechnikov kernel is more than twice the size of the AMISE bandwidth based on the standard normal kernel. Hence, the domain in which boundary bias can be expected using the standard normal kernel is greater than that of the Epanechnikov kernel but not four times greater as indicated in the expressions (2.46) and (2.47). It is merely twice as big. It is evident that the boundary bias effect could be combated far more efficiently by only considering compactly supported kernels. Using the standard normal density as kernel, the boundary bias effect is explained graphically using 10 data points from the standard exponential and the uniform distribution on $[0, 1]$. The output is displayed in Figure 2.4. The kernel weights are superimposed on both of these graphs, showing the cause of boundary bias. Note that for the uniform density the kernel estimate attempts to estimate 1 in the domain $[0, 1]$ and 0 elsewhere. Hence, a compensation is made leading to the appearance of boundary bias.

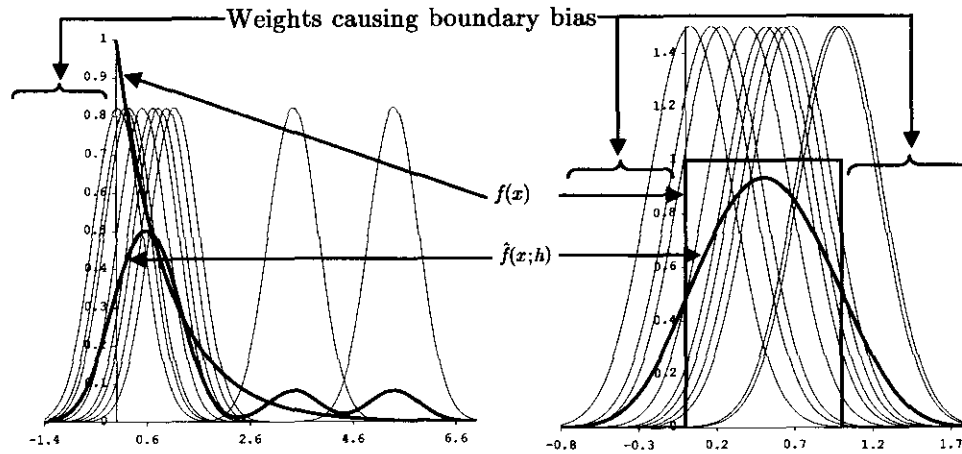


Figure 2.4:
 Left Panel: boundary bias for $n = 10$ values from the standard exponential distribution
 Right Panel: boundary bias for $n = 10$ values from the uniform distribution on $[0, 1]$

Some popular solutions

Since the location of the boundary is usually known, $\hat{f}(x;h)$ can be adapted to achieve better performance in the boundary vicinity. There is an extensive literature on how to correct this boundary effect. From the discussion above it is clear that the boundary effect is more efficiently combated using compactly supported kernels. It is therefore no surprise that the existing literature utilizes these kernels. The plug-in bandwidth selection procedure proposed by Sheather and Jones (1991) may be employed in the discussion below. For convenience, consider the frequently occurring case where the support of the unknown density f is in the interval $[0, \infty)$. For this domain and using a compactly supported kernel function we find that from (2.45)

$$E\hat{f}(x;h) = f(x)v_{0,\alpha}(k) - hf'(x)v_{1,\alpha}(k) + \frac{1}{2}h^2f''(x)v_{2,\alpha}(k) + O(h^3),$$

where $v_{i,\alpha}(k) = \int_{-1}^{\alpha} z^i k(z) dz$ and $\alpha = \min(x/h, 1)$. From the expression above an obvious first solution is to normalize $\hat{f}(x;h)$ by dividing with $v_{0,\alpha}(k)$, this will achieve consistency near the boundary, but still results in $O(h)$ bias there.

Alternatively, we may replace each observation X_i with its positive reflection $-X_i$ and consider the estimation problem of f at and near 0 based on the extended sample

$\{X_1, \dots, X_n, -X_1, \dots, -X_n\}$. This is the ordinary reflection method of Schuster (1985), Silverman (1986) and Cline and Hart (1991). The resulting density estimator utilized in Silverman (1986) is then given by

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n [k_h(x - X_i) + k_h(x + X_i)].$$

Another possible solution is to replace the kernel function with a so-called “boundary kernel”. These boundary kernel functions reduce to ordinary kernel functions in places where no boundary bias is present and changes form for each value of x where boundary bias is present. One family of boundary kernels is given by Gasser and Müller (1979) and is of the form

$$k(z, \alpha) = \frac{v_{2,\alpha}(k) - v_{1,\alpha}(k)z}{v_{0,\alpha}(k)v_{2,\alpha}(k) - v_{1,\alpha}(k)^2} k(z) I_{\{-1 \leq z \leq \alpha\}},$$

where $k(\cdot)$ is any of the “polynomial kernels.” From the expression above it is clear that the boundary kernel will reduce to the ordinary kernel function, $k(\cdot)$, if $\alpha = 1$, i.e., $x > h$. Another boundary kernel that can be used is (see Zhang, Karunamuni and Jones (1999))

$$k_\alpha(z) = \frac{12}{(1 + \alpha)^4} (1 + z) \left\{ (1 - 2\alpha)z + \frac{3\alpha^2 - 2\alpha + 1}{2} \right\} I_{\{-1 \leq z \leq \alpha\}}.$$

Note that if $\alpha = 1$, i.e., $x > h$, then the boundary kernel given above reduces to the Epanechnikov kernel. The reflection idea and the use of the boundary kernel are also easily implemented for the case where f is compactly supported on the interval $[a, b]$, with minor adjustments. For a boundary kernel, the kernel function is varying in the region where the boundary bias is present. Zhang and Karunamuni (1998) proposed to vary the bandwidth in this region as well. Using the boundary kernel $k_\alpha(z)$ the resulting boundary kernel estimator with bandwidth variation function is then defined as

$$\hat{f}_B(x) = \frac{1}{nh_\alpha} \sum_{i=1}^n k_{\alpha/(2-\alpha)} \left(\frac{x - X_i}{h_\alpha} \right),$$

where $h_\alpha = (2 - \alpha)h$. Again note that the estimator given above reduces to the usual kernel estimator in regions where boundary bias does not exist.

Zhang et al. (1999) proposed a more advanced reflection technique where a transformation is used to generate pseudo data beyond the left endpoint of the support of the density. The kernel estimator is then of the form

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n [k_h(x - X_i) + k_h(x + g_n(X_i))],$$

where $g_n(x) = x + d_n x^2 + A d_n^2 x^3$. With the requirement that $3A > 1$ and the recommended value being $A = 0.55$. Zhang et al. (1999) suggest $d_n = (\log f_n^*(h) - \log f_n^*(0)) / h$ where $f_n^*(h) = f_n(h) + 1/n^2$ and $f_n^*(0) = \max(\hat{f}_B(0), 1/n^2)$. Here $f_n(\cdot)$ is the ordinary kernel density estimator as defined in (2.1) and $\hat{f}_B(\cdot)$ is as defined previously in this discussion. Another pseudo data method estimator is that of Cowling and Hall (1996), defined by

$$\hat{f}(x; h) = \frac{1}{nh} \left\{ \sum_{i=1}^n k\left(\frac{x - X_i}{h}\right) + \sum_{i=1}^m k\left(\frac{x - X_{(-i)}}{h}\right) \right\},$$

where $X_{(-i)} = -5X_{(i/3)} - 4X_{(2i/3)} + (10/3)X_{(i)}$ and $X_{(i)}$ is the i^{th} order statistic of X_1, \dots, X_n . The authors suggest using $m = n^{9/10}$.

Jones and Foster (1996) proposed the nonnegative adaptation estimator

$$\hat{f}_{JF}(x; h) = \tilde{f}(x; h) \exp \left\{ \frac{\hat{f}(x; h)}{\tilde{f}(x; h)} - 1 \right\},$$

where $\tilde{f}(x; h) = [1/(nh \int_{-1}^1 k(z) dz)] \sum_{i=1}^n k\left(\frac{x - X_i}{h}\right)$ and where $\hat{f}(x; h)$ may be replaced by a boundary kernel estimate. Alternative boundary correction procedures are proposed by Hjort (1996), Alberts and Karunamuni (2003) and Burnham, Anderson and Laake (1980), among others.

From the discussion above it is clear that the topic of boundary bias has been studied extensively. In this dissertation the transformation method will be employed to combat boundary bias. Marron and Ruppert (1994) utilize the kernel transformation estimator defined in (2.48) below to reduce boundary bias where the support of the unknown density f is in the interval $[0, 1]$. They propose to transform the data, using parametric transformations, to a density that has its first derivative equal to 0 at both boundaries of its support. The density of the transformed data is estimated, and an estimate of the density of the original data is obtained by a change of variable. Although Marron and Ruppert (1994) estimate the density of the transformed data using a reflection estimator, the usual kernel density estimator will be employed in this dissertation. In this dissertation we will consider the transformation $Y_i = t(X_i)$, that transforms the input data to normality. To avoid confusion with the bandwidth, g , used to estimate ψ_m we will use the notation $t(\cdot)$ to indicate the transformation function in the discussion to follow. The transformation function will, however, be redefined in Chapter 4, where $g_\lambda(\cdot)$ will indicate the transformation to normality. Since the normal distribution has unbounded

support, it is clear that the boundary bias problem for the density estimate of the random variable Y is eliminated. Using standard probability theory the density estimate of the input data is then given by

$$\hat{f}_X(x) = t'(x)\hat{f}_Y(t(x)), \quad (2.48)$$

where $\hat{f}_Y(t(x))$ is the usual kernel density estimator of the transformed data as defined in (2.1). A discussion on how to select the AMISE optimal bandwidth required is presented in Chapter 4. The idea of a transformation to normality is conceivable as a more natural solution for boundary bias, since

- Boundary bias is eliminated for the kernel density estimator of the random variable Y , consequently the kernel estimator of the random variable X will be unaffected by boundary bias provided that the transformation $t(\cdot)$ does not inherit the boundary bias. The latter statement is important since, for some nonparametric transformation functions the boundary bias problem is evident in the derivative of the transformation function. The reader is referred to Chapter 4 for more detail.
- The boundary bias problem is addressed automatically and for any bounded support, i.e., $[a, b]$, $[a, +\infty)$ or $(-\infty, b]$. Hence, no special adaptation is required to implement the transformation procedure for different bounded supports.
- Provided that the transformation function $t(\cdot)$ does not inherit the boundary bias, the standard normal kernel function can be used as opposed to one of the “polynomial kernels.”

From the discussion above it is apparent that the transformation $t(\cdot)$ should be well defined and be able to transform any data successfully to normality. The reader is referred to Chapter 3 for a discussion of transformations and the introduction of a new optimal transformation to normality.

2.1.6 Spurious bumps in the tails

In this section the behavior of the kernel density estimator is inspected in regions where data are scarce. To illustrate this behavior, consider the kernel density estimator (2.1) with $k(\cdot) = \phi(\cdot)$ for 10 data points from the kurtotic unimodal distribution (see Section 5.1 for the definition of this density) displayed in Figure 2.5. The tail region of the kernel estimator for the exponential data presented in Figure 2.4 also displays this

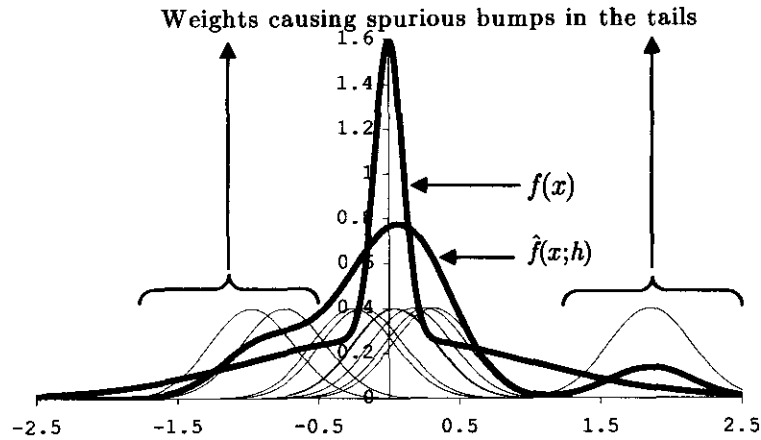


Figure 2.5: *The kernel density estimator for 10 data points from the kurtotic unimodal distribution with the kernel weights superimposed*

behaviour. Recall that from Section 2.1, using the standard normal density function as kernel, one can think of the kernel density estimator (2.1) at a specific point, say x , as the average of n normal density functions, each with mean X_i and standard deviation h . With this realization, the spurious bumps present in Figure 2.5 are easily explained. From arguments leading to (2.47) one can conclude that the weight contribution of the standard normal kernel function at the value X_i to the kernel estimator at a specific point, x , is zero if X_i is further than 4 bandwidths away from x . Similarly, from arguments leading to (2.46) it is clear that the weight contribution of the “polynomial kernel” function at the value X_i to the kernel estimator at a specific point, x , is zero if X_i is further than 1 bandwidth away from x . Consequently, we find that in regions where data are scarce only a limited amount of kernel functions contribute to the kernel estimator. Hence, spurious bumps occur in these regions, usually the tail regions of a distribution. It should be noted that although the “polynomial kernels” are more effective at combating boundary bias, the performance of these kernels is worse than that of the standard normal kernel at combating spurious bumps in the tails!

Some popular solutions

One natural way to deal with the occurrence of spurious bumps is to increase the bandwidth. However, from (2.13) it is clear that an increased bandwidth will result in a

density estimate that is less variable with increased bias, hence an over-smoothed estimate. Abramson (1982) suggests to use broader kernels in regions of low density. To identify regions of low density a pilot kernel estimate at the points X_i is obtained which is then used to scale the global AMISE bandwidth h . The procedure is outlined as follows:

Step 1 Find a pilot kernel estimate $\tilde{f}(x; b)$ that satisfies $\tilde{f}(X_i; b) > 0 \quad \forall i$.

Step 2 Define the local bandwidth factors λ_i by

$$\lambda_i = \left\{ \frac{\tilde{f}(X_i; b)}{g} \right\}^{-\alpha} \quad \text{where} \quad \log g = \frac{1}{n} \sum_{i=1}^n \log \tilde{f}(X_i; b).$$

g is the geometric mean of $\tilde{f}(X_i; b)$ and α is the sensitivity parameter, such that $0 \leq \alpha \leq 1$.

Step 3 Define the adaptive kernel estimate $\hat{f}(x; h)$ by

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i h} k\left(\frac{x - X_i}{\lambda_i h}\right), \quad (2.49)$$

where $k(\cdot)$ is the kernel function and h is the bandwidth.

The procedure outlined above is not sensitive to the pilot estimate $\tilde{f}(x; b)$, thus the usual kernel density estimator (2.1) may be used with a quick and simple bandwidth such as the normal scaled rule of thumb presented in (2.21) and (2.22). In this dissertation, however, the method of Sheather and Jones (1991) is employed, hence, we let $b = h$. The larger the power α , the more sensitive the method will be to variations in the pilot density, and consequently greater differences are observed between bandwidths used in different regions. For the value $\alpha = 0$, the adaptive kernel estimator reduces to the usual kernel estimator defined in (2.1). The recommended value $\alpha = \frac{1}{2}$ is used in this dissertation. The factor g^α has the advantage of freeing the bandwidth factors from the scale of the data and imposes the constraint that the geometric mean of the λ_i 's is equal to one.

Davison and Hall (1997) proposed a variable-bandwidth method that uses approximately equal amounts of information to estimate the density at all points. They argued that the amount of information, or number of data values, in the case of a compactly supported kernel, used to construct the estimate $\hat{f}(x; h)$ is approximately proportional to $nhf(x)$. Therefore, an equal-information argument suggests choosing $h = h(x)$ to ensure that $h\hat{f}(x|h) \approx \varepsilon$, where $\varepsilon > 0$ represents the fraction of total information used to construct

each estimate. This motivates an empirical construction of the bandwidth as a function of ε and is given by the following expression

$$\hat{h}_x = \inf \{ h : h \hat{f}(x|h) \geq \varepsilon \}.$$

Alternatively, the transformation kernel density estimator may be used to address the problem of spurious bumps. Recall that the transformation kernel density estimator (2.48), utilizing the transformation $Y_i = t(X_i)$, is given by

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^n t'(x) k_{h_y}(t(x) - t(X_i)),$$

where h_y is used to show the dependence of h on Y . Then, using the mean-value theorem, i.e., $t'(\eta_i) \approx (t(x) - t(X_i))/(x - X_i)$ where η_i is between x and X_i , we may write

$$\hat{f}_X(x) \approx \frac{1}{n [h_y/t'(x)]} \sum_{i=1}^n k\left(\frac{x - X_i}{h_y/t'(\eta_i)}\right).$$

This shows that the transformation kernel density estimator with a bandwidth h_y is similar to the conventional kernel estimator with bandwidth $h_y/t'(x)$. Hence, the transformation kernel density estimator can also be viewed as an estimator with a variable-bandwidth. In this dissertation transformation to normality is investigated. The new optimal semi-parametric transformation proposed in Section 3.4 has the property that $t'(x)$ is proportional to the density of a parametric pilot transformation of the input data. Therefore, although not the same, striking similarities are evident between the transformation method and the adaptive method of Abramson (1982), since both methods adapt the bandwidth according to density considerations. To avoid unnecessary detail at this stage, the reader is referred to Section 3.4 for the definition of the new transformation and to Chapter 4 for the implementation. For the transformation method, however, a lurking danger exists. To illustrate this, consider the transformation method as presented in (2.48). For a transformation to normality, it should be clear that $\hat{f}_Y(t(x))$, will be an estimate of the normal distribution, which is a light tailed distribution and consequently we can conclude that no spurious bumpiness will be present in the kernel estimate. However, the bumpy behavior can spill over to the derivative $t'(x)$, for certain nonparametric transformations. As a result the estimate $\hat{f}_X(x)$ will have spurious bumps in the tails of the distribution. However, the optimal transformation to normality presented in Section 3.4 is constructed to avoid this from happening.

In conclusion the following remarks: From the discussion presented in Section 2.1.5 and Section 2.1.6 it is clear that the transformation kernel density estimator addresses both the boundary bias and spurious bumps in the tails problems associated with kernel density estimation, *in a natural and automatic manner*. It should therefore be no surprise that this method of density estimation is recommended in this dissertation for both the novice and experienced statistician. Chapter 3 is devoted to the topic of transformation, since identifying the correct transformation is essential for the success of the transformation kernel density estimator.

2.2 Kernel distribution function estimation

Let X_1, \dots, X_n be i.i.d. continuous random variables from the probability law F_X , having a continuous univariate density f_X . Recall that from (2.1) the kernel density estimator is defined as

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

where h is the bandwidth or smoothing parameter and k is the so-called kernel function satisfying

- $k(u) \geq 0, \quad \forall u \in \mathbf{R}$.
- $\int k(u)du = \mu_0(k) = 1$, hence k is a density function.
- $k(-u) = k(u)$, hence k is a symmetric function.
- $\int uk(u)du = \mu_1(k) = 0$.
- $\int u^2k(u)du = \mu_2(k) = a^2 < +\infty$.

The corresponding kernel distribution function estimator is then defined as

$$\hat{F}(x; h) = \int_{-\infty}^x \frac{1}{nh} \sum_{i=1}^n k\left(\frac{t - X_i}{h}\right) dt = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (2.50)$$

where $K(u) = \int_{-\infty}^u k(t)dt$, $\lim_{u \rightarrow -\infty} K(u) = 0$ and $\lim_{u \rightarrow +\infty} K(u) = 1$. Although (2.50) suggests choosing the bandwidth h according to density considerations, i.e., h should be of order $O(n^{-1/5})$ as presented in (2.14), AMISE calculations for the estimator (2.50) suggest that the optimal bandwidth should actually be of order $O(n^{-1/3})$. This is further explored in Section 2.2.1. Using the standard normal distribution function as kernel

function, i.e., $K(\cdot) = \Phi(\cdot)$, the kernel distribution function estimator (2.50) at a point, say x , can be explained as the average of n normal distribution functions each being centered about X_i and having a standard deviation h . Consequently, values of X_i closer to x contribute more to the estimate at the point x . Using 10 data points from the standard normal distribution and $K(\cdot) = \Phi(\cdot)$, the kernel distribution function estimator is graphically explained in Figure 2.6. Another well-known distribution function estimator

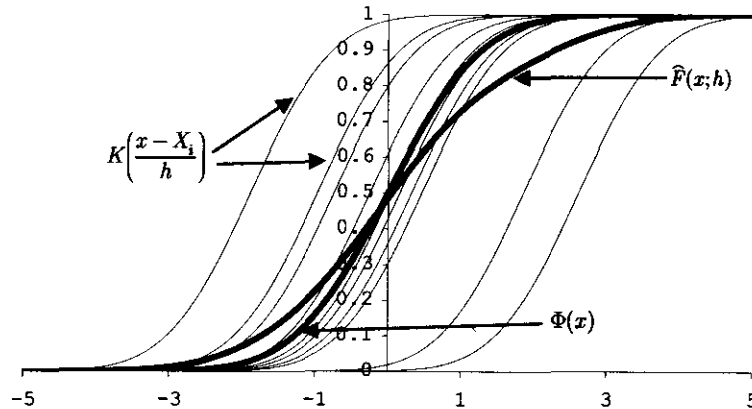


Figure 2.6: Kernel distribution function estimate for 10 data points from the standard normal distribution with the kernel weights superimposed

is the empirical distribution function defined by

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

where $I(\cdot)$ denotes the indicator function. It is easily verified that the expected value is $E[F_n(x)] = \int I(t \leq x) dF(t) = F(x)$, hence the empirical distribution function is an unbiased estimator. Furthermore, it is easily verified that

$$\begin{aligned} \text{Var}[F_n(x)] &= \frac{1}{n} \text{Var}[I(X_1 \leq x)] \\ &= \frac{1}{n} F(x) [1 - F(x)]. \end{aligned}$$

Next, the relationship between the empirical distribution function and the kernel distribution function estimator will be explored. To assist in the establishment of the relationship

note that, as $h \rightarrow 0$, we find

$$K\left(\frac{x - X_i}{h}\right) \rightarrow \begin{cases} K(-\infty) = 0 & \text{if } X_i > x, \\ K(0) = \frac{1}{2} & \text{if } X_i = x, \\ K(+\infty) = 1 & \text{if } X_i < x. \end{cases}$$

Hence, as $h \rightarrow 0$, we may write the kernel distribution function estimator as

$$\widehat{F}(x; 0) = \frac{1}{n} \sum_{i=1}^n \left[\mathbf{I}(X_i < x) + \frac{1}{2} \mathbf{I}(X_i = x) \right].$$

Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be the associated order statistics of X_1, X_2, \dots, X_n then, from the expression above, it is clear that $\widehat{F}(x; 0)$ and $F_n(x)$ differ only when evaluated in these order statistics, i.e.,

$$\widehat{F}(x; 0) = \begin{cases} \frac{i}{n} = F_n(x) & \text{if } x \neq X_{(i)}, \\ \frac{i-1}{n} + \frac{1}{2n} = \frac{i-1/2}{n} & \text{if } x = X_{(i)}. \end{cases} \quad (2.51)$$

The relationship presented in (2.51) is important since the form $\widehat{F}(x; 0)$ is often utilized instead of $F_n(x)$, for example in the construction of qq-plots (see Section 3.1) and in some of the parameter estimation procedures discussed in Section 3.3.3. It should also be pointed out that in some situations, such as finding an optimal transformation, the discrete nature of $F_n(x)$ can be problematic and therefore a smooth estimate such as the kernel distribution function estimator should be implemented. Furthermore, Reiss (1981) shows that kernel smoothing (2.50) reduces the variance, $(1/n)F(x)[1 - F(x)]$, of the empirical distribution function by an amount whose leading term is of the form $-C_1 h/n$, where $C_1 > 0$. However, the kernel estimator is not an unbiased estimator. It can be shown that the bias is of order $O(h^2)$. The reader is referred to the variance and bias expressions leading to (2.55) below for verification of the above-mentioned results. These results provide a theoretical justification for the usefulness of the kernel estimator (2.50).

2.2.1 An appropriate discrepancy measure

It was mentioned previously that AMISE calculations for the estimator (2.50) suggest that the optimal bandwidth is of order $O(n^{-1/3})$. To verify this claim let's proceed by

first calculating the asymptotic mean squared error and then find the global discrepancy measure AMISE. For the mathematical derivation of results presented in this section the reader is referred to van Graan (1982). The usual assumptions concerning the kernel functions k and K are adopted for the following derivation. Usually k is assumed to have a bounded support, but this assumption can be relaxed for certain kernel functions such as the standard normal density function which enjoys unbounded support. In addition it will be assumed that f is continuous, $f'(x)$ exists and that

$$\lim_{n \rightarrow \infty} h = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} nh = \infty.$$

It is well known that

$$\begin{aligned} MSE [\hat{F}(x; h)] &= E [\hat{F}(x; h) - F(x)]^2 \\ &= Var [\hat{F}(x; h)] + \{Bias [\hat{F}(x; h)]\}^2. \end{aligned} \quad (2.52)$$

Using (2.50), notation from Section 1.2, (1a), the i.i.d. assumption and assuming that f and k are both sufficiently smooth, we can calculate the following expectation

$$\begin{aligned} E [\hat{F}(x; h)] &= E \left[K \left(\frac{x - X_1}{h} \right) \right] = \int K \left(\frac{x - y}{h} \right) f(y) dy \\ &= F(x) + \frac{1}{2} h^2 f'(x) \mu_2(k) + O(h^3). \end{aligned} \quad (2.53)$$

Similarly, using the fact that $\int k(z)K(z)dz = \frac{1}{2}$, we find

$$\begin{aligned} E \left[K \left(\frac{x - X_1}{h} \right) \right]^2 &= \int \left[K \left(\frac{x - y}{h} \right) \right]^2 f(y) dy \\ &= F(x) - 2hf(x)C_1 + O(h^2), \end{aligned} \quad (2.54)$$

where $C_1 = \int zk(z)K(z)dz$. Using (2.53) and (2.54) we may write the asymptotic bias and variance terms as

$$\begin{aligned} Bias [\hat{F}(x; h)] &= \frac{1}{2} h^2 f'(x) \mu_2(k) + O(h^3) \\ Var [\hat{F}(x; h)] &= \frac{1}{n} Var \left[K \left(\frac{x - X_1}{h} \right) \right] \\ &= \frac{1}{n} F(x) [1 - F(x)] - \frac{2h}{n} f(x) C_1 + O(n^{-1} h^2). \end{aligned}$$

It should be noted that $2hf(x)C_1/n$ can be increased by choosing larger values of h , which will result in a smaller variance expression but larger bias, hence, a variance-bias

trade-off. Substitution of these values into (2.52) leads to the asymptotic mean squared error

$$AMSE [\hat{F}(x; h)] = \frac{1}{n} F(x) [1 - F(x)] - \frac{2h}{n} f(x) C_1 + \frac{1}{4} h^4 [f'(x)]^2 \mu_2(k)^2. \quad (2.55)$$

We can also define global measures of discrepancy between $\hat{F}(x; h)$ and $F(x)$. From similar arguments as those presented in the context of density estimation, we may define the mean integrated squared error

$$MISE [\hat{F}(\cdot; h)] = \int E [\hat{F}(x; h) - F(x)]^2 dx. \quad (2.56)$$

A more general choice would be to introduce weights so that the resulting discrepancy measure is always bounded, such a measure is given by

$$WMISE [\hat{F}(\cdot; h)] = \int E [\hat{F}(x; h) - F(x)]^2 W(x) dF(x), \quad (2.57)$$

where $W(x)$ is a nonnegative bounded weight function. Jones (1990), Bowman, Hall and Prvan (1998) and Polansky (1997) considered the discrepancy measure (2.56) and in contrast Reiss (1981), Swanepoel (1988), Sarda (1993) and Altman and Léger (1995) considered the discrepancy measure presented in (2.57), which has a more intuitive appeal. We will consider these global discrepancy measures in turn.

THE ASYMPTOTIC MEAN INTEGRATED SQUARED ERROR

If F has two bounded, continuous derivatives, each of which is ultimately monotone in both tails, and has sufficiently many finite moments, then it follows from (2.55), (2.56) and notation from Section 1.2 that

$$\begin{aligned} & AMISE [\hat{F}(\cdot; h)] \\ &= \frac{1}{n} \int F(x) [1 - F(x)] dx - \frac{2h}{n} C_1 + \frac{1}{4} h^4 \mu_2(k)^2 R(f'). \end{aligned}$$

Differentiating this AMISE expression with respect to h and using (2.23) results in the AMISE optimal bandwidth

$$h_{AMISE} = \left(\frac{2C_1}{\mu_2(k)^2 R(f')} \right)^{\frac{1}{3}} n^{-1/3} = \left(\frac{-2C_1}{\mu_2(k)^2 \psi_2} \right)^{\frac{1}{3}} n^{-1/3}. \quad (2.58)$$

To evaluate the performance of the AMISE optimal bandwidth, we substitute (2.58) into the AMISE expression above which yields

$$\inf_{h>0} AMISE [\widehat{F}(\cdot; h)] = \frac{1}{n} \int F(x) [1 - F(x)] dx - \left[\frac{3(2C_1)^{4/3}}{4(\mu_2(k)^2 R(f'))^{1/3}} \right] n^{-4/3}. \quad (2.59)$$

This shows that, if we use $h = h_{AMISE}$ defined in (2.58), the MISE of $\widehat{F}(x; h)$ is asymptotically smaller than that of the empirical distribution function $F_n(x)$.

THE WEIGHTED ASYMPTOTIC MEAN INTEGRATED SQUARED ERROR

Provided that F is sufficiently smooth, it follows from (2.55) and (2.57) that

$$WAMISE [\widehat{F}(\cdot; h)] = \frac{1}{n} \int F(x) [1 - F(x)] W(x) dF(x) - \frac{2h}{n} C_1 C_2 + \frac{1}{4} h^4 \mu_2(k)^2 C_3,$$

where

- $C_1 = \int z k(z) K(z) dz,$
- $C_2 = \int [f(x)]^2 W(x) dx,$
- $C_3 = \int [f'(x)]^2 f(x) W(x) dx.$

Differentiating this WAMISE expression with respect to h results in the WAMISE optimal bandwidth

$$h_{WAMISE} = \left(\frac{2C_1 C_2}{\mu_2(k)^2 C_3} \right)^{\frac{1}{3}} n^{-1/3}. \quad (2.60)$$

It should be noted that the conditions that F is twice differentiable and both F and $|f'|$ are bounded from below on the support of W , imposed by Sarda (1993), are considered not necessary by Altman and Léger (1995). Hence, the bandwidth presented in (2.60) should be preferred over the bandwidth defined in (2.58) as it is derived from a well-defined discrepancy measure. However, in the context of the newly proposed semi-parametric transformation to normality, a distribution function estimate is required based on data that are approximately normally distributed. Therefore, the assumptions made in the construction of h_{AMISE} in (2.58) should be valid and consequently (2.58), which is easier to implement, may be employed. The reader is referred to Section 3.4 for a detailed

discussion of this transformation. To evaluate the performance of the WAMISE optimal bandwidth, we substitute (2.60) into the WAMISE expression, yielding

$$\inf_{h>0} WAMISE [\widehat{F}(\cdot; h)] = \frac{1}{n} \int F(x) [1 - F(x)] W(x) dF(x) - \left[\frac{3(2C_1 C_2)^{4/3}}{4(\mu_2(k)^2 C_3)^{1/3}} \right] n^{-4/3}. \quad (2.61)$$

Estimation of the optimal bandwidth, considering both the discrepancy measures AMISE and WAMISE, will be explored further in Section 2.2.3.

2.2.2 The choice of an appropriate kernel function

In this section the choice of an appropriate kernel function will be explored. In the context of density estimation it was shown that the Epanechnikov kernel is the optimal kernel for estimating densities. The reader is referred to Table 2.2 for detail. With the application of the AMISE and WAMISE optimal bandwidths defined in (2.58) and (2.60) respectively, it is clear that the resulting AMISE and WAMISE presented in (2.59) and (2.61) only depend on the kernel function, $K(\cdot)$, through

$$C(k) = \left(\frac{C_1}{\mu_2(k)^{1/2}} \right)^{4/3}, \quad \text{where } C_1 = \int zk(z)K(z)dz \quad \text{and} \quad \mu_2(k) = \int z^2k(z)dz.$$

Clearly the kernel function that maximizes the functional $C(k)$ will minimize the resulting AMISE and WAMISE and will consequently be the optimal kernel function in the sense of AMISE and WAMISE. Consider the rescaled kernel function

$$k_\delta(\cdot) = \frac{1}{\delta} k\left(\frac{\cdot}{\delta}\right).$$

It should be noted that the functional $C(k)$ is invariant to rescaling of k , i.e., $C(k_{\delta_1}) = C(k_{\delta_2})$ for any $\delta_1, \delta_2 > 0$. Thus, any rescaled version of the optimal kernel function will still have optimal performance in the sense of AMISE and WAMISE. To find such an kernel, van Graan (1982) and Jones (1990) used the Cauchy-Schwarz inequality, i.e.,

$$\left(\int_0^T f(t)g(t)dt \right)^2 \leq \left(\int_0^T f^2(t)dt \right) \left(\int_0^T g^2(t)dt \right),$$

while the proofs presented by Swanepoel (1988) are based on the calculus of variations. Using the Cauchy-Schwarz inequality we will proceed by finding an upper bound for the

functional $\mu_2(k)^{-1/2}C_1$.

$$\begin{aligned}
\mu_2(k)^{-1/2}C_1 &= \mu_2(k)^{-1/2} \int_{-\infty}^{+\infty} zk(z)K(z)dz \\
&= \mu_2(k)^{-1/2} \left[\int_{-\infty}^0 zk(z)K(z)dz + \int_0^{+\infty} zk(z)K(z)dz \right] \\
&= \mu_2(k)^{-1/2} \left[\int_0^{+\infty} (-t)k(-t)K(-t)dt + \int_0^{+\infty} zk(z)K(z)dz \right] \\
&\hspace{15em} [\text{ using the substitution } t = -z] \\
&= \mu_2(k)^{-1/2} \left[- \int_0^{+\infty} zk(z)K(-z)dz + \int_0^{+\infty} zk(z)K(z)dz \right] \\
&\hspace{15em} [\text{ symmetry about zero }] \\
&= \mu_2(k)^{-1/2} \left[- \int_0^{+\infty} zk(z) [1 - K(z)] dz + \int_0^{+\infty} zk(z)K(z)dz \right] \\
&= \mu_2(k)^{-1/2} \left[\int_0^{+\infty} zk(z) [2K(z) - 1] dz \right] \\
&= \mu_2(k)^{-1/2} \left[\int_0^{+\infty} \{zk(z)^{1/2}\} \{k(z)^{1/2} [2K(z) - 1]\} dz \right] \\
&\leq \mu_2(k)^{-1/2} \left[\mu_2(k)^{1/2} \frac{1}{2\sqrt{3}} \right] \\
&\hspace{10em} [\text{ using the Cauchy-Schwarz inequality and 1(j)v from Section 1.2 }] \\
&= \frac{1}{2\sqrt{3}}. \hspace{15em} (2.62)
\end{aligned}$$

It is easy to verify that, for the uniform kernel function, i.e.,

$$k^*(z) = \begin{cases} \frac{1}{2} & -1 \leq z \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$K^*(z) = \begin{cases} 0 & z < -1, \\ \frac{1}{2}(z+1) & -1 \leq z \leq 1, \\ 1 & z > 1, \end{cases}$$

we find $\mu_2(k^*)^{-1/2}C_1 = \mu_2(k^*)^{-1/2} \int_{-1}^{+1} zk^*(z)K^*(z)dz = \frac{1}{2\sqrt{3}}$. Hence, the upper bound is attained with the use of the uniform kernel function, which is the optimal kernel function when estimating a distribution function with the kernel distribution function estimator (2.50). It should be noted that since $C(k)$ is invariant to rescaling the kernel function

k , any rescaled version of the uniform kernel function will also be optimal. To compare other kernel functions to the optimal uniform kernel we define the efficiency measure

$$\text{Eff}(k) = \left\{ \frac{C(k)}{C(k^*)} \right\}^{3/4}.$$

The efficiency measure is displaced in Table 2.3 for different kernel functions. The main

Table 2.3: Kernel functions and their efficiency when estimating distribution functions

Kernel Function	$k(z)$	$K(z)$	Support
Uniform	$\frac{1}{2}$	$\frac{1}{2} [1 + z]$	$-1 \leq z \leq 1$
Epanechnikov	$\frac{3}{4}(1 - z^2)$	$\frac{3}{4} \left[\frac{2}{3} + z - \frac{1}{3}z^3 \right]$	$-1 \leq z \leq 1$
Biweight	$\frac{15}{16}(1 - z^2)^2$	$\frac{15}{16} \left[\frac{8}{15} + z - \frac{2}{3}z^3 + \frac{1}{5}z^5 \right]$	$-1 \leq z \leq 1$
Triangular	$1 - z $	$\frac{1}{2} + z - \text{sign}(z) \frac{1}{2}z^2$	$-1 \leq z \leq 1$
Triweight	$\frac{35}{32}(1 - z^2)^3$	$\frac{35}{32} \left[\frac{16}{35} + z - z^3 + \frac{3}{5}z^5 - \frac{1}{7}z^7 \right]$	$-1 \leq z \leq 1$
Standard Normal	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$	$\Phi(z)$	$-\infty < z < +\infty$

Kernel Function	$\mu_2(k)$	$\int zk(z)K(z)dz$	Eff(k)
Uniform	$\frac{1}{3}$	$\frac{1}{6}$	1
Epanechnikov	$\frac{1}{5}$	$\frac{9}{70}$	0.99591
Biweight	$\frac{1}{7}$	$\frac{25}{231}$	0.99190
Triangular	$\frac{1}{6}$	$\frac{7}{60}$	0.98995
Triweight	$\frac{1}{9}$	$\frac{245}{2574}$	0.98917
Standard Normal	1	$\frac{1}{2\sqrt{\pi}}$	0.97721

message conveyed in Table 2.3 is that many other popular kernel functions result in AMISE and WAMISE negligible worse than that associated with the optimal uniform

kernel function. Also, one should choose the kernel function K using other considerations, such as smoothness of $\widehat{F}(x; h)$ or ease of computation. It is interesting to note that the uniform kernel causes “less smooth” distribution function estimates. For reasons outlined above the standard normal kernel distribution function, $\Phi(\cdot)$, is used in this dissertation.

2.2.3 The choice of a smoothing parameter

In this section we will discuss the choice of a suitable data-dependent smoothing parameter based on the two discrepancy measures generally used to assess the performance of the kernel distribution function estimator. It should be noted that the unweighted discrepancy measure defined in (2.56) will be utilized in this dissertation. However, for general application the weighted discrepancy measure defined in (2.57) is prescribed.

BANDWIDTH SELECTION BASED ON THE ASYMPTOTIC MEAN INTEGRATED SQUARED ERROR

In this section bandwidth selection is considered when the discrepancy measure defined in (2.56), i.e.,

$$MISE[\widehat{F}(\cdot; h)] = \int E[\widehat{F}(x; h) - F(x)]^2 dx,$$

is used. Bowman et al. (1998) proposed a method of cross-validation appropriate for the smoothing of distribution functions. Their method was compared to those of Sarda (1993), Altman and Léger (1995) (these two methods are based on the discrepancy measure (2.57)) and a simple plug-in bandwidth based on (2.58). They found that their cross-validatory proposal works well and that the simple plug-in bandwidth is also quite effective. Based on the good performance of the plug-in method, similar arguments from Altman and Léger (1995), and the success of the Sheather and Jones (1991) l -stage plug-in bandwidth utilized for density estimation, the plug-in method proposed by Polansky (1997) will be discussed and implemented in this dissertation with the standard normal distribution function as the kernel function.

Normal scaled rule of thumb

From expression (2.58) it is clear that the only unknown value is $R(f')$. As in the case of density estimation, the normal scaled rule of thumb involves replacing the only unknown

quantity f with a normal reference with mean μ and variance σ^2 . Using the properties of the normal distribution as discussed in Section 1.2, (2d), (2f) and (2g), we find that

$$\begin{aligned} R(\phi_\sigma^{(1)}(x - \mu)) &= \int \phi_\sigma^{(1)}(x - \mu)\phi_\sigma^{(1)}(x - \mu)dx = -\phi_{\sqrt{2}\sigma}^{(2)}(0) = \frac{1}{4\sqrt{\pi}\sigma^3}, \\ C_1 &= \int x\phi(x)\Phi(x)dx = -\int \phi^{(1)}(x)\Phi(x)dx = R(\phi) = \frac{1}{2\sqrt{\pi}}, \\ \mu_2(k) &= \mu_2(\phi) = \left[\int x^2\phi(x)dx \right] = 1. \end{aligned}$$

Hence, with the normal reference and the standard normal density function as kernel we find the normal scaled rule of thumb, by replacing the quantities calculated above into (2.58), i.e.,

$$h_{NS} = 4^{1/3}\sigma n^{-1/3} = 1.5874\sigma n^{-1/3}. \quad (2.63)$$

To implement (2.63) it is necessary to measure the spread of the data. Here, as in the case of density estimation we will use the robust scale estimator defined in (2.22), i.e.,

$$\hat{\sigma} = \min \left[s, \frac{\hat{q}_3 - \hat{q}_1}{\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right)} \right].$$

The method of Polansky (1997)

Polansky (1997) rightfully observed that an l -stage plug-in estimate of the AMISE optimal bandwidth, similar to that proposed by Sheather and Jones (1991) in the context of density estimation, is obtainable by estimating the functional $(-1)^{m/2}\psi_m = R(f^{(m/2)})$, $m = 0, 2, 4, 6, \dots$ as suggested by Hall and Marron (1987) and Sheather and Jones (1991). The estimation of the functional ψ_m is discussed in detail in Section 2.1.4 and is summarized as follows

- From (2.24)

$$\hat{\psi}_m(g) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n w_g^{(m)}(X_i - X_j),$$

where w is the kernel function.

- From (2.43)

$$g_{AMSE,m} = \left[\frac{-r!w^{(m)}(0)}{\mu_r(w)\psi_{m+r}} \right]^{1/(m+r+1)} n^{-1/(m+r+1)},$$

where r is the order of the kernel function used.

- From (2.43) it is clear that the estimate of ψ_2 depends on ψ_4 which in turn depends on ψ_6 etc., if a second order kernel function, such as the standard normal, is used. In general, an estimate of ψ_{m+2} is required for the estimation of ψ_m . The procedure is therefore recursive.
- From (2.44)

$$\psi_m = \frac{(-1)^{m/2} m!}{(2\sigma)^{m+1} (m/2)! \sqrt{\pi}} \quad \text{for } m \text{ even,}$$

using a normal reference for f .

Polansky (1997), similar to Sheather and Jones (1991), proposed to stop the recursive behavior after l stages by plugging in the normal reference for f in the l^{th} stage, with $l = 2$ being the norm. Hence the normal scaled rule of thumb (2.63) can be considered as a Polansky procedure with $l = 0$. Given the success of the Sheather & Jones procedure in the context of density estimation one can expect that the method proposed by Polansky will result in a highly effective plug-in procedure provided that the assumptions made to derive (2.58) are met. As mentioned previously, these assumptions will be considered valid in this dissertation since distribution function estimation is applied to data already possessing some form of normality. Consequently, the method of Polansky will be applied. The method for $l = 2$, $w(\cdot) = \phi(\cdot)$ and $K(\cdot) = \Phi(\cdot)$ is as follows:

Step 1 Estimate ψ_6 using the normal reference, thus

$$\hat{\psi}_6 = \frac{-15}{16\sqrt{\pi}\hat{\sigma}^7}.$$

[using (2.44), and $\hat{\sigma}$ as defined in (2.22)]

Step 2 Use $\hat{\psi}_6$ to estimate ψ_4 , thus

$$\hat{\psi}_4 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \phi_{\hat{g}_4}^{(4)}(X_i - X_j), \quad \text{where } \phi^{(4)}(x) = (x^4 - 6x^2 + 3) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

[using from Section 1.2, (2e) and (2f)]

In the expression above \hat{g}_4 is obtained through direct application of (2.43), thus

$$\hat{g}_4 = \left[\frac{-2\phi^{(4)}(0)}{\mu_2(\phi)\hat{\psi}_6} \right]^{1/7} n^{-1/7}, \quad \text{where } \mu_2(\phi) = 1 \text{ and } \phi^{(4)}(0) = \frac{3}{\sqrt{2\pi}}.$$

[using from Section 1.2, 2(1)i and 2(1)iii]

Step 3 Use $\hat{\psi}_4$ to estimate $-\psi_2 = R(f')$, thus

$$\hat{\psi}_2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \phi_{\hat{g}_2}^{(2)}(X_i - X_j), \text{ where } \phi^{(2)}(x) = (x^2 - 1) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

[using from Section 1.2, (2e) and (2f)]

In the expression above \hat{g}_2 is obtained through direct application of (2.43), thus

$$\hat{g}_2 = \left[\frac{-2\phi^{(2)}(0)}{\mu_2(\phi)\hat{\psi}_4} \right]^{1/5} n^{-1/5}, \text{ where } \mu_2(\phi) = 1 \text{ and } \phi^{(2)}(0) = \frac{-1}{\sqrt{2\pi}}.$$

[using from Section 1.2, 2(1)i and 2(1)ii]

Step 4 Use $\hat{\psi}_2$ to calculate the AMISE optimal bandwidth, h , (direct plug-in) thus

$$\hat{h}_{DPI,2} = \left[\frac{-2C_1}{\mu_2(\phi)^2 \hat{\psi}_2} \right]^{1/3} n^{-1/3}, \text{ where } \mu_2(\phi) = 1 \text{ and } C_1 = \frac{1}{2\sqrt{\pi}}.$$

[using from Section 1.2, 2(1)i and 2(1)xii]

BANDWIDTH SELECTION BASED ON THE WEIGHTED ASYMPTOTIC MEAN INTEGRATED SQUARED ERROR

In this section bandwidth selection is considered when the discrepancy measure defined in (2.57), i.e.,

$$WMISE[\hat{F}(\cdot; h)] = \int E[\hat{F}(x; h) - F(x)]^2 W(x) dF(x),$$

is used. Recall from (2.60) that the asymptotic optimal bandwidth using this discrepancy measure is given by

$$h_{WAMISE} = \left(\frac{2C_1 C_2}{\mu_2(k)^2 C_3} \right)^{1/3} n^{-1/3},$$

where (as before)

- $C_1 = \int z k(z) K(z) dz,$
- $C_2 = \int [f(x)]^2 W(x) dx,$
- $C_3 = \int [f'(x)]^2 f(x) W(x) dx.$

Chu (1995) proposed the selection of the smoothing parameter via bootstrapping. Sarda (1993) proposed a leave-one-out cross-validation procedure to select the smoothing parameter. However, Altman and Léger (1995) showed that the leave-one-out procedure is asymptotically equivalent to a leaving-none-out procedure. In addition, they showed that the expected value of the derivative, with respect to h , of the leave-none-out criterion is asymptotically positive, which suggests that the criterion is increasing and that for sufficiently large samples the *smallest* available bandwidth will always be selected. Consequently, Altman and Léger (1995) proposed a plug-in estimator using the asymptotic optimal bandwidth (2.60).

Normal scaled rule of thumb

In this section the normal scaled rule of thumb will be discussed with $W(x) = 1$ and $K(\cdot) = \Phi(\cdot)$. The unknown density function f is replaced with a normal reference with mean μ and variance σ^2 . Using the properties of the normal distribution as discussed in Section 1.2, (2l), we find

$$C_1 = \frac{1}{2\sqrt{\pi}}, \quad C_2 = \frac{1}{2\sqrt{\pi}\sigma}, \quad C_3 = \frac{1}{3\sqrt{3}}(2\pi)^{-1}\sigma^{-4} \quad \text{and} \quad \mu_2(k) = 1.$$

Replacing these quantities into (2.60) the normal scaled rule of thumb is given by

$$h_{NS} = (3\sqrt{3})^{1/3}\sigma n^{-1/3} = 1.7321\sigma n^{-1/3}, \quad (2.64)$$

where the robust scale estimator defined in (2.22) is utilized to estimate the scale parameter σ . Note that the normal scaled rule of thumb given in (2.64) based on the weighted discrepancy measure (2.57), is slightly larger than the normal scaled rule of thumb defined in (2.63) based on the unweighted discrepancy measure (2.56). Hence, smoother distribution function estimates can be expected with the former bandwidth. It is, however, a difficult task to compare these two bandwidths directly since they are based on different discrepancy measures.

The method of Altman & Léger

In this section the notation $w_1(\cdot)$, $w_2(\cdot)$ and $w_3(\cdot)$ will be used to denote possible different kernel functions, and g_1 , g_2 and g_3 will be the associated smoothing parameters. Altman and Léger (1995) proposed to estimate the unknown quantity $C_2 = \int [f(x)]^2 W(x) dx$ by

$$\hat{C}_2 = \frac{1}{n^2 g_1} \sum_{i=1}^n \sum_{j=1}^n w_1 \left(\frac{X_i - X_j}{g_1} \right) W(X_i),$$

and the quantity $C_3 = \int [f'(x)]^2 f(x)W(x)dx$ by

$$\hat{C}_3 = \frac{1}{n^3 g_2^4} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n w_2' \left(\frac{X_i - X_j}{g_2} \right) w_2' \left(\frac{X_i - X_k}{g_2} \right) W(X_i).$$

Substitution of these estimates into (2.60) results in the Altman & Léger plug-in estimator

$$h_{AL} = \left(\frac{2C_1 \hat{C}_2}{\mu_2(k)^2 \hat{C}_3} \right)^{1/3} n^{-1/3}.$$

In order to find the asymptotically optimal bandwidth, g_2 , Altman and Léger (1995) calculated the asymptotic mean squared error of \hat{C}_3 under the assumptions

1. The kernel function w_2 has mean 0, finite variance and $w_2'(0) = 0$.
2. The density f has a bounded fourth derivative.
3. The bandwidth g_2 is a non-random sequence of positive numbers, satisfying

$$\lim_{n \rightarrow \infty} g_2 = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} n g_2 = \infty.$$

Hence, g_2 can be of the form $g_2 = cn^{-t}$, where $0 < t < 1$ and c is a positive constant.

The asymptotic mean squared error of \hat{C}_3 is given by

$$\begin{aligned} AMSE[\hat{C}_3] &= \frac{2C(w_2)}{n^2 g_2^5} \int [f(x)]^4 W(x) dx \\ &+ \left[g_2^2 \mu_2(w_2) \int f'(x) f'''(x) f(x) W(x) dx + \frac{R(w_2')}{n^2 g_2^3} \int [f(x)]^2 W(x) dx \right]^2, \end{aligned} \quad (2.65)$$

where $C(w_2) = \iiint w_2'(s) w_2'(t) w_2'(u) w_2'(t+u-s) ds dt du$. From (2.65) it easily follows that the asymptotically optimal bandwidth is given by:

$$g_2 = \left[\frac{5C(w_2) \int [f(x)]^4 W(x) dx}{2\mu_2(w_2)^2 \Psi_{1,3}^2} \right]^{1/9} n^{-2/9}, \quad (2.66)$$

where $\Psi_{m,p} = \int f^{(m)}(x) f^{(p)}(x) f(x) W(x) dx$. The effect of the AMSE optimal bandwidth given in (2.66) on the AMSE presented in (2.65) can be measured by replacing (2.66) into (2.65). This shows that the AMSE converges to zero at the rate of $O(n^{-8/9})$. Although not discussed by Altman and Léger (1995), one is tempted to set the weight function $W(x) = 1$. With this choice, a number of simplifications and improvements for the proposed method is possible. These are summarized as follows:

- The unknown quantity $C_2 = \int [f(x)]^2 dx$ may be estimated using the l -stage estimator as suggested by Hall and Marron (1987) and Sheather and Jones (1991), i.e., $\hat{\psi}_0(g_1)$; see (2.23), (2.24), (2.43) and (2.44).
- The unknown quantity $C_3 = \int [f'(x)]^2 f(x) dx$ may also be estimated using an l -stage estimator, with a normal reference utilized at stage l . Note that a general estimator for the quantity $\Psi_{m,p} = \int f^{(m)}(x) f^{(p)}(x) f(x) dx$ is given by

$$\hat{\Psi}_{m,p} = \frac{1}{n^3 g_2^{m+1} g_3^{p+1}} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n w_2^{(m)} \left(\frac{X_i - X_j}{g_2} \right) w_3^{(p)} \left(\frac{X_i - X_k}{g_3} \right),$$

where $w_2(\cdot)$ and $w_3(\cdot)$ are possibly different kernel functions with different associated bandwidths g_2 and g_3 respectively.

The $l = 0$ stage procedure results in the normal scaled rule of thumb presented in (2.64). Next, the procedure for $l = 1$, $w_1(\cdot) = w_2(\cdot) = w_3(\cdot) = \phi(\cdot)$, $K(\cdot) = \Phi(\cdot)$, $W(x) = 1$ and a normal reference with mean μ and variance σ^2 for the unknown density f will be described:

Step 1 Estimate ψ_2 using the normal reference, thus

$$\hat{\psi}_2 = \frac{-1}{4\sqrt{\pi}\hat{\sigma}^3}.$$

[using (2.44), and $\hat{\sigma}$ as defined in (2.22)]

Step 2 Use $\hat{\psi}_2$ to estimate $C_2 = \psi_0 = R(f)$, thus

$$\hat{C}_2 = \hat{\psi}_0 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \phi_{\hat{g}_1}(X_i - X_j).$$

In the expression above \hat{g}_1 is obtained through direct application of (2.43), thus

$$\hat{g}_1 = \left[\frac{-2\phi(0)}{\mu_2(\phi)\hat{\psi}_2} \right]^{1/3} n^{-1/3}, \quad \text{where } \mu_2(\phi) = 1 \text{ and } \phi(0) = \frac{1}{\sqrt{2\pi}}.$$

Step 3 Estimate $C_3 = \Psi_{1,1} = \int [f'(x)]^2 f(x) dx$ by

$$\hat{C}_3 = \hat{\Psi}_{1,1} = \frac{1}{n^3 \hat{g}_2^4} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \phi' \left(\frac{X_i - X_j}{\hat{g}_2} \right) \phi' \left(\frac{X_i - X_k}{\hat{g}_2} \right),$$

where, using the normal reference and (2.66)

$$\hat{g}_2 = \frac{1}{2}[405]^{1/9} \hat{\sigma} n^{-2/9} = 0.9743 \hat{\sigma} n^{-2/9}.$$

[see Section 1.2, 2(1)ix, 2(1)xi and 2(1)xii]

Step 4 Use \hat{C}_2 and \hat{C}_3 to calculate the WAMISE optimal bandwidth, h , (direct plug-in), thus

$$\hat{h}_{DPI,1} = \left(\frac{2C_1 \hat{C}_2}{\mu_2(k)^2 \hat{C}_3} \right)^{1/3} n^{-1/3}, \quad \text{where } \mu_2(\phi) = 1 \text{ and } C_1 = \frac{1}{2\sqrt{\pi}}.$$

[see Section 1.2, 2(1)i and 2(1)xii]

Conclusion

In this section we presented the bandwidth selection procedure proposed by Polansky (1997), based on the discrepancy measure (2.56) and presented a slight alteration to the procedure proposed by Altman and Léger (1995), based on the weighted discrepancy measure (2.57). We conclude that the latter is based on a more sound discrepancy measure and should be favored for general application. However, in the context of the transformation kernel density estimator a parametric pre-transformation will be employed for the newly proposed semi-parametric transformation to normality discussed in Chapter 3. Hence, the data used for distribution function estimation, will be approximately normally distributed and consequently adhere to the assumptions necessary in setting up the discrepancy measure (2.56). For this reason we advocate the use of the method proposed by Polansky (1997), since this method is faster and simpler to implement.

3

Transformation of data

In this chapter transformation of data to normality will be considered. The relationship between qq-plots and an optimal transformation will be discussed in Section 3.1. Section 3.3 can be viewed as a general introduction to *parametric transformations*. New contributions include:

- Establishing a new transformation to any distribution (Section 3.2).
- Using polynomials to approximate transformation functions (Section 3.2).
- Introducing two alternative parametric transformation parameter estimation techniques (Section 3.3.3).
- Defining a new optimal semi-parametric transformation to normality (Section 3.4).

3.1 QQ-plots: key to transformations

To understand the logic behind transformations one must understand qq-plots. This section is devoted to this topic and explains the relationship between qq-plots and transformations. Let X_1, \dots, X_n be i.i.d. random variables from the continuous distribution function F_X , with the associated order statistics $X_{(1)}, \dots, X_{(n)}$. Also, let U_1, \dots, U_n be i.i.d. uniformly distributed random variables on $[0, 1]$, with the associated order statistics

$U_{(1)}, \dots, U_{(n)}$. Suppose that the distribution function of X , F_X , is unknown and of the form

$$F_X(t) = F_0\left(\frac{t - \mu_x}{\sigma_x}\right),$$

then F_X belongs to the class of location-scale distributions, where F_0 is a standardized distribution function. With F_X unknown, one can speculate as to what the true distribution function is. Suppose that the hypothesized distribution is G_X , then, a qq-plot can be used to test this claim graphically. The qq-plot is constructed using the hypothesis $H_0 : F_X(t) = G_X(t)$. Using the location-scale definition, this is equivalent to

$$H_0 : F_X(t) = G_0\left(\frac{t - \mu_x}{\sigma_x}\right).$$

If the null hypothesis is correct and since both $F_X(\cdot)$ and $G_0(\cdot)$ are monotonic increasing functions and $\sigma_x > 0$ it follows that for $i = 1, \dots, n$,

$$\frac{X_{(i)} - \mu_x}{\sigma_x} = G_0^{-1}\left(F_X(X_{(i)})\right), \quad (3.1)$$

which is equivalent (in distribution) to

$$\frac{X_{(i)} - \mu_x}{\sigma_x} = G_0^{-1}\left(U_{(i)}\right). \quad (3.2)$$

Since the density of $U_{(i)}$ is given by

$$f_{U_{(i)}}(u) = n \binom{n-1}{i-1} u^{i-1} (1-u)^{n-i} \quad \text{for } 0 \leq u \leq 1,$$

it is easy to verify that

$$EU_{(i)} = \frac{i}{n+1}. \quad (3.3)$$

Taking expectations on both sides of the equation in (3.2), using (3.3) and a first-order Taylor expansion about $EU_{(i)}$ we have,

$$\begin{aligned} E\left(\frac{X_{(i)} - \mu_x}{\sigma_x}\right) &= EG_0^{-1}(U_{(i)}) \\ &\approx E\left\{G_0^{-1}(EU_{(i)}) + (U_{(i)} - EU_{(i)}) \left[\frac{d}{du}G_0^{-1}(u)\right]_{EU_{(i)}}}\right\} \\ &= G_0^{-1}(EU_{(i)}) = G_0^{-1}\left(\frac{i}{n+1}\right). \end{aligned} \quad (3.4)$$

The meaning of (3.4) is twofold:

- If the null model is correct we can expect a straight line with intercept 0 and gradient 1 when plotting

$$Z_{(i)} = \frac{X_{(i)} - \hat{\mu}_x}{\hat{\sigma}_x} \text{ vs. } G_0^{-1}\left(\frac{i}{n+1}\right), \quad i = 1, \dots, n,$$

where $\hat{\mu}_x$ and $\hat{\sigma}_x$ are estimates of μ_x and σ_x respectively, based on the data X_1, \dots, X_n .

- If the null model is correct we can expect a straight line with intercept $\hat{\mu}_x$ and gradient $\hat{\sigma}_x$ when plotting

$$X_{(i)} \text{ vs. } G_0^{-1}\left(\frac{i}{n+1}\right), \quad i = 1, \dots, n.$$

The null model implied by the qq-plot representation is therefore given by

$$X_{(i)} \approx \hat{\mu}_x + \hat{\sigma}_x G_0^{-1}\left(\frac{i}{n+1}\right). \quad (3.5)$$

Note that the convention $i/(n+1)$ is not unique and is sometimes replaced with $(i-0.5)/n$. This is equivalent to using the kernel distribution function estimator, (2.50), with the bandwidth, h equal to zero, i.e., $\hat{F}_X(x; 0)$. For a discussion on this relationship the reader is referred to Section 2.2, (2.51). The qq-plot used in this dissertation is to plot $Z_{(i)}$ vs. $G_0^{-1}\left(\frac{i-0.5}{n}\right)$. The reason for using the standardized order statistics is that this plot allows for the creation of convex to convex, concave to concave, convex to concave and concave to convex transformations (see Section 3.3.2 for more detail on this), since the curvature of the parametric transformations considered can change form about zero.

Example 1: In the rest of this section, for illustration purposes, 50 data points drawn from a standard lognormal distribution, $X \sim \text{Logn}(0, 1) \equiv F_X$, will be considered. The data is then standardized, i.e., $Z = (X - \hat{\mu}_x)/\hat{\sigma}_x \sim F_Z$, for reasons explained above. We know that the ideal transformation to transform standard lognormal data to normality is given by $y = \ln(x)$ and that the optimal back transformation is $x = e^y$. The curvature of these optimal transformations can be used as a reference for comparison with alternative transformations. The aim is therefore to transform the input data to normality, i.e., $X \rightarrow Z \rightarrow Y \sim G_0(\cdot) \equiv \Phi(\cdot)$. Subsequently, we will show how to obtain this transformation.

What is the relationship between qq-plots and the ideal transformation?

To explain this, consider the data from Example 1, and argue that we want to transform this data so that the transformed data will have a normal distribution. In this case $G_0(\cdot) = \Phi(\cdot)$. The curvature of the transformation function required to transform this data to normality should be similar to the curvature of the qq-plot. It should be noted that the traditional qq-plot cannot be used to model the ideal transformation, since the dependent variable is given by $G_0^{-1}\left(\frac{i}{n+1}\right)$, which is non-stochastic. For this reason we need to replace the traditional qq-plot with an “ideal” qq-plot. The ideal qq-plot will be defined as a plot of

$$Z_{(i)} \text{ vs. } G_0^{-1}\left(U_{(i)}\right), \quad i = 1, \dots, n.$$

The ideal transformation function, henceforth denoted by g , is defined by

$$g(t) = G_0^{-1}\left(F_Z(t)\right), \quad (3.6)$$

where $F_Z(t)$ is (as before) the distribution function of $Z = (X - \hat{\mu}_x) / \hat{\sigma}_x$. The transformation function g is called “ideal” because if we set

$$\tilde{Y}_{(i)} = g\left(Z_{(i)}\right), \quad (3.7)$$

then

$$\begin{aligned} \tilde{Y}_{(i)} &= G_0^{-1}\left(F_Z\left(Z_{(i)}\right)\right) \\ &\stackrel{d}{=} G_0^{-1}\left(U_{(i)}\right), \end{aligned}$$

which implies that $\tilde{Y}_{(i)}$ is distributed as the i^{th} order statistic of a random sample from G_0 .

In view of the discussion given above, we will show in the section below how to derive an estimated ideal transformation function, say \hat{g} , henceforth called the optimal transformation function. This \hat{g} can therefore be applied practically to transform any given data set to a set of data having distribution function G_0 .

3.2 A new transformation to any distribution**3.2.1 The transformation**

In this section a new method of transformation will be described that has the advantage that it can assume convex to convex, concave to concave, convex to concave, concave

to convex and **any** combination of these forms into one transformation function (these shapes as well as where they are applicable are discussed in Section 3.3.2). Traditional transformation methods, for example the Box-Cox transformation, are not armed with this flexibility. This property improves the p-value for a goodness-of-fit test dramatically. The proposed transformation can also be iterated to improve the p-value after each iteration. We impose the following restrictions that should be satisfied by the newly proposed transformation:

1. The transformation should transform the data with a high p-value when testing the goodness-of-fit of the transformed data.
2. The transformation should be a monotonic increasing function. This will ensure the existence of the back transformation, more specific a one-to-one mapping from input data to transformed data.
3. The transformation should (preferably) be written in a simple mathematical form.

It should be noted that the ideal transformation function g , given in expression (3.6), cannot be applied in practice, since F_Z is (under the alternative hypothesis) unknown. We therefore suggest estimating g by

$$\hat{g}(t) = G_0^{-1}(\hat{F}_Z(t)), \quad (3.8)$$

where $\hat{F}_Z(\cdot)$ is an appropriate estimate of $F_Z(\cdot)$. Henceforth, \hat{g} will be referred to as the optimal transformation function. The transformed data will be defined as (see (3.7))

$$Y_{(i)} = \hat{g}(Z_{(i)}), \quad i = 1, \dots, n. \quad (3.9)$$

For the data from Example 1, the estimated ideal qq-plot and optimal transformation function are shown in Figure 3.1.

Properties of the transformation (3.8) are summarized in the following list:

- Since $G_0^{-1}(\cdot)$ is a continuous function the transformation $\hat{g}(\cdot)$ will also be continuous provided that $\hat{F}_Z(\cdot)$ is continuous.
- Furthermore, the transformation $\hat{g}(\cdot)$ will be a monotonic increasing function provided that the estimate, $\hat{F}_Z(\cdot)$, is monotonic increasing.

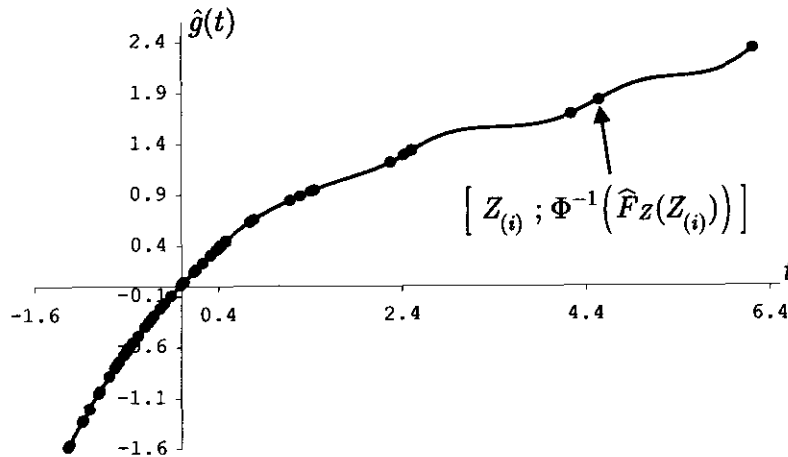


Figure 3.1: *Estimated ideal qq-plot, showing the optimal transformation function to transform standardized standard lognormal data ($n = 50$) to normality.*

- Perhaps the most important property of this transformation is that $\hat{g}(\cdot)$ can take on any monotonic increasing form. The transformation function can therefore change form from convex to convex, concave to concave, concave to convex, convex to concave or any combination of the above, depending on the shape of $\hat{F}_Z(\cdot)$.
- The transformation can also be iterated to improve the p-value when testing the goodness of fit. This is discussed in Section 3.4.

In order to apply $\hat{g}(\cdot)$, defined in (3.8), we need to define a suitable estimator $\hat{F}_Z(\cdot)$. Three such estimators will now be discussed.

One method is to set up a wide and exhaustive range of possible distribution functions, then perform goodness-of-fit tests to find a suitable probability model for Z . Let's call this the **goodness-of-fit approach**. After the initial model selection we need to estimate, using standard estimation techniques (e.g. maximum likelihood), the unknown parameter(s) θ of the model selected. The problems here are, firstly to find a suitable set of candidate distributions, secondly, to find a suitable goodness-of-fit test. Should we use a standard χ^2 goodness-of-fit test or should we incorporate different techniques associated with each distribution in our set of distributions? A simple well-known example here is the Shapiro and Wilk (1965) test for normality. One feels that the goodness-of-fit

approach should be implemented by a specialist (or in supervision of) in goodness-of-fit techniques, since there are so many gray areas. Nevertheless, once the probability model is decided upon, the optimal transformation in (3.8) is given by

$$\hat{g}(t) = G_0^{-1} \left(F_{Z, \hat{\theta}}(t) \right).$$

A second approach is to replace $\hat{F}_Z(\cdot)$ by the empirical distribution function

$$F_{n,Z}(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}(Z_i \leq t).$$

Let's call this the **empirical approach**. At first glance $F_{n,Z}(t)$ appears to be a good candidate in the current context since, $EF_{n,Z}(t) = \int_{-\infty}^{+\infty} \mathbf{I}(x \leq t) dF_Z(x) = F_Z(t)$, thus we have an unbiased estimator. However, after careful inspection of the transformation implied by this estimate, viz.

$$\hat{g}(t) = G_0^{-1} (F_{n,Z}(t)),$$

it is clear that the empirical approach has poor properties in the current context. These properties are summarized in the following list:

- $F_{n,Z}(t)$ is a discrete step function, hence not strictly monotonic. This implies that $\hat{g}(t)$ is also a discrete step function.
- For any $t < \min(Z_1, \dots, Z_n)$ we have that $F_{n,Z}(t) = 0$ and also for any $t \geq \max(Z_1, \dots, Z_n)$ we have that $F_{n,Z}(t) = 1$. Note that $G_0^{-1}(p)$ approaches $+\infty$ and $-\infty$ if $G_0(\cdot) = \Phi(\cdot)$ and $p \rightarrow 1$ and $p \rightarrow 0$ respectively. Hence, the range of $\hat{g}(t)$ is a bounded interval, determined by the values of $Z_{(1)}$ and $Z_{(n)}$.

It should be clear that we can improve the procedure by replacing the empirical estimate $F_{n,Z}(t)$ by a smooth version such as the adaptive kernel distribution function estimator

$$\hat{F}_Z(t; h) = \frac{1}{n} \sum_{i=1}^n K \left(\frac{t - Z_i}{h\lambda_i} \right),$$

where K is the so-called kernel function, h is the bandwidth and λ_i is a scaling factor. Let's call this the **kernel approach**. For the kernel distribution function estimator it can be shown that $E\hat{F}_Z(t; h) = F_Z(t) + O(h^2)$. The reader is referred to Section 2.2.1 for more detail. We use the adaptive kernel distribution function estimate here instead of the traditional kernel estimate since $G_0^{-1}(p)$ approaches $+\infty$ and $-\infty$ if $G_0(\cdot) = \Phi(\cdot)$

and $p \rightarrow 1$ and $p \rightarrow 0$ respectively. The result is that artificial outliers are created in the transformed data. This phenomenon has a negative effect on the estimation of densities when the transformation kernel density estimator (2.48) is applied. For this reason we developed the adaptive kernel distribution function estimator. The reader is referred to Section 4.2 for a more detailed discussion on the choice of λ_i . The smoothing parameter h is chosen according to the methods described in Section 2.2.3. With this kernel approach the optimal transformation function (using (3.8)) is given by

$$\begin{aligned}\hat{g}(t) &= G_0^{-1}(\hat{F}_Z(t; \hat{h})) \\ &= G_0^{-1}\left[\frac{1}{n} \sum_{i=1}^n K\left(\frac{t - Z_i}{\hat{h}\lambda_i}\right)\right].\end{aligned}\quad (3.10)$$

The transformation procedures in (3.8) and (3.9) can be iterated to improve the p-value. This is discussed further in Section 3.4. Using the data from Example 1, the optimal transformation function, using $\lambda_i = 1 \quad \forall i$, is shown in Figure 3.1.

Sometimes one is interested not only in the forward transformation but also in the backward transformation. In the latter case we have (compare with (3.8))

$$\hat{g}^{-1}(t) = \hat{F}_Z^{-1}(G_0(t)), \quad (3.11)$$

so that, if evaluated in the transformed data (see (3.9)), we obtain the original data

$$Z_{(i)} = \hat{g}^{-1}(Y_{(i)}) = \hat{F}_Z^{-1}(G_0(Y_{(i)})). \quad (3.12)$$

It should be noted that quantile estimation is required to implement the optimal backward transformation. For an excellent account of kernel quantile estimation procedures, the reader is referred to Sheather and Marron (1990). The backward transformation and corresponding qq-plot are displayed in Figure 3.2.

3.2.2 Polynomial approximation of the optimal transformation function

Sometimes it is convenient to approximate the optimal transformation function $\hat{g}(t)$, defined in (3.8), by a suitable polynomial, viz.

$$\hat{g}(t) \approx \hat{b}_0 + \hat{b}_1 t + \hat{b}_2 t^2 + \cdots + \hat{b}_p t^p,$$

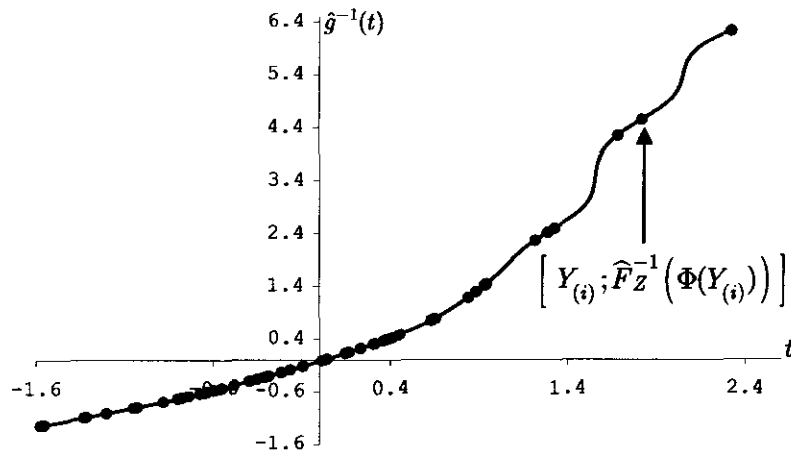


Figure 3.2: Backward transformation for the data of Example 1.

where p is the order and $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_p$ are estimated coefficients. In the estimation procedures presented below t_1, \dots, t_m are equally-spaced grid points and $G_0^{-1}(\hat{F}_Z(t_i; h))$, $i = 1, \dots, m$ are the corresponding optimal transformation values. For each fixed p , the estimates $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_p$ can, for example, be derived by using the least-squares estimation method. The degree p of the polynomial can then be determined by minimizing the AIC_C criterion with respect to p , which is given by

$$AIC_C = \log \hat{\sigma}^2 + \frac{2(p+1)}{m-p-2},$$

where

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^m \left[G_0^{-1}(\hat{F}_Z(t_i; h)) - (\hat{b}_0 + \hat{b}_1 t_i + \dots + \hat{b}_p t_i^p) \right]^2.$$

The interested reader is referred to Hurvich and Simonoff (1998) for a discussion of this and related criteria. Alternatively, the degree of the polynomial may be selected by fitting the lowest degree polynomial achieving a specified R^2 value. Two useful Fortran subroutines are "RPOLY" and "RCURV". For the data from Example 1 the polynomial approximation of $\hat{g}(t)$ is shown in Figure 3.3 and is given by

$$y = 0.0024254 + 0.99306t - 0.1874t^2 + \dots + (3.833 \times 10^{-8})t^{17},$$

where

$$t = \frac{x - 1.25972}{0.94842}.$$

Comparing Figure 3.3 with Figure 3.1, shows that the polynomial approximation of $\hat{g}(t)$ is exceedingly well.

From (3.10) it is clear that a successful transformation depends solely on the efficacy of the

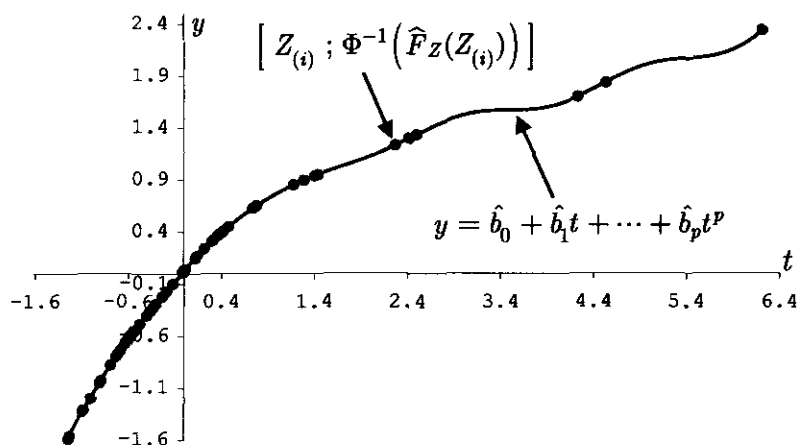


Figure 3.3: *Polynomial approximation.*

adaptive kernel distribution function estimator \hat{F}_Z . To improve the performance of this estimator a parametric pre-transformation on Z is implemented. This technique together with a motivation of the chosen transformation distribution, $\Phi(\cdot)$, will be discussed next.

3.3 Parametric transformations

3.3.1 Overview

It can be shown, Terrell (1990), that the beta(4,4) density is the easiest to estimate using kernel density estimation (see Section 2.1.2 for a more detailed discussion and the definition of the efficiency measure). The efficiency of the kernel estimator for estimating this density is 1 (1 = easiest to estimate, 0 = most difficult to estimate using kernel density estimation). Since the normal density is similar to the beta(4, 4) density in shape and is also easy to estimate (efficiency measure = 0.908), we decided to transform all random variables, for which a density estimate is required, to the normal distribution, primarily because of ease of use. For this reason the following paragraphs will be devoted to transformation to *normality*.

Next, a short summary of possible parametric transformations will be presented. Tukey

(1957) introduced a family of power transformations defined by

$$g_\lambda(x) = \begin{cases} x^\lambda & \lambda \neq 0, \\ \log(x) & \lambda = 0, \end{cases} \quad (3.13)$$

for $x > 0$. This family contains a discontinuity at $\lambda = 0$. Box and Cox (1964) modified this transformation to take account of the discontinuity at $\lambda = 0$ and defined

$$g_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \log(x) & \lambda = 0, \end{cases} \quad (3.14)$$

for $x > 0$. Box and Cox (1964) also proposed the shifted power transformation defined by

$$g_\lambda(x) = \begin{cases} \frac{(x + \lambda_1)^{\lambda_2} - 1}{\lambda_2} & \lambda_2 \neq 0, \\ \log(x + \lambda_1) & \lambda_2 = 0. \end{cases} \quad (3.15)$$

This transformation can be applied to negative values of x . λ_2 is called the transformation parameter and λ_1 is chosen such that $\lambda_1 > -\min\{X_1, \dots, X_n\}$. In the context of transformation kernel density estimation, Wand, Marron and Ruppert (1991) considered an alternative version of this transformation defined by

$$g_\lambda(x) = \begin{cases} \text{sign}(\lambda_2)(x + \lambda_1)^{\lambda_2} & \lambda_2 \neq 0, \\ \log(x + \lambda_1) & \lambda_2 = 0. \end{cases}$$

Manley (1976) defined the following transformation

$$g_\lambda(x) = \begin{cases} \frac{e^{\lambda x} - 1}{\lambda} & \lambda \neq 0, \\ x & \lambda = 0. \end{cases} \quad (3.16)$$

This transformation can be applied to negative values of x and is effective in turning skew unimodal distributions into nearly symmetric normal-like distributions.

John and Draper (1980) defined the modulus transformation, which is considered to normalize distributions already possessing some measure of approximate symmetry. This transformation is given by

$$g_\lambda(x) = \begin{cases} \frac{\text{sign}(x) \{(|x| + 1)^\lambda - 1\}}{\lambda} & \lambda \neq 0, \\ \text{sign}(x) \{\log(|x| + 1)\} & \lambda = 0. \end{cases} \quad (3.17)$$

Bickel and Doksum (1981) suggested the following transformation

$$g_\lambda(x) = \frac{\text{sign}(x)|x|^\lambda - 1}{\lambda}, \quad (3.18)$$

with $\lambda > 0$. This transformation is designed to handle kurtosis rather than skewness.

Burdige, Magee and Robb (1988) compared the inverse hyperbolic sine transformation proposed by Johnson (1949) (a member of a transformation family) with the transformation proposed by Bickel and Doksum (1981), in a study to reduce the influence of the extreme observations. This transformation is defined over the whole real line and given by

$$\begin{aligned} g_\theta(x) &= \log(\theta x + (\theta^2 x^2 + 1)^{1/2})/\theta \\ &= \sinh^{-1}(\theta x)/\theta. \end{aligned}$$

They concluded that when the input data can take on the value 0, an immediate advantage of the inverse hyperbolic sine transformation is apparent, since the likelihood function of the Bickel and Doksum (1981) transformation will not be defined when such observations occur.

Yang and Marron (1999) reparameterized the Johnson (1949) system of transformations. This family is a versatile set of transformations since they can assume numerous forms, and they are able to address both kurtosis and skewness. The family of transformations is given by

$$g_{\gamma,\lambda}(x) = \begin{cases} \log(1 + cJx)/cJ & 0 < \lambda \leq \lambda_M, \gamma = 1, \\ \log(cx + (c^2x^2 + 1)^{1/2})/c & 0 < \lambda \leq \lambda_M, \gamma = 2, \\ \log\left(\frac{1+cx}{1-cx}\right)/(2c) & 0 < \lambda \leq \lambda_M, \gamma = 3, \\ x & \lambda = 0, \quad \gamma = 1, 2, 3, \end{cases} \quad (3.19)$$

where $J = \pm 1$, $c = \frac{\lambda^p}{1-\lambda^p} \in [0, \infty)$, $\lambda_M \in (0, 1)$ and $p \geq \frac{1}{2}$ is a tuning constant which is needed for smoothness of $g_{\gamma,\lambda}(x)$ as a function of λ . For this transformation to be valid there must be some restrictions on c . These restrictions are summarized in Table 3.1.

Table 3.1: Restrictions on c for the Johnson transformation.

Parameter Choices	Range of the input data	Restriction
$\gamma = 1, J = +1$	$x \in (0, \infty)$	$c > 0$
	$x \in (-\infty, 0)$	$0 < c < -(\min(x))^{-1}$
	$x \in (-\infty, \infty)$	$0 < c < -(\min(x))^{-1}$
$\gamma = 1, J = -1$	$x \in (0, \infty)$	$0 < c < (\max(x))^{-1}$
	$x \in (-\infty, 0)$	$c > 0$
	$x \in (-\infty, \infty)$	$0 < c < (\max(x))^{-1}$
$\gamma = 2$	$x \in \mathbf{R}$	$c > 0$
$\gamma = 3$	$x \in (0, \infty)$	$0 < c < (\max(x))^{-1}$
	$x \in (-\infty, 0)$	$0 < c < -(\min(x))^{-1}$
	$x \in (-\infty, \infty)$	$0 < c < \min \left[\frac{-1}{\min(x)}, \frac{1}{\max(x)} \right]$

Yeo and Johnson (2000) introduced a new power transformation:

$$g_\lambda(x) = \begin{cases} \frac{(x+1)^\lambda - 1}{\lambda} & (x \geq 0, \lambda \neq 0), \\ \log(x+1) & (x \geq 0, \lambda = 0), \\ \frac{-\{(-x+1)^{2-\lambda} - 1\}}{(2-\lambda)} & (x < 0, \lambda \neq 2), \\ -\log(-x+1) & (x < 0, \lambda = 2). \end{cases} \quad (3.20)$$

This transformation is well-defined on the whole real line and is appropriate for reducing skewness and to approximate normality.

In the context of transformation kernel density estimation, (2.48), Ruppert and Wand (1992) introduced the following kurtosis reducing transformation

$$g_\alpha(x) = \alpha x + (1 - \alpha) \hat{\sigma}_x \sqrt{2\pi} \left\{ \Phi \left(\frac{x}{\hat{\sigma}_x} \right) - \frac{1}{2} \right\},$$

for $0 \leq \alpha \leq 1$, where $\Phi(\cdot)$ is the standard normal distribution function.

3.3.2 Transformation curvature

To fully understand how these parametric transformations work and to see where they could be applied one needs to understand the curvature of these transformations. Orig-

nally, attempts were made to transform data repeatedly using a Box-Cox transformation. This failed drastically. To understand why, consider once again the kernel qq-plot for the data from example 1 of Section 3.1. The output given in Figure 3.1, suggests a transformation that is concave in form. This is exactly what is produced by the shifted Box-Cox transformation with $\lambda_2 < 1$. The p-value of the transformed data for example 1, utilizing the Shapiro-Wilk test for normality is 0.86914. Although successful, one is tempted to repeat this Box-Cox transformation, i.e., to transform the transformed data again to improve the p-value for normality. This appears to be a novel idea, but it is doomed to failure.

The reason is that the Box-Cox transformation family can only model transformations which assume shapes that are convex or concave. This is rather limited considering the huge number of probability models from which the data might be drawn. A convex or concave transformation model cannot be used where the qq-plot suggests a transformation that changes shape from convex to concave or vice versa (e.g. data from a skewed unimodal or kurtotic unimodal density), not even to mention suggested transformations that must change shape more than once. The Box-Cox transformation can therefore not be used without first considering the density of the input data. This is where kernel density estimation can come in handy. Since the curvature of the qq-plot determines the curvature of the transformation an inspection of the available transformations is needed. Hence, convex, concave, concave to convex and convex to concave transformations will be inspected next. Note that the newly proposed optimal transformation can take any form and from there its versatility.

Convex or concave transformations

Transformations of this class can take the form described in Figure 3.4.

Where should these transformations be applied?

Convex

This type of transformation is applied to data that are skewed to the left, for example, data from a skewed unimodal distribution (see Section 5.1 for a graph). The left tail of this transformation will contract the left tail of the input data whilst the right tail of the transformation will protract the right tail of the input data.

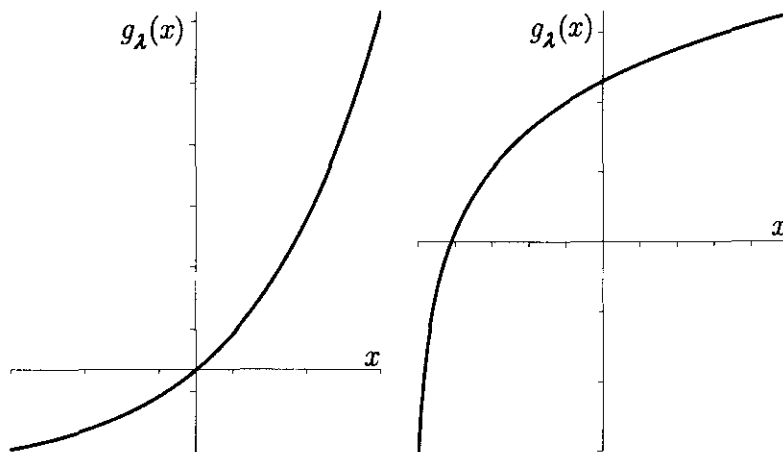


Figure 3.4: *Left panel: Convex transformation.*
Right panel: Concave transformation.

Concave

This type of transformation is applied to data that are skewed to the right, for example, lognormal data. The left tail of this transformation will protract the left tail of the input data whilst the right tail of the transformation will contract the right tail of the input data.

Which parametric transformations are convex or concave?

To inspect the curvature of the different transformations it is useful to inspect the second derivative of these transformations. From the transformations presented earlier, those having a convex or concave shape are summarized in Table 3.2. It is clear that the shape of the transformation is determined by the parameters of the transformation. The first derivative is also presented, since it is needed when applying the transformation kernel density estimate and setting up the profile log-likelihood function for estimating the transformation parameters. Note that the power transformation renders a curvature that is monotonic decreasing for $\lambda < 0$. This is unacceptable when estimating densities using the transformation kernel density estimator, since this will result in negative density estimates. The Johnson (1949) family of transformations is the only family that can assume a number of different shapes (see Tables 3.2 and 3.3).

Transformations that do not change shape are therefore summarized in the following

list:

- Tukey's power transformation,
- Box-Cox one parameter transformation,
- Shifted Box-Cox transformation,
- Manley's transformation,
- Yeo and Johnson's transformation,
- Johnson (1949) transformation with $\gamma = 1$.

Convex to concave or concave to convex transformations

Transformations of this class can take the shape displayed in Figure 3.5. Note that these transformations change shape around the origin. Therefore, to make full use of the change in curvature it is recommended to use these transformations for input data that can assume both negative and positive values. If, however, the input data can assume only negative or only positive values, this class of transformations reduces to convex or concave transformations.

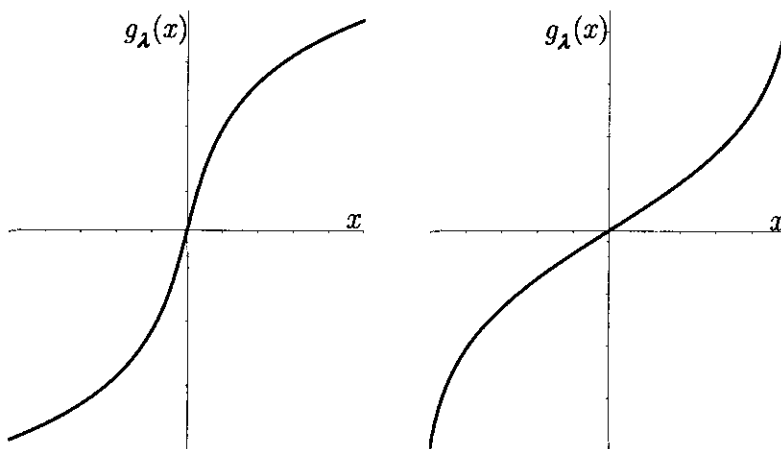
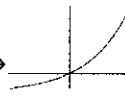
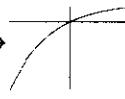
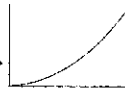
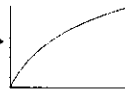
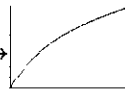
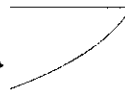
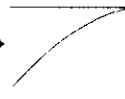
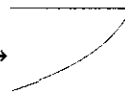

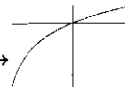


Figure 3.5: *Left panel: Convex to concave transformation.*
Right panel: Concave to convex transformation.

Table 3.2: Summary of the convex and concave transformations.

Transformation	$g'_\lambda(x)$	$g''_\lambda(x)$	Curvature
Power transformation			
$\lambda \neq 0$	$\lambda x^{\lambda-1}$	$\lambda(\lambda-1)x^{\lambda-2}$	$\left\{ \begin{array}{l} \text{convex if } \lambda < 0 \forall x \in (0, \infty) \rightarrow \text{graph} \\ \text{convex if } \lambda > 1 \forall x \in (0, \infty) \rightarrow \text{graph} \\ \text{concave if } 0 < \lambda < 1 \forall x \in (0, \infty) \rightarrow \text{graph} \end{array} \right.$
$\lambda = 0$	x^{-1}	$-x^{-2}$	concave $\forall x \in (0, \infty) \rightarrow \text{graph}$
Box-Cox 1 parameter			
$\lambda \neq 0$	$x^{\lambda-1}$	$(\lambda-1)x^{\lambda-2}$	$\left\{ \begin{array}{l} \text{convex if } \lambda > 1 \forall x \in (0, \infty) \rightarrow \text{graph} \\ \text{concave if } \lambda < 1 \forall x \in (0, \infty) \rightarrow \text{graph} \end{array} \right.$
$\lambda = 0$	x^{-1}	$-x^{-2}$	concave $\forall x \in (0, \infty) \rightarrow \text{graph}$
Shifted Box-Cox			
$\lambda_2 \neq 0$	$(x + \lambda_1)^{\lambda_2-1}$	$(\lambda_2 - 1)(x + \lambda_1)^{\lambda_2-2}$	$\left\{ \begin{array}{l} \text{convex if } \lambda_2 > 1 \rightarrow \text{graph} \\ \text{concave if } \lambda_2 < 1 \rightarrow \text{graph} \end{array} \right.$
$\lambda_2 = 0$	$(x + \lambda_1)^{-1}$	$-(x + \lambda_1)^{-2}$	concave $\forall x \in \mathbf{R} \rightarrow \text{graph}$

Transformation	$g'_\lambda(x)$	$g''_\lambda(x)$	Curvature
Manley	$e^{\lambda x}$	$\lambda e^{\lambda x}$	convex if $\lambda > 0 \forall x \in \mathbf{R} \rightarrow$  concave if $\lambda < 0 \forall x \in \mathbf{R} \rightarrow$ 
Yeo & Johnson			
$x \geq 0, \lambda \neq 0$	$(x+1)^{\lambda-1}$	$(\lambda-1)(x+1)^{\lambda-2}$	$\left\{ \begin{array}{l} \text{convex if } \lambda > 1 \rightarrow \end{array} \right.$  $\left\{ \begin{array}{l} \text{concave if } \lambda < 1 \rightarrow \end{array} \right.$ 
$x \geq 0, \lambda = 0$	$(x+1)^{-1}$	$-(x+1)^{-2}$	concave \rightarrow 
$x < 0, \lambda \neq 2$	$(-x+1)^{1-\lambda}$	$-(1-\lambda)(-x+1)^{-\lambda}$	$\left\{ \begin{array}{l} \text{convex if } \lambda > 1 \rightarrow \end{array} \right.$  $\left\{ \begin{array}{l} \text{concave if } \lambda < 1 \rightarrow \end{array} \right.$ 
$x < 0, \lambda = 2$	$(-x+1)^{-1}$	$(-x+1)^{-2}$	convex \rightarrow 
Johnson			
$\gamma = 1$	$(1+cJx)^{-1}$	$-cJ(1+cJx)^{-2}$	convex if $J = -1 \forall x \in \mathbf{R} \rightarrow$  concave if $J = +1 \forall x \in \mathbf{R} \rightarrow$ 

Where should these transformations be applied?**Convex to concave**

This type of transformation is applied to data that have long tails to the left and right, for example, data from a kurtotic unimodal distribution (see Section 5.1 for a graph). The left tail of this transformation will contract the left tail of the input data whilst the right tail of the transformation will contract the right tail of the input data.

Concave to convex

This type of transformation is applied to data that have short tails to the left and right, for example, data from a uniform, $U(0, 1)$, distribution. The left tail of this transformation will protract the left tail of the input data whilst the right tail of the transformation will protract the right tail of the input data.

Which parametric transformations are convex to concave or concave to convex?

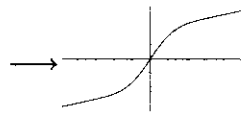
From the transformations presented earlier, those having a convex to concave or concave to convex shape are summarized in Table 3.3. Note that the Bickel and Doksum transformation has a serious drawback that renders it unusable in the context of transformation kernel density estimation. For this transformation we have that $g'_\lambda(0) = \infty$ for $0 < \lambda < 1$, this will result in a density estimate that is infinite in the point $x = 0$. Also, the derivative of the John & Draper transformation involves absolute values that imply a non-smooth derivative. For this reason this transformation is also not suitable for application in the context of transformation kernel density estimation.

Transformations that change shape around zero are therefore summarized in the following list:

- John and Draper's transformation,
- Bickel and Doksum's transformation,
- Johnson (1949) transformation with $\gamma = 2$ or $\gamma = 3$,
- Ruppert and Wand transformation.

Table 3.3: Summary of the convex to concave and concave to convex transformations.

Transformation	$g'_\lambda(x)$	$g''_\lambda(x)$	Curvature
John & Draper			
$\lambda \neq 0$	$(x + 1)^{\lambda-1}$	$\text{sign}(x)(\lambda - 1)(x + 1)^{\lambda-2}$	$\left\{ \begin{array}{l} \lambda > 1 \\ \lambda < 1 \end{array} \right. \begin{cases} \text{concave if } x < 0 \\ \text{convex if } x > 0 \end{cases} \rightarrow \begin{array}{c} \text{Graph 1} \\ \text{Graph 2} \end{array}$
$\lambda = 0$	$(x + 1)^{-1}$	$-\text{sign}(x)(x + 1)^{-2}$	$\left\{ \begin{array}{l} \text{convex if } x < 0 \\ \text{concave if } x > 0 \end{array} \right. \rightarrow \begin{array}{c} \text{Graph 3} \end{array}$
Bickel & Doksum			
	$ x ^{\lambda-1}$	$\text{sign}(x)(\lambda - 1) x ^{\lambda-2}$	$\left\{ \begin{array}{l} \lambda > 1 \\ 0 < \lambda < 1 \end{array} \right. \begin{cases} \text{concave if } x < 0 \\ \text{convex if } x > 0 \end{cases} \rightarrow \begin{array}{c} \text{Graph 4} \\ \text{Graph 5} \end{array}$
Johnson			
$\gamma = 2$	$(1 + c^2x^2)^{-\frac{1}{2}}$	$-c^2x(1 + c^2x^2)^{-\frac{3}{2}}$	$\left\{ \begin{array}{l} \text{convex if } x < 0 \\ \text{concave if } x > 0 \end{array} \right. \rightarrow \begin{array}{c} \text{Graph 6} \end{array}$
$\gamma = 3$	$(1 - c^2x^2)^{-1}$	$2c^2x(1 - c^2x^2)^{-2}$	$\left\{ \begin{array}{l} \text{concave if } x < 0 \\ \text{convex if } x > 0 \end{array} \right. \rightarrow \begin{array}{c} \text{Graph 7} \end{array}$

Transformation	$g'_\lambda(x)$	$g''_\lambda(x)$
Ruppert & Wand	$\alpha + (1 - \alpha)\sqrt{2\pi} \phi\left(\frac{x}{\hat{\sigma}_x}\right)$	$-(1 - \alpha)\frac{\sqrt{2\pi} x}{\hat{\sigma}_x^2} \phi\left(\frac{x}{\hat{\sigma}_x}\right)$
Curvature:	$\begin{cases} \text{convex if } x < 0 \\ \text{concave if } x > 0 \end{cases}$ 	

Black Box

To transform data successfully to normality a black box of possible pilot transformations is utilized. For this black box to be exhaustive in the number of shapes that the transformations can possess, the input data is standardized so that transformations that change shape will be included automatically. When standardizing data one must take into account the possible effect of outliers on the location and scale estimates. These estimates can also be greatly affected by data that are highly skewed to the left or right. For this reason one must turn to more robust methods for estimating the location and scale parameters. The method of standardizing data used in this dissertation is given by

$$Z = \frac{X - \hat{\mu}_x}{\hat{\sigma}_x},$$

where $\hat{\mu}_x$ is the sample median of the data X_1, \dots, X_n and

$$\hat{\sigma}_x = \min \left\{ s_x, (\hat{q}_3 - \hat{q}_1) / \left(\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right) \right) \right\},$$

where \hat{q}_1 and \hat{q}_3 are the first and third sample quartiles respectively and s_x^2 is the usual unbiased sample variance. *Note that this robust standardization will only be applied to the input data X , after which the usual sample mean will be employed for any subsequent standardizations.*

Hence, the mapping used in this dissertation will be as follows:

$$X \rightarrow Z \rightarrow Y,$$

where the linear standardization described above will be used to proceed from X to Z . Then one of the pilot transformations from the black box, defined below, will be applied

to transform the data from Z to Y . After these two initial transformations the newly proposed optimal transformation will be employed. *The black box of transformations decided upon is given in the following list:*

- The shifted Box and Cox (1964) transformation.
- The Yeo and Johnson (2000) transformation.
- The John and Draper (1980) transformation (not used in the context of transformation kernel density estimation).
- The Johnson transformation as given by Yang and Marron (1999).
- The Ruppert and Wand (1992) transformation.

Note that this black box contains three transformations that cannot change shape (Box-Cox, Yeo & Johnson and the Johnson transformation with $\gamma = 1$) and three transformations that change shape around zero (John & Draper, Johnson with $\gamma = 2, 3$ and the Ruppert & Wand transformation). Tukey's power transformation and the one parameter Box-Cox transformation are only valid for positive data and since the input data are standardized these transformations are not an option anymore. Manley's transformation was left out of the black box since enough transformations that cannot change shape are included. The transformation of Bickel & Doksum was omitted on account of the poor behaviour of its first derivative at $x = 0$.

3.3.3 Parameter estimation and transformation selection

In this section attention is given to the estimation of the parameters of the transformations nominated to be in the black box. The profile maximum likelihood method as well as two new methods, namely the minimum residual and minimum distance methods will be discussed. In addition, we will define the procedure to select an optimal transformation from the black box. It should be noted that this selection procedure is performed after, and independently of the parameter estimation procedures.

Markovitch and Krieger (2000) studied four methods of estimating the Box-Cox parameter used to transform data to normality. Although they only studied the one parameter Box-Cox transformation, the results are well worth mentioning. Three of the methods

were based on optimizing test statistics for normality. They are tests based on skewness, kurtosis and the Shapiro-Wilk test. The fourth method utilized the profile maximum likelihood procedure as described in Box and Cox (1964) and Atkinson (1985). Markovitch and Krieger (2000) found that the estimator based on the Shapiro and Wilk (1965) statistic generally gives rise to the best transformation, while the maximum likelihood estimator performs almost as well. They also concluded that estimators based on optimizing skewness and kurtosis do not perform well in general.

The Shapiro-Wilk test is considered by some to be the standard test of normality to date. In order to assess the performance of the different transformations in the black box the Shapiro-Wilk test statistic will be used. *That transformation with the highest p-value based on the Shapiro-Wilk test will be selected as the optimal pilot transformation.* It seems logical that the parameter estimation technique and the method of selecting the optimal pilot transformation from the black box should differ. Thus, care is taken not to introduce bias into the selection procedure. For this reason the profile likelihood procedure will, among others, be applied to estimate the unknown transformation parameters.

Note, however, that the Shapiro-Wilk p-value can only be calculated for sample sizes ≤ 2000 when using the Fortran subroutine "SPWLK". For any sample size > 2000 we propose the selection of the optimal transformation from the black box, using the difficulty measure defined in Section 2.1.2 (see (2.16)), i.e.,

$$\sigma_y^5 R(f_Y''),$$

where $R(f_Y'')$ can be estimated using the procedure discussed in Section 2.1.4 and σ_y with the robust scale estimator (2.22). The transformation selection procedure defined above will ensure that the selected transformation renders data from some density which is easy to estimate. Alternatively, Tan, Gan and Chang (2004) proposed the use of a normal quantile plot to select a suitable transformation function. However, their procedure is not automatic but requires visual inspection of the resulting qq-plots. Hence, this procedure will not be utilized in this dissertation. Instead, we propose in Section 3.3.3 a fully automatic procedure based on qq-plots.

The transformation selection procedure in this dissertation is summarized as follows.

Select the transformation that:

- maximizes the Shapiro-Wilk p-value of the transformed data if $n \leq 2000$,
- minimizes an estimate of the functional $\sigma_y^5 R(f_Y'')$ if $n > 2000$.

Note that this selection procedure is implemented after parameter estimation which is discussed next.

In the context of transformation kernel density estimation Wand et al. (1991) used a transformation that is scale preserving. Motivation for this scale preservation follows from the discrepancy measure used to assess the performance of the kernel estimator, i.e., the MISE (see Section 2.1.1 for more detail on this measure). The mean integrated squared error (MISE) on the transformed scale, $MISE_Y[\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})]$, is easier to apply when searching for the optimal smoothing h and transformation parameter(s), than the mean integrated squared error based on the original scale, $MISE_X[\hat{f}_X(\cdot; h, \boldsymbol{\lambda})]$. The MISE on the transformed scale is however not invariant to the scale transformation $Y \rightarrow cY$ ($c \neq 0$). For this reason Wand et al. (1991) ensured that the scale is preserved when mapping from X to Y . (See Section 4.1 for a more detailed discussion of the above statements.) To accomplish the scale invariance, the authors suggested the transformation

$$g_{\boldsymbol{\lambda}}(Z) = (\sigma_x/\sigma_y) \tilde{g}_{\boldsymbol{\lambda}}(Z), \quad (3.21)$$

where $\tilde{g}_{\boldsymbol{\lambda}}(\cdot)$ is one of the parametric transformations from the black box; σ_x and σ_y are scale parameters. Since σ_y is a function of $\boldsymbol{\lambda}$, the question arises as to what the influence of the unknown parameter in this scale measure will be when using maximum likelihood. It turns out that the scale change in (3.21) does not affect the estimation procedure, the proof of which is to follow.

Let X_1, \dots, X_n be *i.i.d.* random variables. Let $Z_i = (X_i - \mu_x)/\sigma_x$, where μ_x and σ_x are the location and scale parameters of X respectively, estimated using robust methods. Define for $i = 1, \dots, n$,

$$\begin{aligned} Y_i &:= g_{\boldsymbol{\lambda}}(Z_i), \\ \tilde{Y}_i &:= c_{\boldsymbol{\lambda}} g_{\boldsymbol{\lambda}}(Z_i), \end{aligned}$$

where $g_\lambda(z)$ is a monotonic real-valued function and c_λ is some positive constant, both depending on the parameter(s) λ . Denote the location and scale parameters of Y and \tilde{Y} by (μ_y, σ_y) and $(\mu_{\tilde{y}}, \sigma_{\tilde{y}})$ respectively. Also, suppose ζ consists of the location-scale densities, i.e.,

$$\zeta = \left\{ f : f(t) = \frac{1}{\sigma} f_o \left(\frac{t - \mu}{\sigma} \right), -\infty < \mu < \infty, \sigma > 0 \right\},$$

for some fixed known density function f_o .

Theorem 3.1: *The profile likelihood function of λ based on the Y_i 's, with density $f_{Y_1} \in \zeta$, attains its maxima at the same λ -values as those that maximize the profile likelihood function of λ based on the \tilde{Y}_i 's, with density $f_{\tilde{Y}_1} \in \zeta$.*

Proof: Note that μ_x and σ_x do not depend on λ , and can therefore be seen as fixed when setting up the likelihood function for estimating λ . The likelihood function of λ based on the X_i 's, and written in terms of the Y_i 's, is

$$\begin{aligned} L(\lambda, \mu_y, \sigma_y) &:= \prod_{i=1}^n f_{Y_1}(g_\lambda(Z_i)) g'_\lambda(Z_i) \sigma_x^{-1} \\ &= \prod_{i=1}^n \sigma_y^{-1} f_o \left(\frac{g_\lambda(Z_i) - \mu_y}{\sigma_y} \right) g'_\lambda(Z_i) \sigma_x^{-1}. \end{aligned} \quad (\text{A})$$

The likelihood function of λ written in terms of the \tilde{Y}_i 's is

$$\begin{aligned} \tilde{L}(\lambda, \mu_{\tilde{y}}, \sigma_{\tilde{y}}) &:= \prod_{i=1}^n f_{\tilde{Y}_1}(c_\lambda g_\lambda(Z_i)) c_\lambda g'_\lambda(Z_i) \sigma_x^{-1} \\ &= \prod_{i=1}^n \sigma_{\tilde{y}}^{-1} f_o \left(\frac{c_\lambda g_\lambda(Z_i) - \mu_{\tilde{y}}}{\sigma_{\tilde{y}}} \right) g'_\lambda(Z_i) c_\lambda \sigma_x^{-1} \\ &= \prod_{i=1}^n (\sigma_{\tilde{y}}/c_\lambda)^{-1} f_o \left(\frac{g_\lambda(Z_i) - \mu_{\tilde{y}}/c_\lambda}{\sigma_{\tilde{y}}/c_\lambda} \right) g'_\lambda(Z_i) \sigma_x^{-1} \\ &= L(\lambda, \mu_{\tilde{y}}/c_\lambda, \sigma_{\tilde{y}}/c_\lambda). \end{aligned} \quad (\text{from (A)})$$

Hence,

$$\begin{aligned} \max \tilde{L}(\lambda, \mu_{\tilde{y}}, \sigma_{\tilde{y}}) &= \max L(\lambda, \mu_{\tilde{y}}/c_\lambda, \sigma_{\tilde{y}}/c_\lambda) \\ &= \max L(\lambda, (c_\lambda \mu_{\tilde{y}})/c_\lambda, (c_\lambda \sigma_{\tilde{y}})/c_\lambda) \\ &= \max L(\lambda, \mu_y, \sigma_y), \end{aligned}$$

which completes the proof of the theorem.

The effect of this is that the likelihood function may be maximized for the transformation $g_\lambda(\cdot)$ after which the transformed data may be scaled in order to preserve the scale of the input data.

The profile likelihood method

Next, a general formula for the profile likelihood function for transformation to normality will be presented. This formula will be used to construct the profile likelihood functions of the chosen transformations. From (A) the likelihood function of λ is given by:

$$L(\lambda, \mu_y, \sigma_y) = \prod_{i=1}^n \sigma_y^{-1} f_o \left(\frac{g_\lambda(Z_i) - \mu_y}{\sigma_y} \right) g'_\lambda(Z_i) \sigma_x^{-1}.$$

Suppose that a transformation to normality is required then $f_o(\cdot) = \phi(\cdot)$, where ϕ is the standard normal density. The likelihood function will then be

$$L(\lambda, \mu_y, \sigma_y) = (2\pi\sigma_y^2)^{-\frac{n}{2}} \sigma_x^{-n} \prod_{i=1}^n e^{-\frac{1}{2} \left(\frac{g_\lambda(Z_i) - \mu_y}{\sigma_y} \right)^2} g'_\lambda(Z_i).$$

The log-likelihood function is given by

$$l(\lambda, \mu_y, \sigma_y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma_y^2 - n \log \sigma_x - \frac{n}{2\sigma_y^2} \frac{1}{n} \sum_{i=1}^n (g_\lambda(Z_i) - \mu_y)^2 + \sum_{i=1}^n \log g'_\lambda(Z_i).$$

The dilemma here is that the location and scale parameters μ_y and σ_y both depend on the unknown transformation parameter λ . This is where the profile likelihood function comes in. The assumption made above is that for a **given** λ , the transformed data $Y_i = g_\lambda(Z_i)$ will be normally distributed for $i = 1, \dots, n$, with mean μ_y and variance σ_y^2 . Using this assumption the maximum likelihood estimates of μ_y and σ_y^2 , for a given λ , are found and given by

$$\hat{\mu}_y = \frac{1}{n} \sum_{i=1}^n g_\lambda(Z_i),$$

$$\hat{\sigma}_y^2 = \frac{1}{n} \sum_{i=1}^n (g_\lambda(Z_i) - \hat{\mu}_y)^2.$$

The profile log-likelihood function is constructed by replacing these estimates into the log-likelihood function and removing terms that does not depend on λ , rendering

$$\tilde{l}(\lambda, \hat{\mu}_y, \hat{\sigma}_y) = -\frac{n}{2} \log \hat{\sigma}_y^2 + \sum_{i=1}^n \log g'_\lambda(Z_i) \quad (3.22)$$

Maximizing $\tilde{l}(\boldsymbol{\lambda}, \hat{\mu}_y, \hat{\sigma}_y)$ with respect to $\boldsymbol{\lambda}$ yields $\hat{\boldsymbol{\lambda}}_{mle}$. Expression (3.22) is a general form of the profile log-likelihood function needed to transform data to normality. Using Table 3.2 and Table 3.3 from Section 3.3.2 in conjunction with expression (3.22), one can construct the profile log-likelihood functions for all of the transformations considered in the black box. The result is shown in Table 3.4. Of course, when optimizing the profile

Table 3.4: Profile log-likelihood functions for transformations in the black box.

Transformation	$\tilde{l}(\boldsymbol{\lambda}, \hat{\mu}_y, \hat{\sigma}_y)$
Shifted Box-Cox	$-\frac{n}{2} \log \hat{\sigma}_y^2 + (\lambda_2 - 1) \sum_{i=1}^n \log(Z_i + \lambda_1)$
Yeo & Johnson	$-\frac{n}{2} \log \hat{\sigma}_y^2 + (\lambda - 1) \sum_{i=1}^n \text{sign}(Z_i) \log(Z_i + 1)$
Johnson	
$\gamma = 1$	$-\frac{n}{2} \log \hat{\sigma}_y^2 - \sum_{i=1}^n \log(1 + cJZ_i)$
$\gamma = 2$	$-\frac{n}{2} \log \hat{\sigma}_y^2 - \frac{1}{2} \sum_{i=1}^n \log(1 + c^2 Z_i^2)$
$\gamma = 3$	$-\frac{n}{2} \log \hat{\sigma}_y^2 - \sum_{i=1}^n \log(1 - c^2 Z_i^2)$
John & Draper	$-\frac{n}{2} \log \hat{\sigma}_y^2 + (\lambda - 1) \sum_{i=1}^n \log(Z_i + 1)$
Ruppert & Wand	$-\frac{n}{2} \log \hat{\sigma}_y^2 + \sum_{i=1}^n \log [\alpha + (1 - \alpha) \sqrt{2\pi} \phi(Z_i / \hat{\sigma}_z)]$

log-likelihood function one must adhere to the parameter restrictions as outlined in the definitions of the transformations.

Although maximum likelihood can be employed for the one parameter Box-Cox transformation, this method breaks down for the shifted Box-Cox power transformation as the problem is nonregular. The problem is nonregular since for the shifted Box-Cox transformation to apply we require $\lambda_1 > -\min\{Z_1, \dots, Z_n\}$, so that the support of the distribution of Z must be $(-\lambda_1, \infty)$. With the range of the observations dependent on the unknown shift parameter λ_1 , the distribution of the maximum likelihood estimate $\hat{\lambda}_1$ cannot be assumed to be close to normality. The effect of nonregularity on likelihood inference is discussed in Section 9.3 of Atkinson (1985), where numerical results show

that the ordinary likelihood function is unbounded at the edge of the parameter space, and may or may not have a local maximum somewhere else. Thus, there is an unbounded global maximum at $-\min\{Z_1, \dots, Z_n\}$. Atkinson, Pericchi and Smith (1991) show further that, as $\lambda_1 \rightarrow -\min\{Z_1, \dots, Z_n\}$, then $\lambda_2 \rightarrow 0$, meaning that, regardless of the data, the log transformation is chosen. To solve this problem Atkinson et al. (1991) proposed the use of grouped likelihood, which removes the unbounded maximum of the likelihood. However, to apply this method special attention should be given to numerical details, in specific, Atkinson et al. (1991) use approximate conditional probabilities in order to avoid inaccurate differencing of normal integrals. A related procedure is the maximum product of spacings proposed by Cheng and Amin (1983), for which Titterton (1985) showed that this may be interpreted as a form of grouped likelihood. *In this dissertation neither the likelihood nor the grouped likelihood estimation procedures are considered for the shifted Box-Cox transformation.* Two new methods are proposed to estimate transformation parameters. The first method minimizes a quantile discrepancy measure and will be called the **minimum residual** method, whilst the second method minimizes a probability discrepancy measure and is called the **minimum distance** method.

The minimum residual method

Let X_1, \dots, X_n be *i.i.d.* random variables distributed according to the probability law F_X and let $Z_i = (X_i - \mu_x) / \sigma_x$, $i = 1, \dots, n$, be the standardized data. Also, let $g_\lambda(\cdot)$ be a monotonic increasing transformation that takes as input Z and transform this data, such that the output data follow the probability law G_Y . Let G_Y be a location-scale distribution, thus of the form $G_Y(y) = G_0\left(\frac{y - \mu_y}{\sigma_y}\right)$, where G_0 is a known distribution function. Finally, let $Y_{(1)}, \dots, Y_{(n)}$ be the order statistics of Y_1, \dots, Y_n .

The purpose of the transformation model $g_\lambda(\cdot)$ is to relocate the input data in such a way that the output (transformed) data follow the probability law G_Y . From this point of view transformation can be seen as a matter of relocating data. The success of this relocation can be measured with a qq-plot of the standardized order statistics of the transformed data versus the quantiles of G_0 , i.e., $\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}$ vs. $G_0^{-1}\left(\frac{i-0.5}{n}\right)$. A successful transformation will render a qq-plot where the data points $\left[\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}; G_0^{-1}\left(\frac{i-0.5}{n}\right)\right]$ are as close as possible to the straight line with intercept zero and gradient one. Figure 3.6 shows the resulting plot using the data from Example 1 of Section 3.1, and applying the

shifted Box-Cox transformation with an arbitrarily chosen λ to assist in the explanation of this method. Note that here $G_0(\cdot) = \Phi(\cdot)$. For a correct parameter choice λ , the

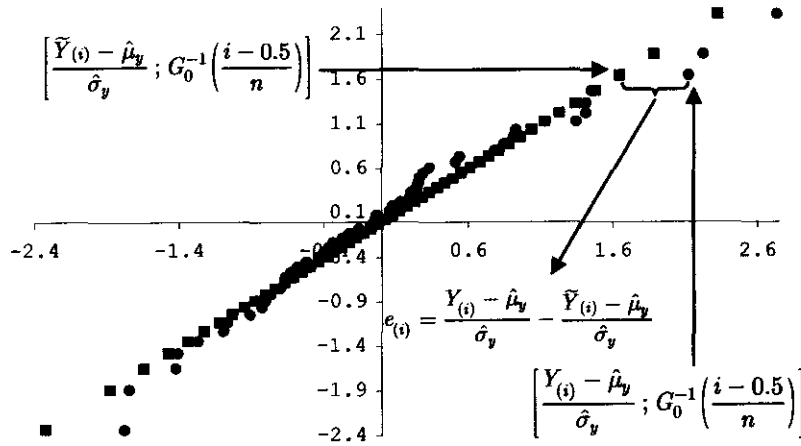


Figure 3.6: Normal qq-plot of the transformed data.

implied null model is $Y_{(i)} \approx \hat{\mu}_y + \hat{\sigma}_y G_0^{-1} \left(\frac{i-0.5}{n} \right)$. For an incorrect parameter choice, the implied null model will not be valid, thus the straight line connotation does not apply. Using the straight line one can, however, find the position $\frac{\tilde{Y}_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}$ on the y-scale for which the model $\tilde{Y}_{(i)} \approx \hat{\mu}_y + \hat{\sigma}_y G_0^{-1} \left(\frac{i-0.5}{n} \right)$ is correct. This is explained graphically in Figure 3.6. The idea is to find that parameter choice that would move $\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}$ as close as possible to $\frac{\tilde{Y}_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}$. This will be accomplished by minimizing a discrepancy measure between $\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}$ and $\frac{\tilde{Y}_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y}$, where $\tilde{Y}_{(i)}$ is replaced by the implied null model. Define the residuals

$$e_{(i)} = \frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} - \frac{\tilde{Y}_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} \approx \frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} - G_0^{-1} \left(\frac{i-0.5}{n} \right), \quad i = 1, \dots, n.$$

Mathematically, the minimum residual method entails the minimization of the following discrepancy measure

$$\hat{C}_n(\lambda) = \sum_{i=1}^n \left(\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} - G_0^{-1} \left(\frac{i-0.5}{n} \right) \right)^2,$$

with respect to λ , where $g_\lambda(\cdot)$ is any transformation from the black box. When transforming data to normality, $G_0 = \Phi$, and the resulting discrepancy measure is given by

$$\hat{C}_n(\boldsymbol{\lambda}) = \sum_{i=1}^n \left(\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} - \Phi^{-1} \left(\frac{i - 0.5}{n} \right) \right)^2 \quad (3.23)$$

The minimum distance method

Using the same notation as described in the minimum residual method, one may argue that an estimate of the distribution of a successful transformation must be close to the desired distribution G_Y . This can be quantified by using a weighted mean integrated squared error, WMISE, as described in Section 2.2. The distribution estimate used in this section, however, will be the empirical distribution function of the transformed data. The discrepancy measure is therefore defined as follows:

$$\hat{C}_n(\boldsymbol{\lambda}) = \int_{-\infty}^{+\infty} \left(F_{n,Y}(t) - G_0 \left(\frac{t - \hat{\mu}_y}{\hat{\sigma}_y} \right) \right)^2 w \left(G_0 \left(\frac{t - \hat{\mu}_y}{\hat{\sigma}_y} \right) \right) dG_0 \left(\frac{t - \hat{\mu}_y}{\hat{\sigma}_y} \right),$$

where $F_{n,Y}(t) = \frac{1}{n} \sum_{i=1}^n I(Y_i \leq t)$ is the empirical distribution function of the transformed data. Using the substitution $x = \frac{t - \hat{\mu}_y}{\hat{\sigma}_y}$ and using the definition of the empirical distribution function, the WMISE is given by

$$\begin{aligned} \hat{C}_n(\boldsymbol{\lambda}) &= \int_{-\infty}^{+\infty} \left\{ \frac{1}{n} \sum_{i=1}^n I(Y_i \leq \hat{\mu}_y + \hat{\sigma}_y x) - G_0(x) \right\}^2 w(G_0(x)) dG_0(x) \\ &= \int_{-\infty}^{+\infty} \left\{ \frac{1}{n} \sum_{i=1}^n I \left(\frac{Y_i - \hat{\mu}_y}{\hat{\sigma}_y} \leq x \right) - G_0(x) \right\}^2 w(G_0(x)) dG_0(x) \\ &= \int_{-\infty}^{+\infty} \left\{ \frac{1}{n} \sum_{i=1}^n I \left[G_0 \left(\frac{Y_i - \hat{\mu}_y}{\hat{\sigma}_y} \right) \leq t \right] - t \right\}^2 w(t) dt. \end{aligned}$$

(using the substitution $t = G_0(x)$)

Remarks:

- If $w(t) = 1$, then $\hat{C}_n(\boldsymbol{\lambda})$ is the so-called ‘‘Cramér-von Mises’’ discrepancy measure. To minimize $\hat{C}_n(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is equivalent to minimizing

$$\hat{C}_n(\boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^n \left\{ G_0 \left(\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} \right) - \frac{i - 1/2}{n} \right\}^2 + \frac{1}{12n^2}.$$

See D’Agostino and Stephens (1986) (p. 101) for verification.

- If $w(t) = \frac{1}{t(1-t)}$ then $\hat{C}_n(\boldsymbol{\lambda})$ is the so-called ‘‘Anderson-Darling’’ discrepancy measure. To minimize $\hat{C}_n(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is equivalent to minimizing

$$\hat{C}_n(\boldsymbol{\lambda}) = -\frac{1}{n^2} \sum_{i=1}^n \left[(2i-1) \log G_0 \left(\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} \right) + (2n+1-2i) \log \left(1 - G_0 \left(\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} \right) \right) \right] - 1.$$

See D'Agostino and Stephens (1986)(p. 101) for verification.

In this dissertation the unweighted distance measure ($w(t) = 1$) is considered. Also note that for transformation to normality $G_0(\cdot) = \Phi(\cdot)$. The discrepancy measure to minimize is then given by

$$\hat{C}_n(\boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^n \left\{ \Phi \left(\frac{Y_{(i)} - \hat{\mu}_y}{\hat{\sigma}_y} \right) - \frac{i-1/2}{n} \right\}^2 + \frac{1}{12n^2} \quad (3.24)$$

Alternative parameter estimation techniques

In the context of transformation kernel density estimation, Wand et al. (1991) and Yang and Marron (1999) estimate the transformation parameters such that the performance of the density estimate is asymptotically optimized. The reader is advised to first look at Section 2.1 (kernel density estimation) before reading the rest of this section. Also, note that in this section only a short outline of the methods proposed by the above-mentioned authors will be discussed, for a more detailed discussion the reader is referred to Chapter 4. The asymptotic mean integrated squared error (AMISE_Y) is used in order to find an optimal bandwidth. The discrepancy measure is based on the Y-scale for computational reasons as discussed in Section 4.1. Recall that from Section 2.1 (see (2.13)) the asymptotic mean integrated squared error was found to be

$$\text{AMISE}_Y[\hat{f}_Y(\cdot; h)] = \frac{1}{nh} R(k) + \frac{1}{4} h^4 \mu_2(k)^2 R(f_Y''). \quad (3.25)$$

Minimizing the right-hand side of (3.25) with respect to h results in the asymptotic optimal bandwidth

$$h_{\text{AMISE}_Y} = \left[\frac{R(k)}{\mu_2(k)^2 R(f_Y'') n} \right]^{\frac{1}{5}}. \quad (3.26)$$

Having found the asymptotic optimal bandwidth in (3.26), it is substituted back into (3.25), yielding

$$\inf_{h>0} \text{AMISE}_Y[\hat{f}_Y(\cdot; h)] = \frac{5}{4} C(k) R(f_Y'')^{1/5} n^{-4/5}, \quad (3.27)$$

where $C(K) = R(k)^{4/5} \mu_2(k)^{2/5}$ is a constant that only depends on the kernel k . In order to minimize the right-hand side of (3.27) one can minimize the constant $C(k)$ and the value $R(f_Y'')^{1/5}$. The effect of minimizing the constant $C(k)$ is discussed in Section 2.1.3. The quantity $R(f_Y'')$ is a measure of how easy the density f_Y can be estimated. This quantity is however not scale invariant. A scale invariant version of this quantity is given by

$$\sigma_y^5 R(f_Y''). \quad (3.28)$$

Terrell (1990) showed that the expression in (3.28) is minimized by the beta(4,4) density, thus the beta(4,4) density can be considered to be the easiest to estimate in terms of AMISE. For a more detailed discussion on the quantity in (3.28), the reader is referred to Chapter 2, Section 2.1.2. Wand et al. (1991) constructed the transformation in such a way that the transformed data are scale preserving (see equation 3.21), and therefore they chose the transformation parameter λ that minimizes an estimate of $R(f_Y'')^{1/5}$. Yang and Marron (1999), however, did not construct the transformation in a scale preserving manner, and therefore they chose the transformation parameter λ that minimizes an estimate of $\sigma_y R(f_Y'')^{1/5}$. Note that these two methods are equivalent, thus an attempt is made to choose λ in such a way that λ is optimal for the density estimate in terms of AMISE_Y . The idea is that the chosen transformation parameter estimate, $\hat{\lambda}$, will force the transformed data to have a distribution that is as close as possible to the beta(4,4) distribution. Note that this method is asymptotically optimal estimating densities and not to transform data to normality. In this dissertation, however, an attempt is made to transform data to normality, therefore, the above-mentioned procedure will only be used for reference purposes.

3.4 A new optimal semi-parametric transformation to normality

In this section results from Section 2.2, 3.2 and 3.3 will be utilized to define the optimal semi-parametric transformation to normality. Thereafter, graphs will be presented to illustrate the efficacy of the new transformation. Data from a uniform distribution on $[0, 1]$, negatively skewed unimodal distribution, positively skewed unimodal distribution, skewed bimodal distribution, kurtotic unimodal distribution, bimodal distribution, trimodal distribution and the claw distribution (a distribution with five modes) will be

considered. The distributions considered above were carefully selected to put the newly developed method under strain. Let X_1, \dots, X_n be i.i.d. random variables distributed according to the probability law F_X with the associated order statistics $X_{(1)}, \dots, X_{(n)}$.

Notation

Suppose that $Z_{x,j} = \frac{X_j - \hat{\mu}_x}{\hat{\sigma}_x}$, $j = 1, \dots, n$, then, for ease of notation, the subscript j will be suppressed, i.e., we write Z_x instead of $Z_{x,j}$. The little x will therefore indicate that the X -data are standardized. Also, if $Y_{i,j}$, $j = 1, \dots, n$, denote the transformed data at the i -th iteration step, we will simply write Y_i (omitting the index j). The index i will therefore be used as an iteration step index throughout the discussion below. The method is summarized as follows:

Step 1 Standardize the input data, X , using the robust standardization method as described in Section 3.3. Thus

$$Z_x = \frac{X - \hat{\mu}_x}{\hat{\sigma}_x},$$

where $\hat{\mu}_x = \hat{q}_2$ and $\hat{\sigma}_x = \min \left\{ s_x, (\hat{q}_3 - \hat{q}_1) / \left(\Phi^{-1}\left(\frac{3}{4}\right) - \Phi^{-1}\left(\frac{1}{4}\right) \right) \right\}$. Here $\hat{q}_1, \hat{q}_2, \hat{q}_3$ are the sample quartiles and s_x the usual sample standard deviation.

Step 2 Apply the black box of transformations to Z_x , thus

$$Y_0 = g_{\lambda}(Z_x),$$

where $g_{\lambda}(\cdot)$ is any of the transformations in the black box, i.e., the

- Shifted Box-Cox transformation,
- Yeo-Johnson transformation,
- John-Draper transformation,
- Johnson family of transformations,
- Ruppert-Wand transformation.

The parameter λ is estimated using the following methods:

- Profile maximum likelihood,
- Minimum residual,

- Minimum distance.

The ultimate transformation and estimation method chosen from this black box, is the transformation-estimation combination that has the highest p-value when testing for normality using the Shapiro-Wilk test statistic for $n \leq 2000$, and the combination that renders the easiest to estimate density if $n > 2000$. The reader is referred to Section 3.3.3 for a more detailed discussion on these transformation selection procedures.

Step 3 For each $i = 1, 2, \dots, r$ iterate the nonparametric transformation given in (3.10) by following steps one and two described below:

1. Calculate the standardized data

$$Z_{y_{i-1}} = \frac{Y_{i-1} - \hat{\mu}_{y_{i-1}}}{\hat{\sigma}_{y_{i-1}}},$$

where $\hat{\mu}_{y_{i-1}}$ and $\hat{\sigma}_{y_{i-1}}$ are the sample mean and robust scale estimate based on the transformed data from the previous iteration.

2. Apply the nonparametric transformation given by

$$Y_i = \Phi^{-1} \left[\hat{F}_{Z_{y_{i-1}}} (Z_{y_{i-1}}; \hat{h}_{i-1}) \right].$$

The bandwidth required may be selected according to the procedure of Polansky (1997) as described in Section 2.2.3. Also, no bandwidth adaptation, i.e., $\lambda_j = 1 \quad \forall j$ in (3.10) renders the highest p-value.

Step 4 The last iteration from Step 3 produces the transformed data Y_r . If scale preservation is required, then, the scale of the transformed variable produced in the last iteration is changed to the scale of the original input data. This is done by replacing Y_r with

$$\frac{\hat{\sigma}_x}{\hat{\sigma}_{y_r}} Y_r,$$

where $\hat{\sigma}_x$ and $\hat{\sigma}_{y_r}$ are the scale estimates of the original and transformed data respectively.

3.5 Application of the optimal transformation to simulated data

Next, the newly proposed semi-parametric transformation to normality will be applied to a random sample ($n = 400$) from some carefully selected distributions. The iteration number is fixed at $r = 2$ throughout the following illustrations. We will compare the new semi-parametric transformation with the “best” parametric transformation. The parametric transformation is selected from the black box according to the Shapiro-Wilk p-value. No bandwidth adaptation was performed, i.e., we used $\lambda_j = 1 \quad \forall j$ in (3.10). The legend used to describe the transformation results (first column of Table 3.6) is given in Table 3.5. The second column of Table 3.6 represents the density from which the data

Table 3.5: Descriptive legend for the first column of Table 3.6.

Distribution from which data are drawn
Curvature of the distribution utilized
Parametric transformation selected
Parameter estimation method selected
Estimated parameters
Shapiro-Wilk p-value (parametric)
Shapiro-Wilk p-value (semi-parametric)

are drawn. The third column represents the semi-parametric transformation function represented in red and the “best” parametric transformation represented in blue. The fourth column represents an ordinary kernel density estimate of the transformed data. If the transformation is successful, we would expect a normal like density in this column. The estimated density of the semi-parametric transformed data is presented in red and that of the parametric transformed data in blue.

Conclusions:

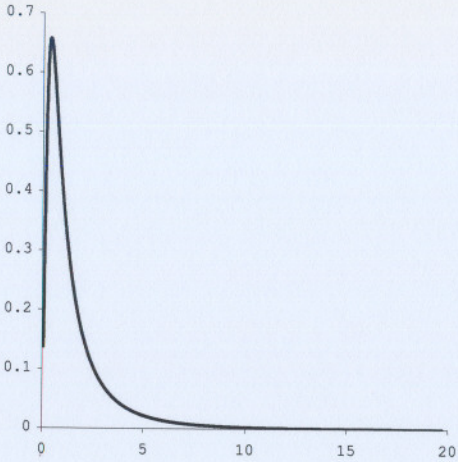
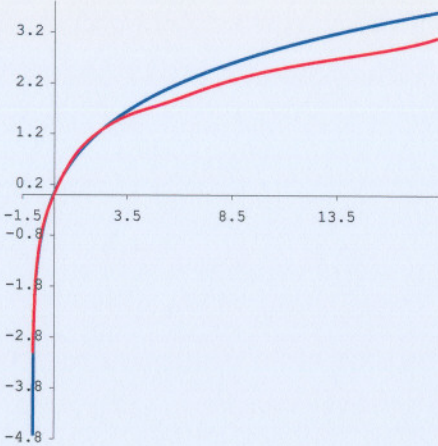
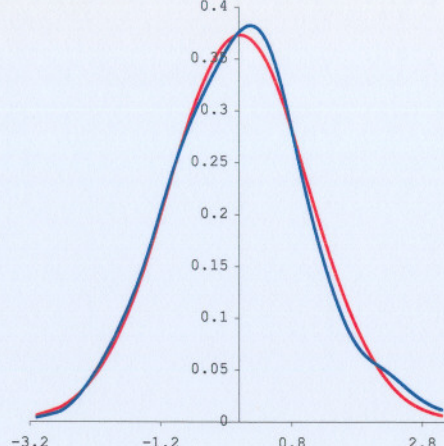
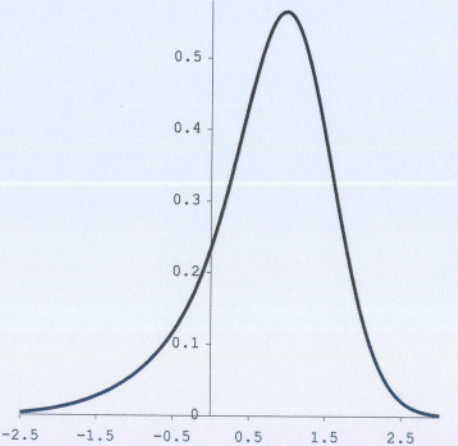
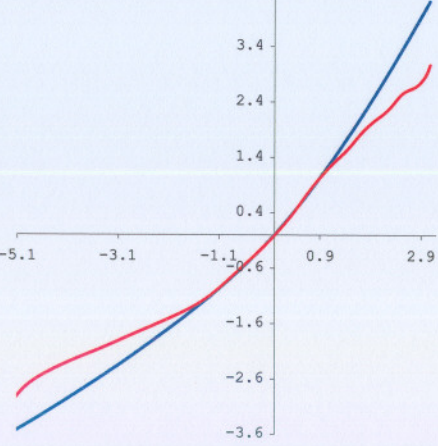
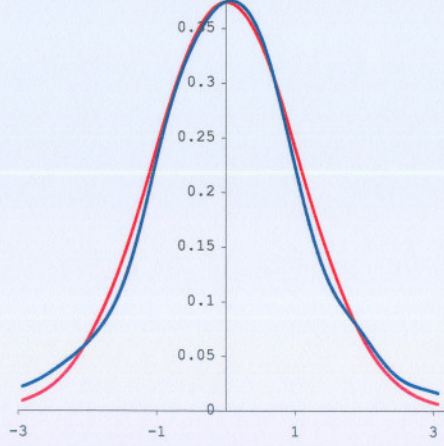
- The parametric transformation performs well in cases where only a convex, concave, convex to concave or concave to convex transformation is required. However, this transformation fails drastically when more than one shape change is required.

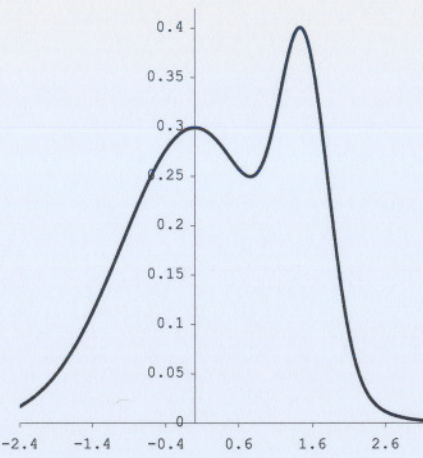
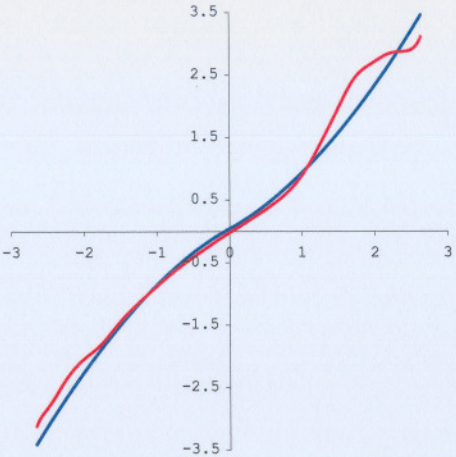
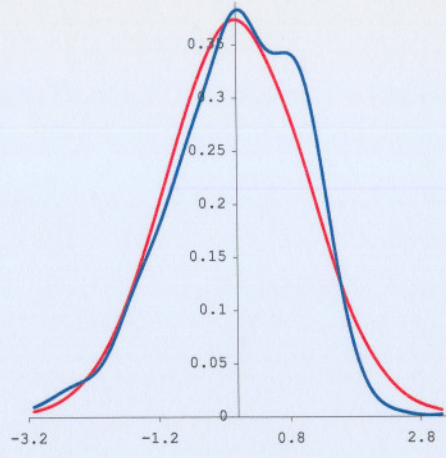
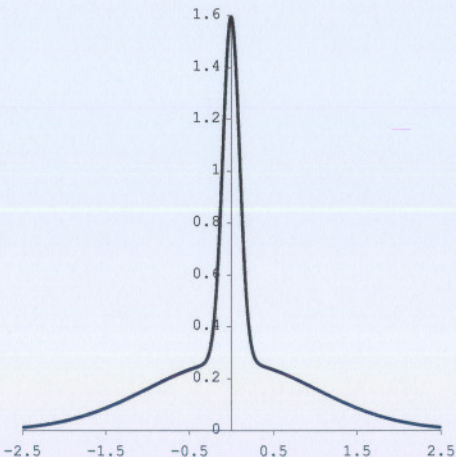
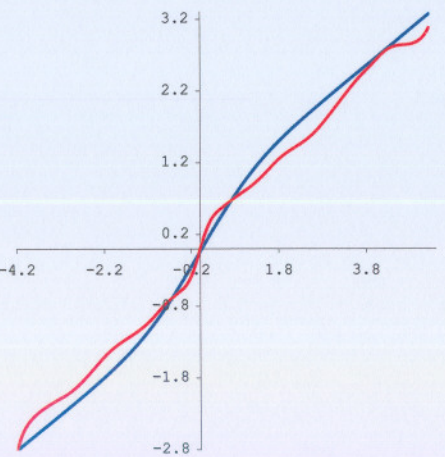
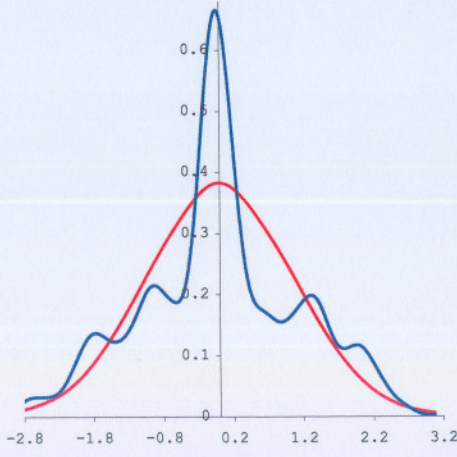
- For the data considered, the semi-parametric transformation outperformed the parametric counterpart significantly in all cases except for the lognormal data, where the two transformations performed almost identically. In this case a higher iteration value ($r > 2$) may be considered. The new transformation can assume all the shapes possible for a parametric transformation and any combination of these. Hence, input data are relocated more significantly which results in the high p-value when testing normality of the transformed data. In addition, the density estimate of the semi-parametric transformed data, appears to be much closer to the normal density than the density estimate of the parametric transformed data. Hence, using the newly proposed procedure will render data for which the density is easier to estimate.
- In the context of *transformation kernel density estimation*, the derivative of the transformation is required. From the graphs presented in Table 3.6 it should be clear that bandwidth adaptation is required to stabilize these derivatives in tail regions. This adaptation will be explored in Chapter 4.
- We propose $r = 2$ iterations in cases where the parametric transformation fails. However, $r = 0$ and $r = 1$ may also be considered otherwise.

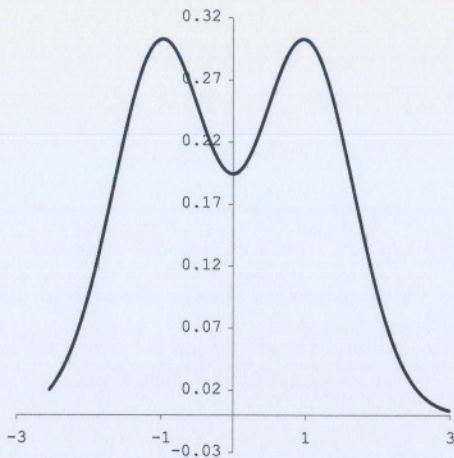
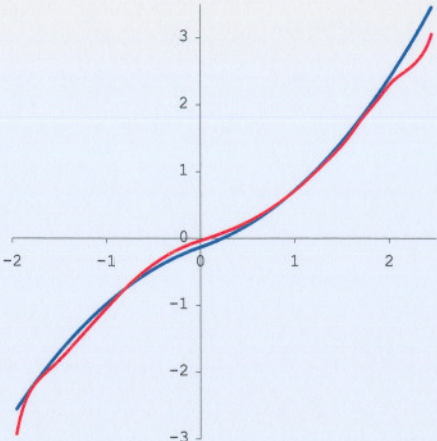
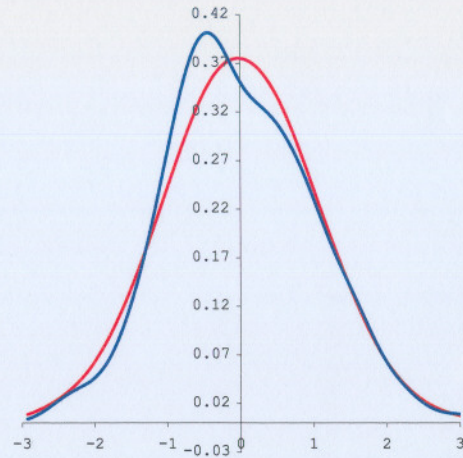
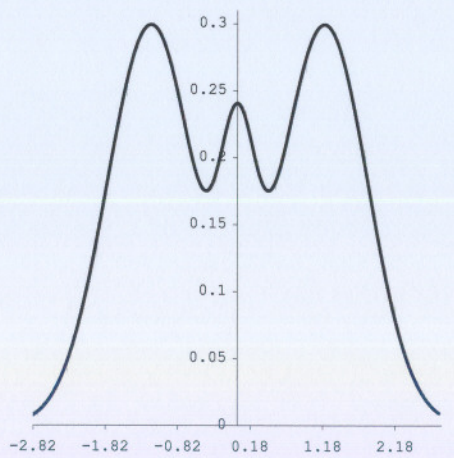
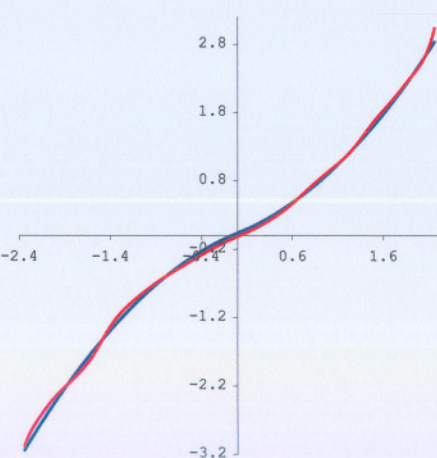
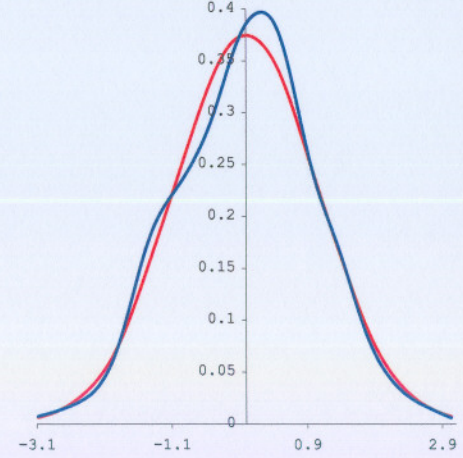
Remarks

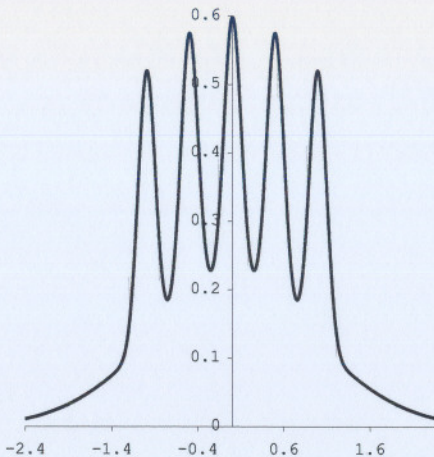
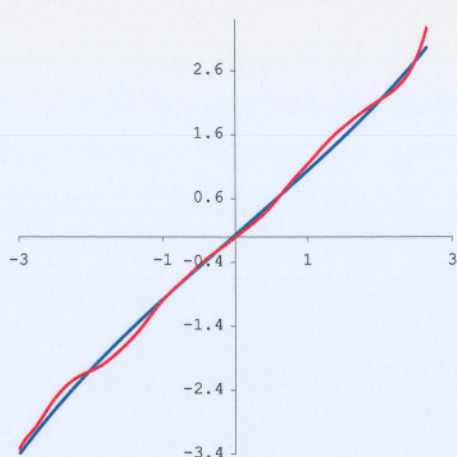
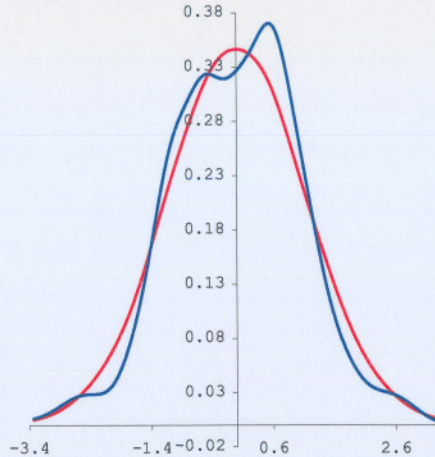
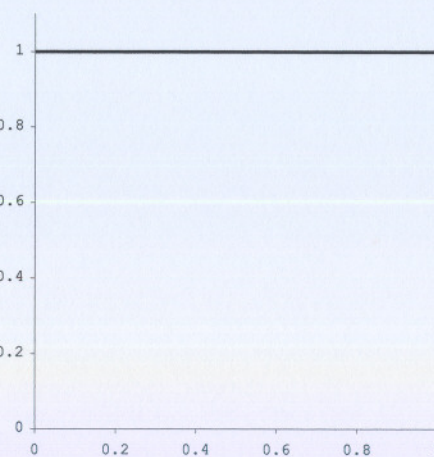
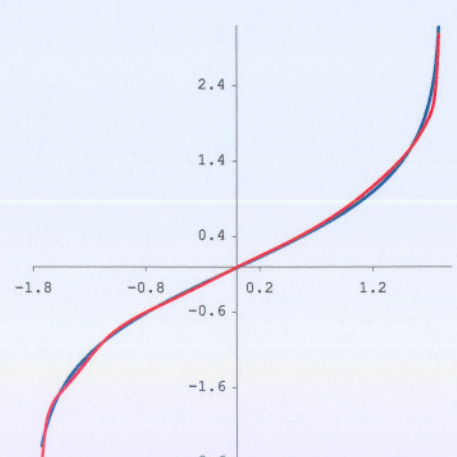
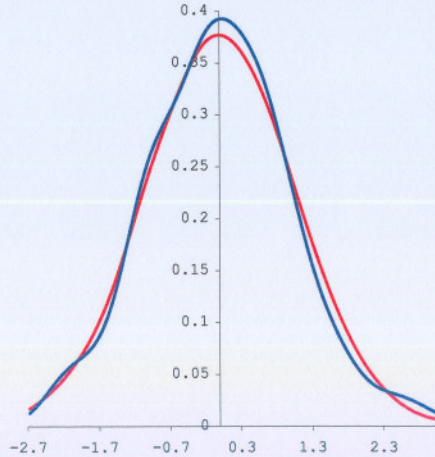
The reader is referred to tables displayed in Sections 5.1.1-5.1.13 where interesting Monte Carlo results are provided regarding mean p-values, the percentage number of times a certain transformation is selected and the percentage number of times a parameter estimation technique is chosen. These results are discussed in more detail in Section 5.1.

Table 3.6: Transformation of data.

Description	Density	Transformation	Density Estimate
<ul style="list-style-type: none"> • Standard Lognormal • Right Skewed • Shifted Box-Cox • Minimum Distance • $\lambda_1 = 0.986144$ • $\lambda_2 = 0.15$ • $p(\text{par}) = 0.991193$ • $p(\text{semi}) = 0.897014$ 			
<ul style="list-style-type: none"> • Skewed Unimodal • Negative Skew & One Mode • Yeo & Johnson • Minimum Residual • $\lambda = 1.339198$ • $p(\text{par}) = 0.389448$ • $p(\text{semi}) = 0.819035$ 			

Description	Density	Transformation	Density Estimate
<ul style="list-style-type: none"> • Skewed Bimodal • Negative Skew & Two Modes • John & Draper • Minimum Distance • $\lambda = 1.855469$ • $p(\text{par}) = 0.100705$ • $p(\text{semi}) = 0.903434$ 			
<ul style="list-style-type: none"> • Kurtotic Unimodal • High Kurtosis • Ruppert & Wand • Profile Maximum Likelihood • $\alpha = 0.482518$ • $p(\text{par}) = 0.000226$ • $p(\text{semi}) = 0.777237$ 			

Description	Density	Transformation	Density Estimate
<ul style="list-style-type: none"> • Bimodal • Two Modes • John & Draper • Minimum Distance • $\lambda = 2.338867$ • $p(\text{par}) = 0.250258$ • $p(\text{semi}) = 0.818493$ 			
<ul style="list-style-type: none"> • Trimodal • Three Modes • John & Draper • Profile Maximum Likelihood • $\lambda = 2.177734$ • $p(\text{par}) = 0.545100$ • $p(\text{semi}) = 0.861042$ 			

Description	Density	Transformation	Density Estimate
<ul style="list-style-type: none"> • Claw • Five Modes • Johnson System ($\gamma = 3$) • Minimum Distance • $c = 0.200664$ • $p(\text{par}) = 0.356219$ • $p(\text{semi}) = 0.840307$ 			
<ul style="list-style-type: none"> • Uniform • Constant Density • Johnson System ($\gamma = 3$) • Profile Maximum Likelihood • $c = 0.552268$ • $p(\text{par}) = 0.114273$ • $p(\text{semi}) = 0.655378$ 			

4

Transformation kernel density estimation

In this chapter we will discuss the transformation kernel density estimator which has the property of automatically addressing boundary bias and spurious bumps in the tails, both which are characteristics of the ordinary kernel density estimator. These issues enjoyed serious attention in the literature. The interested reader is referred to Chapter 2 for an account. In addition, the advantage of using transformations is that it still allows the use of a global bandwidth, although on a transformed scale. The different amount of smoothing needed at different locations is absorbed in the transformation function, making it possible to use a global bandwidth effectively. This is important since much is known about global bandwidth selection (see Section 2.1.4), but not for local bandwidth choice. *New contributions in this dissertation to the field of kernel density estimation are:*

- A new adaptation scheme that incorporates both an initial density and distribution function estimator as opposed to the procedure proposed by Abramson (1982) (see (2.49) for more detail).
- Modification of the newly proposed optimal transformation to normality as discussed in Section 3.4, to utilize it in a robust way for density estimation.

The layout of this chapter is as follows:

- Section 4.1 contains a discussion of the transformation kernel density estimator with special reference to the selection of an appropriate smoothing parameter.
- Section 4.2 is devoted to a newly proposed transformation kernel density estimator (TKDE).

4.1 The transformation kernel density estimator

The use of transformations in kernel density estimation has been proposed by Devroye and Györfi (1985), Silverman (1986) and Wand et al. (1991). Related work were presented by Park, Chung and Seog (1992), Ruppert and Wand (1992), Marron and Ruppert (1994), Ruppert and Cline (1994), Hössjer and Ruppert (1995), Yang and Marron (1999), Markovitch and Krieger (2000) and Bolancé, Guillen and Nielsen (2003). Let X_1, \dots, X_n be a sample having density f_X , also let $g_\lambda(x)$ be a monotonic increasing transformation. The transformation kernel density estimator (henceforth referred to as TKDE) is given by

$$\hat{f}_X(x; h, \lambda) = g'_\lambda(x) \hat{f}_Y(g_\lambda(x); h, \lambda) = \frac{1}{n} \sum_{i=1}^n g'_\lambda(x) k_h \{g_\lambda(x) - g_\lambda(X_i)\}, \quad (4.1)$$

where λ is the transformation parameter(s), h is the smoothing parameter and $k_h(\cdot) = k(\cdot/h)/h$. It is important that the transformation function is sufficiently smooth so that \hat{f}_Y inherits the smoothness properties of \hat{f}_X . Estimation of the transformation parameter(s) and transformation selection are discussed in Section 3.3.3. It is important to note that the derivative of the John & Draper transformation function entails absolute values which render this transformation unusable in the context of the TKDE. The global discrepancy measure MISE will be employed to find a suitable smoothing parameter, h . Hence, we are interested in minimizing

$$\text{MISE}_X [\hat{f}_X(\cdot; h, \lambda)] = E \int \{\hat{f}_X(x; h, \lambda) - f_X(x)\}^2 dx, \quad (4.2)$$

with respect to h . Hence, utilizing (4.1) we find

$$\begin{aligned} & \text{MISE}_X [\hat{f}_X(\cdot; h, \lambda)] \\ &= E \int \{\hat{f}_Y [g_\lambda(x); h, \lambda] - f_Y [g_\lambda(x); \lambda]\}^2 [g'_\lambda(x)]^2 dx \\ &= E \int \{\hat{f}_Y [y; h, \lambda] - f_Y [y; \lambda]\}^2 [g'_\lambda (g_\lambda^{-1}(y))] dy \\ & \quad \left[\text{using the substitution } y = g_\lambda(x) \right] \end{aligned}$$

$$\begin{aligned}
&= \int \left\{ \text{Var}_Y \hat{f}_Y [y; h, \boldsymbol{\lambda}] + \text{Bias}_Y^2 \left(\hat{f}_Y [y; h, \boldsymbol{\lambda}] \right) \right\} \left[g'_\lambda \left(g_\lambda^{-1}(y) \right) \right] dy \\
&\approx \int \left\{ \frac{R(k) f_Y(y; \boldsymbol{\lambda})}{nh} + \frac{1}{4} h^4 \mu_2(k)^2 f_Y''(y; \boldsymbol{\lambda})^2 \right\} \left[g'_\lambda \left(g_\lambda^{-1}(y) \right) \right] dy \\
&\hspace{20em} \left[\text{using (2.10) and (2.11)} \right] \\
&= \text{AMISE}_X \left[\hat{f}_X(\cdot; h, \boldsymbol{\lambda}) \right]. \tag{4.3}
\end{aligned}$$

Minimizing (4.3) with respect to h yields the asymptotic optimal bandwidth

$$h_{\text{AMISE}_X} = \left[\frac{R(k) E_Y \left[g'_\lambda \left(g_\lambda^{-1}(y) \right) \right]}{\mu_2(k)^2 \mathcal{L}_Y(\boldsymbol{\lambda})} \right]^{1/5} n^{-1/5}, \tag{4.4}$$

where $\mathcal{L}_Y(\boldsymbol{\lambda}) = \int f_Y''(y; \boldsymbol{\lambda})^2 \left[g'_\lambda \left(g_\lambda^{-1}(y) \right) \right] dy$. Substituting the asymptotic optimal bandwidth in (4.4) into (4.3) we find that, for a fixed $\boldsymbol{\lambda}$, the smallest possible AMISE_X is given by

$$\inf_{h>0} \text{AMISE}_X \left[\hat{f}_X(\cdot; h, \boldsymbol{\lambda}) \right] = \frac{5}{4} C(k) \left\{ E_Y \left[g'_\lambda \left(g_\lambda^{-1}(y) \right) \right] \right\}^{4/5} \mathcal{L}_Y(\boldsymbol{\lambda})^{1/5} n^{-4/5},$$

where $C(k) = \mu_2(k)^{2/5} R(k)^{4/5}$ is a constant only depending on the kernel function k . It should be noted that the bandwidth in (4.4) is difficult to implement in practice, since the integral $\mathcal{L}_Y(\boldsymbol{\lambda})$ is difficult to estimate. However, it might be argued that, for a transformation to normality, the unknown $f_Y''(y; \boldsymbol{\lambda})$ may be replaced with a normal reference, which will render a simple calculation of $\mathcal{L}_Y(\boldsymbol{\lambda})$. Such an argument comprises two flaws. Firstly, the transformation $g_\lambda(x)$ should be successful in transforming the input data to normality, if this is not the case the resulting bandwidth can be considered to be in the class of the quick and simple normal scaled rule of thumb bandwidths. Secondly, the newly proposed semi-parametric transformation to normality, discussed in Section 3.4, has the ability to transform any input data successfully to normality, however, the inverse transformation involves quantile estimation. Such an exercise may be performed using kernel quantile estimators, which in turn requires additional bandwidth selection. For this reason we will not pursue the calculation of the optimal bandwidth by means of expression (4.4).

As an alternative we may consider the MISE_Y of $\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})$. However, MISE_Y is not invariant to the scale transformation $Y \mapsto cY$ with $c \neq 0$, hence, we need to ensure that the scale of the input data is preserved when mapping from X to Y . This is easily

accomplished by dividing the transformed data with an estimate of its scale measure, and then multiply the result with an estimate of the scale measure of the original data. Applying the asymptotic results presented in Section 2.1.1, in specific (2.10), (2.11) and (2.13) we find

$$\begin{aligned} \text{MISE}_Y [\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})] &= E \int \{ \hat{f}_Y(y; h, \boldsymbol{\lambda}) - f_Y(y; \boldsymbol{\lambda}) \}^2 dy \\ &\approx \frac{R(k)}{nh} + \frac{1}{4} h^4 \mu_2(k)^2 R(f_Y''(\cdot; \boldsymbol{\lambda})) = \text{AMISE}_Y [\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})]. \end{aligned} \quad (4.5)$$

Minimizing (4.5) with respect to h results in the asymptotic optimal bandwidth

$$h_{\text{AMISE}_Y} = \left[\frac{R(k)}{\mu_2(k)^2 R(f_Y''(\cdot; \boldsymbol{\lambda}))} \right]^{1/5} n^{-1/5}, \quad (4.6)$$

where, for a given $\boldsymbol{\lambda}$, the density functional $R(f_Y''(\cdot; \boldsymbol{\lambda}))$ may be estimated according to the methods presented in Section 2.1.4. Hence, the bandwidth selection procedure proposed by Sheather and Jones (1991) may be utilized. Substituting (4.6) into (4.5) we find that the smallest possible AMISE_Y , for a given $\boldsymbol{\lambda}$, is given by

$$\inf_{h>0} \text{AMISE}_Y [\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})] = \frac{5}{4} C(k) R(f_Y''(\cdot; \boldsymbol{\lambda}))^{1/5} n^{-4/5}. \quad (4.7)$$

From (4.7) it is clear that one may select the transformation parameter $\boldsymbol{\lambda}$ to minimize $R(f_Y''(\cdot; \boldsymbol{\lambda}))$ or a scale invariant version thereof (see (2.16) for more detail). This is the route followed by most authors. The logic behind this procedure is to minimize the curvature in the density estimate of the transformed data, since $R(f_Y''(\cdot; \boldsymbol{\lambda}))$ is a global measure of the curvature present in a density, as discussed in Section 2.1.2. Terrell (1990) showed that the beta(4,4) density minimizes this roughness measure. Consequently, by choosing the transformation parameter(s) to minimize $R(f_Y''(\cdot; \boldsymbol{\lambda}))$ is an attempt to transform the data to have a density similar in shape than the beta(4,4) density. Alternatively, one might argue that this is an attempt to choose $\boldsymbol{\lambda}$ that yields the “easiest to estimate” density over the family of transformed densities indexed by $\boldsymbol{\lambda}$. This transformation parameter(s) selection procedure was previously discussed in Section 3.3.3. However, for the newly proposed TKDE procedure an attempt is made to transform the data to normality and therefore the transformation parameter(s) needs to be selected on different considerations. The transformation parameter(s) estimation procedures for the newly proposed TKDE’s can be found in Section 3.3.3. It should be clear that the asymptotic optimal

bandwidth in (4.6), is easier to implement in practice than the bandwidth in (4.4). However, keep in mind that we seek good performance of \hat{f}_X . Hence, we will establish the relationship between the discrepancy measures based on the X - and Y - data.

$$\begin{aligned} \text{MISE}_Y [\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})] &= E \int \left\{ \hat{f}_Y(y; h, \boldsymbol{\lambda}) - f_Y(y; \boldsymbol{\lambda}) \right\}^2 dy \\ &= E \int \left\{ \hat{f}_X(g_{\boldsymbol{\lambda}}^{-1}(y); h, \boldsymbol{\lambda}) \frac{dg_{\boldsymbol{\lambda}}^{-1}(y)}{dy} - f_X(g_{\boldsymbol{\lambda}}^{-1}(y)) \frac{dg_{\boldsymbol{\lambda}}^{-1}(y)}{dy} \right\}^2 dy \\ &= E \int \left\{ \hat{f}_X(x; h, \boldsymbol{\lambda}) - f_X(x) \right\}^2 \frac{1}{g'_{\boldsymbol{\lambda}}(x)} dx. \\ &\quad \left[\text{using the substitution } y = g_{\boldsymbol{\lambda}}(x) \right] \end{aligned}$$

The MISE_Y expression given above may be compared to the MISE_X expression based on the X - data, i.e.,

$$\text{MISE}_X [\hat{f}_X(\cdot; h, \boldsymbol{\lambda})] = E \int \left\{ \hat{f}_X(y; h, \boldsymbol{\lambda}) - f_X(y; \boldsymbol{\lambda}) \right\}^2 [g'_{\boldsymbol{\lambda}}(g_{\boldsymbol{\lambda}}^{-1}(y))] dy.$$

Such a comparison lead to the conclusion that $\text{MISE}_Y [\hat{f}_Y(\cdot; h, \boldsymbol{\lambda})]$ can be interpreted as a weighted X - data MISE, and vice versa. Wand et al. (1991) comment that in the simulated examples they considered, little practical difference was found between the density estimators attempting to minimize MISE_X and MISE_Y . Based on the above-mentioned and computational considerations, bandwidth selection will be executed by using MISE_Y .

Next, a short summary of the existing literature will be presented. Perhaps the most informative article on the topic was written by Wand et al. (1991). The transformation considered is the shifted Box-Cox of the form

$$\tilde{y} = \tilde{g}_{\boldsymbol{\lambda}}(x) = \begin{cases} (x + \lambda_1)^{\lambda_2} \text{sign}(\lambda_2) & \lambda_2 \neq 0, \\ \ln(x + \lambda_1) & \lambda_2 = 0. \end{cases}$$

Note that the shape of this transformation is convex if $\lambda_2 > 1$ and concave if $\lambda_2 < 1$. Hence, the ability of the proposed transformation to transform any given data is limited. However, the authors primarily considered density estimation for positive, skewed data. The interested reader is referred to Section 3.3.2 for a more detailed discussion concerning the curvature of transformations. The transformation is then scaled to preserve the scale of the original input data, i.e.,

$$y = g_{\boldsymbol{\lambda}}(x) = \frac{\hat{\sigma}_x}{\hat{\sigma}_y} \tilde{g}_{\boldsymbol{\lambda}}(x). \quad (4.8)$$

This rescaling is necessary since the transformation parameter estimation technique proposed by Wand et al. (1991) is based on MISE_Y which is not scale invariant. Hence, the transformation parameters are selected based on the criterion $R\left(f_Y''(\cdot; \lambda)\right)^{1/5}$. From Section 3.3.3 we note that for the shifted Box-Cox transformation the parameter λ_1 is extremely difficult to estimate. To address this, Wand et al. (1991) found a reparametrization of (λ_1, λ_2) , which will be described next.

Let $q_1 < q_2 < q_3$ be three points in the range of X , where q_1, q_2, q_3 might be the three quartiles. Let

$$\delta_i = \tilde{g}'_{\lambda}(q_i), \quad i = 1, 2, 3.$$

It was shown in Section 2.1.6 that the TKDE defined in (4.1) can be viewed as the conventional density kernel estimator with bandwidth $h/\tilde{g}'_{\lambda}(x)$. Based on this observation Wand et al. (1991) defined

$$d_i = \frac{\delta_i}{\delta_2} = \left(\frac{q_i + \lambda_1}{q_2 + \lambda_1}\right)^{\lambda_2 - 1}, \quad i = 1, 2, 3. \quad (4.9)$$

Note that $d_2 = 1$. Since the effective bandwidth at q_i is h/δ_i , the reparametrization has intuitive appeal in the sense that d_1 is the ratio of the effective bandwidth at q_2 to the effective bandwidth at q_1 , hence, $d_1 = \frac{h}{\delta_2}/\frac{h}{\delta_1} = \delta_1/\delta_2$. The parameter d_3 has a similar interpretation. Thus, one can choose h to control the degree of smoothing at the center of the X -data (near q_2) and then choose d_1 and d_3 to control the smoothing in the right and left tails, respectively. If $\lambda_2 > 1$, the shape of $\tilde{g}_{\lambda}(x)$ is convex and consequently $\tilde{g}'_{\lambda}(x)$ will increase as x increases, hence, $d_1 < 1 < d_3$. Conversely, if $\lambda_2 < 1$, $\tilde{g}_{\lambda}(x)$ is concave, hence, $d_1 > 1 > d_3$. Note that not all pairs (d_1, d_3) are possible, since we have the restrictions $-\min(X_1, \dots, X_n) < \lambda_1 < +\infty$ and $-\infty < \lambda_2 < +\infty$. The algebraic manipulation to follow is greatly simplified if $q_2 = 1$ and $\min(X_1, \dots, X_n) = 0$. This can be accomplished by using the linear preliminary transformation

$$\tilde{X} = \frac{X - \min(X_1, \dots, X_n)}{q_2 - \min(X_1, \dots, X_n)}, \quad (4.10)$$

where q_2 is the median based on the data X_1, \dots, X_n . Utilizing (4.10) we note that

$$\tilde{q}_2 = 1 \text{ and } \min(\tilde{X}_1, \dots, \tilde{X}_n) = 0.$$

Note that from (4.9) we have

$$\tilde{d}_i = \left(\frac{\tilde{q}_i + \tilde{\lambda}_1}{1 + \tilde{\lambda}_1}\right)^{\tilde{\lambda}_2 - 1}, \quad i = 1, 2, 3. \quad (4.11)$$

Next we will find the bounds for \tilde{d}_1 and \tilde{d}_3 . With some algebraic manipulation of the constraint $\tilde{\lambda}_1 > 0$ we find

$$\ln(\tilde{q}_3) > \ln\left(\frac{\tilde{q}_3 - 1}{1 + \tilde{\lambda}_1} + 1\right) > 0 \quad \text{and} \quad (4.12)$$

$$0 > 1/\ln(\tilde{q}_1) > 1/\ln\left(\frac{\tilde{q}_1 - 1}{1 + \tilde{\lambda}_1} + 1\right). \quad (4.13)$$

From (4.12) and (4.13) it follows that

$$\frac{\ln(\tilde{d}_3)}{\ln(\tilde{d}_1)} = \frac{\ln\left(\frac{\tilde{q}_3 - 1}{1 + \tilde{\lambda}_1} + 1\right)}{\ln\left(\frac{\tilde{q}_1 - 1}{1 + \tilde{\lambda}_1} + 1\right)} < \frac{\ln(\tilde{q}_3)}{\ln(\tilde{q}_1)}. \quad (4.14)$$

To determine a lower bound for $\ln(\tilde{d}_3)/\ln(\tilde{d}_1)$ note that from (4.11)

$$\tilde{\lambda}_2 = \frac{\ln(\tilde{d}_3)}{\ln(\epsilon_{\tilde{q}_3} + 1)} + 1 \quad \text{and} \quad (4.15)$$

$$\tilde{\lambda}_2 = \frac{\ln(\tilde{d}_1)}{\ln(\epsilon_{\tilde{q}_1} + 1)} + 1, \quad (4.16)$$

where $\epsilon_{\tilde{q}_i} = (\tilde{q}_i - 1)/(1 + \tilde{\lambda}_1)$. Utilizing (4.15) and (4.16) and some simple algebraic manipulation we find

$$\psi(\epsilon_{\tilde{q}_3}) = (\epsilon_{\tilde{q}_3} + 1)^{r_d} - r_q \epsilon_{\tilde{q}_3} - 1 = 0, \quad (4.17)$$

where $r_d = \ln(\tilde{d}_1)/\ln(\tilde{d}_3)$ and $r_q = (\tilde{q}_1 - 1)/(\tilde{q}_3 - 1)$. Hence, the root of equation (4.17), will produce a value for $\tilde{\lambda}_1$, which may be substituted into expression (4.15) or (4.16) to find $\tilde{\lambda}_2$. From the restriction $\tilde{\lambda}_1 > 0$ we find that

$$0 < \epsilon_{\tilde{q}_3} < \tilde{q}_3 - 1. \quad (4.18)$$

From this restriction and the fact that $\psi(0) = 0$ it follows that $\psi(x)$ should have a turning point between 0 and $\epsilon_{\tilde{q}_3}$. This yields that

$$\frac{\ln(\tilde{d}_3)}{\ln(\tilde{d}_1)} > -\left(\frac{\tilde{q}_3 - 1}{1 - \tilde{q}_1}\right). \quad (4.19)$$

The restrictions (4.14) and (4.19) determine the bounds, which are given by

$$-\left(\frac{\tilde{q}_3 - 1}{1 - \tilde{q}_1}\right) < \frac{\ln(\tilde{d}_3)}{\ln(\tilde{d}_1)} < \frac{\ln(\tilde{q}_3)}{\ln(\tilde{q}_1)}. \quad (4.20)$$

After observing that a rectangular grid of pairs \tilde{d}_1, \tilde{d}_3 based on (4.20) is unsatisfactory, Wand et al. (1991) suggested to use a trigonometric grid design based on (4.20). Let

$$\ln \tilde{d}_1 = r \cos \theta \quad \text{and} \quad \ln \tilde{d}_3 = r \sin \theta, \quad (4.21)$$

or

$$\theta = \tan^{-1} \left(\frac{\ln \tilde{d}_3}{\ln \tilde{d}_1} \right) \quad \text{and} \quad r = \sqrt{(\ln \tilde{d}_1)^2 + (\ln \tilde{d}_3)^2}.$$

Hence, from (4.20) we find

$$\tan^{-1} \left[- \left(\frac{\tilde{q}_3 - 1}{1 - \tilde{q}_1} \right) \right] < \theta < \tan^{-1} \left[\frac{\ln(\tilde{q}_3)}{\ln(\tilde{q}_1)} \right]. \quad (4.22)$$

It should be noted that when $\tilde{g}_\lambda(\cdot)$ is convex, then $d_1 < 1 < d_3$ and conversely if $\tilde{g}_\lambda(\cdot)$ is concave, then $d_1 > 1 > d_3$. This information should be incorporated in the polar grid design, hence, we will subtract π from the bounds given in expression (4.22) for the case where $\tilde{g}_\lambda(\cdot)$ is convex. Let θ_L and θ_R be the left and right sides of (4.22), respectively. A grid of points are generated via

$$\theta_i = \theta_L + \frac{(\theta_U - \theta_L) i}{10} \quad \text{and} \quad r_j = \frac{\sqrt{2} j}{10}, \quad i, j = 1, \dots, 10. \quad (4.23)$$

A grid point (θ, r) corresponds to $(\tilde{d}_1, \tilde{d}_3)$ through the relation (4.21). Putting together the ideas discussed above, the following general procedure should be followed for the implementation of the Wand et al. (1991) procedure:

- Step 1* Apply the linear preliminary transformation (4.10) to the input data.
- Step 2* Determine the grid (θ_i, r_j) with $i, j = 1, \dots, 10$, according to (4.22) and (4.23).
- Step 3* For each grid point (θ_i, r_j) , calculate $(\tilde{d}_1, \tilde{d}_3)$ according to (4.21).
- Step 4* For each pair $(\tilde{d}_1, \tilde{d}_3)$, find the root of (4.17), and solve $\epsilon_{\tilde{q}_3} = (\tilde{q}_3 - 1)/(1 + \tilde{\lambda}_1)$ to find $\tilde{\lambda}_1$. Also find $\tilde{\lambda}_2$ using (4.15) or (4.16).
- Step 5* For each pair $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ calculate the transformed value $\tilde{y} = \tilde{g}_{\tilde{\lambda}}(x)$ and preserve scale as outlined in (4.8), rendering the transformed value y .
- Step 6* The optimal transformation parameters are then selected according to MISE_Y considerations, that is, the pair $(\tilde{\lambda}_1, \tilde{\lambda}_2)$ that minimizes an estimate of $R(f_Y''(\cdot; \tilde{\lambda}))$. Note that if scale is not preserved a scale invariant version of this quantity is

required, i.e., the version defined in (2.16). Kernel estimation is discussed in Section 2.1.4 (see (2.24) with associated bandwidth calculated according to (2.43)). This leads to the l-stage high-tech procedure of Sheather and Jones (1991) as discussed in Section 2.1.4. In addition, the robust scale estimator (2.22) may be utilized.

Step 7 Estimate the density $f_X(x)$ using (4.1) with associated bandwidth calculated according to the procedure outlined in (4.6). Bandwidth selection is based on MISE_Y as pointed out previously.

Park et al. (1992) investigated the performance of the TKDE based on the procedure proposed by Wand et al. (1991) through a simulation study. The authors considered distributions which have support $[0, \infty)$ and report that the method works quite well, but is not so effective for distributions with relatively high density near zero. As a remedy, they proposed to restrict the range of λ_1 to $\lambda_1 > c$ ($c > 0$) in the minimization procedure. This is equivalent to considering a preliminary transformation $X \mapsto X + c$, where c is not involved in the minimization procedure but is predetermined. The authors reported that the preliminary shift transformation significantly improves the performance of the TKDE for distributions with relatively high density near zero. For other distributions, preliminary shift is not harmful in the sense of mean integrated squared error. They suggested that one should always make a preliminary shift transformation before attempting to find the best shifted power transformation. However, no guidance is given how to choose the preliminary shift parameter c . In Section 4.2 a preliminary shift parameter will be introduced for the implementation of the TKDE.

Ruppert and Wand (1992) considered densities with high kurtosis. The kurtosis of a probability density function f_X can be described in terms of its “peakedness” in the center and “heaviness” in the tails. A density with a high kurtosis generally has a sharp peak in the center and long tails, e.g., the Cauchy density. Ruppert and Wand (1992) made the assumption that f_X is symmetric around 0 and unimodal. The center of symmetry is achieved by subtracting the sample median from each observation, rendering the linear preliminary transformation

$$\widetilde{X} = X - q_2.$$

To reduce kurtosis a convex to concave transformation is required. Such a transformation has the effect of taking probability mass from both the peak and the tails and moving it to shoulders, which reduces peakedness and lightens the tails. Ruppert and Wand (1992) proposed the use of the transformation

$$\tilde{y} = \tilde{g}_\alpha(x) = \alpha x + (1 - \alpha)\hat{\sigma}_x \sqrt{2\pi} \left\{ \Phi(x/\hat{\sigma}_x) - \frac{1}{2} \right\}, \quad (4.24)$$

for $0 \leq \alpha \leq 1$, where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Decreasing α strengthens the kurtosis reduction of the transformation. When $\alpha = 1$, $\tilde{g}_\alpha(\cdot)$ has no effect, since the identity transformation is returned. The scale of the input data is preserved according to (4.8), i.e.,

$$y = g_\alpha(x) = \frac{\hat{\sigma}_x}{\hat{\sigma}_y} \tilde{g}_\alpha(x). \quad (4.25)$$

For the scale estimates the robust scale estimator (2.22) is recommended. In accordance with Wand et al. (1991), Ruppert and Wand (1992) chose the transformation parameter α according to MISE considerations, i.e., that α that minimizes an estimate of $R(f_Y''(\cdot; \alpha))$. The estimation of this quantity can be performed utilizing the high-tech procedure proposed by Sheather and Jones (1991), see Section 2.1.4 for more detail. The TKDE is then given by (4.1) with associated bandwidth calculated according to the procedure in (4.6). From a small simulation study, Ruppert and Wand (1992) concluded that the TKDE based on the transformation (4.24) is superior to the ordinary kernel density estimate for densities with high kurtosis, whilst similar performance can be expected for densities close to normality, i.e., densities that are easy to estimate.

Marron and Ruppert (1994) utilized the TKDE to reduce boundary bias where the support of the unknown density is in the interval $[0, 1]$. They proposed to transform the data to have a density with first derivative equal to zero at both boundaries of its support.

Ruppert and Cline (1994) were the first authors to propose the use of a nonparametric transformation to a predetermined target distribution $G_0(\cdot)$ to reduce bias in kernel density estimation. The associated target density will be denoted by $g_0(\cdot)$. The transformation is then given by

$$y = G_0^{-1} \left[\hat{F}_X(x) \right], \quad (4.26)$$

where $\hat{F}_X(\cdot)$ is a smooth estimate of the cumulative distribution function F_X of X_1, \dots, X_n . From arguments leading to the bias expression in (2.10) it follows that the bias for the

ordinary kernel density estimator defined in (2.1) at x has a formal asymptotic expansion of the form

$$\sum_{j=1}^{\infty} h^{2j} f_X^{(2j)}(x) \int u^{2j} k(u) du / (2j)!.$$

The authors argued that a uniform target distribution (where $G_0(\cdot)$ is the identity function) is particularly interesting, since its density has all derivatives equal to 0 so that bias is asymptotically negligible. However, for distributions with bounded support and for estimation of the density of the transformed data, boundary bias will occur. For a detailed discussion concerning boundary bias the reader is referred to Section 2.1.5. Moreover, spurious bumps in the tails will also be evident of the procedure. The reader is referred to Section 2.1.6 for a detailed discussion concerning spurious bumps in the tails. To confirm the occurrence of boundary bias and spurious bumps in the tails, consider the iterative procedure proposed by Ruppert and Cline (1994). Let t be the number of iterations. For $j = 1, \dots, t$, $\hat{f}_j(\cdot)$ and $\hat{F}_j(\cdot)$ are defined as follows:

$$\hat{f}_1(x) = \hat{f}_X(x; h_1) = \frac{1}{n} \sum_{i=1}^n k_{h_1}(x - X_i),$$

and

$$\hat{F}_1(x) = \int_{-\infty}^x \hat{f}_1(u) du,$$

if $t \geq 2$ then, from the definition of the TKDE in (4.1), for $j = 2, \dots, t$,

$$\begin{aligned} \hat{f}_j(x) &= (G_0^{-1})'(\hat{F}_{j-1}(x)) \hat{f}_{G_0^{-1}(\hat{F}_{j-1})} [G_0^{-1}(\hat{F}_{j-1}(x)); h_j] \\ &= \frac{\hat{f}_{j-1}(x)}{g_0[G_0^{-1}(\hat{F}_{j-1}(x))]} \hat{f}_{G_0^{-1}(\hat{F}_{j-1})} [G_0^{-1}(\hat{F}_{j-1}(x)); h_j], \end{aligned} \quad (4.27)$$

and

$$\hat{F}_j(x) = \int_{-\infty}^x \hat{f}_j(u) du.$$

Note that if $G_0(\cdot)$ is the uniform target distribution then (4.27) reduce to

$$\hat{f}_j(x) = \hat{f}_{j-1}(x) \hat{f}_{\hat{F}_{j-1}} [\hat{F}_{j-1}(x); h_j]. \quad (4.28)$$

From (4.27) and (4.28) the following disappointing conclusions can be reached:

- The transformation derivative $\hat{f}_{j-1}(x) / g_0[G_0^{-1}(\hat{F}_{j-1}(x))]$ inherits both boundary bias and spurious bumps in the tails, since $\hat{f}_{j-1}(x)$ is present in the derivative. This behavior was mentioned in Sections 2.1.5 and 2.1.6.

- If $G_0(\cdot)$ is the uniform target distribution then, from (4.28) it is clear that boundary bias will be present in the estimation of the density of the transformed data. For this reason Ruppert and Cline (1994) utilize boundary kernels to estimate the density of the transformed data.
- The derivative $\hat{f}_{j-1}(x) / g_0 \left[G_0^{-1} \left(\hat{F}_{j-1}(x) \right) \right]$ appearing in (4.27) can become very large in the tail regions. This will cause explosive behavior of the final density estimate in these regions. This phenomenon will be addressed subsequently in Section 4.2.
- Another point of criticism is that the authors use the bandwidths $h_j, j = 1, \dots, t$, appearing in the density estimates \hat{f}_j also as bandwidths for the distribution function estimates \hat{F}_j . Surely, appropriate bandwidths for \hat{F}_j can differ from appropriate bandwidths for \hat{f}_j .

For reasons outlined above, the TKDE developed by Ruppert and Cline (1994) will not be included in the simulation study. Despite the occurrence of these potentially harming effects, Ruppert and Cline (1994) implemented the procedure with $G_0(\cdot)$ being the uniform distribution. Their findings are:

- The nonparametric TKDE seems to be highly effective at capturing interesting features such as multiple modes and densities with sharp peaks.
- For extremely skewed or heavy-tailed densities, the poor performance of the initial kernel density estimate seriously degrades the performance of the TKDE. In such situations they recommend using an initial parametric transformation as discussed in Wand et al. (1991) and in Ruppert and Wand (1992). This idea is essentially pursued in the newly proposed TKDE's (see Section 4.2).
- For densities with compact support, the boundary bias of the initial kernel density estimate will persist in later iterations. To avoid this problem, they recommend using the parametric TKDE developed by Marron and Ruppert (1994) to get the initial estimator.

For the newly proposed TKDE's (Section 4.2) the target distribution $G_0(\cdot)$ will be the normal distribution function, $\Phi(\cdot)$. Hence, assuming some measure of normality of the data used in density estimates, boundary bias should not be a problem. Hössjer and

Ruppert (1995) studied the asymptotics for the nonparametric TKDE proposed by Ruppert and Cline (1994).

Yang and Marron (1999) proposed the use of the TKDE, where the transformation utilized is defined as a reparametrization of the versatile transformation family proposed by Johnson (1949). This family includes convex, concave, convex to concave and concave to convex transformations (see (3.19)). Note that the curvature of the transformation changes around zero. The reader is referred to Section 3.3.2, Table 3.3 for confirmation. For this reason we will employ the linear preliminary transformation

$$\tilde{X} = X - q_2.$$

Let $y = g_c(x)$ denote the transformation, with c the transformation parameter. Unlike Wand et al. (1991), Yang and Marron (1999) do not preserve the scale of the input data as described in (4.8). If the transformation parameter selection rule is based on MISE considerations, i.e., choose the parameter that minimizes an estimate of $R(f_Y''(\cdot; c))$, then, a scale invariant version of this quantity is required. From Section 2.1.2 (see (2.16)) we define the following scale invariant version

$$L(c) = \sigma_y R(f_Y''(\cdot; c))^{1/5}. \quad (4.29)$$

Kernel estimation of $R(f_Y''(\cdot; c))$ is discussed in Section 2.1.4, see (2.24), with associated bandwidth calculated according to (2.43). The robust scale estimator in (2.22) is utilized to estimate σ_y . Yang and Marron (1999) proposed to estimate the parameter c for each member of the transformation family, and to choose the family member for which $\hat{L}(\hat{c})$ is the smallest. This will ensure that the member selected is optimal in the sense that it produces the easiest to estimate density. The authors also noted that $L(c)$ can be regarded as a global roughness measure or global measure of curvature, and less curvature makes estimation easier. The TKDE, (4.1), is then applied with MISE_Y optimal bandwidth calculated according to (4.6). The bandwidth selection procedure of Sheather and Jones (1991) is employed. In addition it should be noted that the original input data can be transformed more than once, utilizing the Johnson family of transformations. This is a result of the richness of the family (the curvature it can assume). As a consequence the procedure proposed by Yang and Marron (1999) can be iterated any number of times. The authors found that transforming the data twice yields an estimate much superior to

the estimate without transformation, and in most cases, little improvement is achieved after two transformation steps. Specifically, they report that the transformation density estimate is better than the untransformed estimate in overall smoothness and the capturing of peaks.

Markovitch and Krieger (2000) considered TKDE to study WWW (World Wide Web) - traffic measurements since different traffic characteristics can be modelled by long-tail distributed random variables. The data considered have a support $[0; \infty)$. The transformation proposed is

$$y = g(x) = \frac{2}{\pi} \tan^{-1}(x) \quad \text{with} \quad g'(x) = \frac{2}{\pi(1+x^2)}.$$

This transformation is convex to concave and changes curvature around zero. For the data considered the transformation is concave. The proposed transformation is somewhat disappointing, since it does not depend on parameters to be estimated. Nevertheless, the input data will be mapped to the bounded domain $[0; 1)$, hence, boundary bias will be present in the density estimate of the transformed data. Markovitch and Krieger (2000) employed reflection to overcome this obstacle (see Section 2.1.5 for more detail on this technique). The usual TKDE defined in (4.1) is used.

Bolancé et al. (2003) estimate actuarial loss distributions based on a symmetrized version of the transformation approach proposed by Wand et al. (1991). From an actuarial background, the authors gave valuable motivational insight for the application of the TKDE in the actuarial context. The loss distribution is the probability distribution of the amount to be paid to the insured for the damage. Traditional methods for loss distributions use parametric models. Two of the most popular distributions are the lognormal (overall shape) and the pareto (tail behavior). Loss distributions have typically one mode for the low loss values and then a long heavy tail. The authors showed by means of a simulation study that the proposed method is able to estimate all three possible kind of tails, as defined in Embrechts, Klüppelberg and Mikosch (1997), namely the Fréchet type, the Gumbel type, and the Weibull type, which makes the methodology extremely powerful for actuaries in various disciplines. The authors also showed that the TKDE is able to estimate a heavy-tailed distribution beyond the data. Consider the following short summary of the three classes of tails:

- Fréchet Type* This class includes the Pareto, Burr, loggamma, Cauchy and t-distributions as well as various mixture models. We will refer to distributions in this class as *heavy-tailed*. Not all moments are finite for distributions in the Fréchet class.
- Gumbel Type* This class contains distributions whose tails decay roughly exponentially and we call these distributions *medium-tailed*. All moments exist for distributions in the Gumbel class. Examples are the normal, lognormal, exponential and gamma.
- Weibull Type* This class contains distributions that are *short-tailed* for example the uniform and beta.

It should be noted from the summary given above, that the Fréchet type distributions are of vital importance for financial applications. In conclusion, the authors mentioned that the TKDE can be useful to estimate the tail index (α), which is the shape parameter of the extreme value distribution. They reported that the closest possible Pareto distribution fit in the tail, using the weighted discrepancy measure

$$WISE = \int_{-\infty}^{+\infty} \{ \hat{f}(x; \hat{h}) - f(x; \alpha) \}^2 x^2 dx,$$

where $\hat{f}(x; \hat{h})$ represents the TKDE and $f(x; \alpha)$ the extreme value density, is much better than the Hill estimator (Hill (1975)) that is widely used in actuarial science and finance. Note that the use of a weighted discrepancy measure places more importance on the tail. An additional advantage of the proposed procedure is that all the data are used and consequently the estimation technique does not have to bother about where the tail begins such as the Hill estimator has to. For a more detailed discussion concerning extreme value theory and related topics, the reader is referred to Beirlant, Teugels and Vynckier (1996) and Embrechts et al. (1997). The TKDE proposed by Bolancé et al. (2003) basically applied the procedure proposed by Wand et al. (1991), but, with a slight deviation since they only considered transformations that give a symmetric distribution. The authors also mentioned that Pareto tail shape belongs to the class of transformations they considered, i.e., the shifted Box-Cox. Hence, we can expect that the TKDE behaves extremely well in the tails of distributions with heavy tails. Bolancé et al. (2003) gave the following motivations for a symmetric transformation distribution:

- A transformation that results in a symmetric distribution is bound to have a significant influence on a possibly heavy tail of the original distribution.
- Bandwidth selection for estimating the density of the transformed data is simplified, since a simple rule of thumb may be employed.
- The boundary problem will more or less disappear since the transformed distribution can be expected to level off slowly. Therefore, boundary kernels or other forms of correcting at the boundary may be ignored.

The procedure proposed by Bolancé et al. (2003) is exactly the same as that proposed by Wand et al. (1991) with the slight alteration that they restrict the set of parameters λ to give approximately zero skewness for the transformed data Y_1, \dots, Y_n . Skewness is defined as

$$\hat{\gamma}_Y = \left\{ n^{-1} \sum_{i=1}^n (Y_i - \hat{\mu}_Y)^3 \right\} / \left\{ n^{-1} \sum_{i=1}^n (Y_i - \hat{\mu}_Y)^2 \right\}^{3/2}.$$

This restriction is easily incorporated at Step 6, in the stepwise procedure provided in the discussion of the method proposed by Wand et al. (1991).

4.2 The new optimal semi-parametric TKDE

From the literature study presented in Section 4.1 it should be clear that the TKDE has the ability to overcome the boundary bias and spurious bumps in the tails commonly associated with the ordinary kernel density estimator, (2.1), and in addition is able to capture density curvature more prominently. The transformation density estimation method is able to do all that, provided that the correct transformation is selected and correctly applied. It is widely accepted that there is no single best method to estimate densities. However, in this section we will present a *new semi-parametric procedure* that will automatically detect and address the problems commonly associated with the ordinary kernel density estimator. The newly proposed procedure consists of a parametric transformation selected from a black box of transformations and a non-parametric counterpart. The parametric transformation is responsible for removing spurious bumps in the tails as well as minimizing boundary bias. The non-parametric transformation is responsible for minimizing boundary bias, removing spurious bumps in the tails as well as capturing density curvature more effectively. The TKDE can be considered similar to the original kernel estimator with a variable bandwidth (see Section 2.1.6 and Section 4.1 for

more detail) as a direct consequence of the use of the transformation. Hence, the transformations (both parametric and non-parametric) can be seen as an attempt to adapt the bandwidth (make it a more local choice) according to the grid position where the density estimate is required. In addition we will also introduce a generalized adaptation scheme that adapt the bandwidth according to the data, i.e., kernels with larger variances will be utilized in places where data are scarce, and conversely utilization of kernels with smaller variances where data are abundant. The use of such an adaptation scheme is vital for consistent density estimation in the tail regions. It turns out that the newly proposed adaptation scheme includes the adaptation scheme proposed by Abramson (1982), (see (2.49)), as a special case. A new location shift procedure will also be introduced.

From the broad overview given above it should be clear that the newly proposed procedure combines a number of high-tech procedures into one package in a natural way, rendering the possibility of optimal density estimation. These procedures include:

- Parametric transformations to normality, with a variety of transformation parameter estimation techniques.
- The newly proposed optimal non-parametric transformation, that may be iterated.
- A newly proposed generalized adaptation scheme.
- A newly proposed shift procedure.
- Utilization of the high-tech bandwidth selection procedures proposed by Sheather and Jones (1991) and Polansky (1997).
- The transformation kernel density estimator.

In this section it will become apparent that at the core of a well executed transformation kernel density estimate lies the selection of a suitable transformation function and the estimation of the transformation function derivative. Hence, the broader the range of distributions that can be transformed successfully, the broader the application potential.

We claim that our procedure handles density estimation in an automatic and quite natural way, hence the combination of the parametric and nonparametric transformations handles a wide variety of distributions. To prove this claim, the following densities will be included in the simulation study (Chapter 5):

- Densities from the *Fréchet*- and *Gumbel Type*, to show the removal of spurious bumps in the tails and the boundary bias correction ability.
- Densities with bounded support, such as the uniform density, to show the boundary bias correction ability.
- Densities with more than one mode, to show the ability of capturing density curvature.
- Densities that are easy to estimate using the ordinary kernel estimator, to show the behavior where little or no transformation is required.

If the newly proposed procedure is successful, as claimed, to estimate densities of the various forms described above, one would be tempted to challenge the general consensus that there is no single best method of density estimation, at least using the kernel method. Furthermore, the procedure will be valuable to practitioners from a non-mathematical audience, since all potentially harmful side effects of the density estimation process are automatically handled. We will now proceed with a discussion of the proposed procedure.

In essence, the newly proposed procedure utilizes the optimal semi-parametric transformation to normality (see Section 3.4) after which the ordinary TKDE defined in (4.1) is applied. Let X_1, \dots, X_n be i.i.d. random variables distributed according to the probability law F_X . The zero-step iteration of the optimal transformation is given (using the “ease of notation” by suppressing subscripts as described in Section 3.4) by

$$Y_0 = g_{\hat{\lambda}}(Z_x) \quad \text{and} \quad Z_x = \frac{X - \hat{\mu}_x}{\hat{\sigma}_x}.$$

The first-step iteration of the optimal transformation is then given by

$$Y_1 = \hat{g}_1(Z_{y_0}) = \Phi^{-1}[\hat{F}_{Z_{y_0}}(Z_{y_0}; \hat{h}_0)], \quad \text{where} \quad Z_{y_0} = \frac{Y_0 - \hat{\mu}_{y_0}}{\hat{\sigma}_{y_0}}. \quad (4.30)$$

Location-scale and bandwidth parameters are estimated according to the conventions described in Section 3.4. The parametric transformation function is given by $g_{\hat{\lambda}}(\cdot)$, selected from the proposed black box of transformations with parameter estimation and transformation selection as described in Section 3.4. Note that scale preservation is strictly speaking not necessary since transformation parameter(s) estimation is not based on

MISE considerations. Implementation of the TKDE in a predetermined grid point, say x_0 , requires the derivative of the transformation in (4.30):

$$\frac{d\hat{g}_1(z_{y_0})}{dx_0} = \frac{\hat{f}_{Z_{y_0}}(z_{y_0}; \hat{h}_0)}{\phi(\Phi^{-1}[\hat{F}_{Z_{y_0}}(z_{y_0}; \hat{h}_0)])} \frac{dg_{\hat{\lambda}}(z_x)}{dx_0} \frac{1}{\hat{\sigma}_{y_0}}. \quad (4.31)$$

From (4.31) it should be clear that this derivative could be highly explosive in the tail regions, i.e., where $\hat{F}_{Z_{y_0}}(z_{y_0}; \hat{h}_0)$ could be close to 0 or 1. In these cases $\phi(\cdot)$ approaches zero and the derivative will explode, having a devastating effect on the TKDE in these regions. For subsequent iterations the occurrence of this possible explosive behavior will continue. Observation of this potentially unfortunate behavior could raise doubt concerning a transformation to normality as one may favour a transformation to an alternative bounded distribution, such as the uniform distribution proposed by Ruppert and Cline (1994), for which this behaviour will certainly be less severe. However, the normal distribution was chosen based on ease of estimation (see Section 2.1.2), and more importantly, since this distribution has an unbounded support and its density approach 0 in the tail regions. Hence, boundary bias will be eliminated in the resulting TKDE. Next, we will focus on finding a sufficient procedure to estimate the derivative presented in expression (4.31).

First, observe that theoretically the derivative in (4.31) requires density and distribution function estimation using the same bandwidth, \hat{h}_0 . However, from AMISE considerations (see Chapter 2, specifically (2.14), (2.58) and (2.60)) we observe that the AMISE optimal bandwidth required for density estimation is of the order $O(n^{-1/5})$, while the AMISE optimal bandwidth for distribution function estimation is of the order $O(n^{-1/3})$. We can therefore replace \hat{h}_0 in the density estimate appearing in the numerator of the right-hand side in (4.31) by a bandwidth specifically designed for density estimation, such as the Sheather and Jones (1991) procedure. However, our numerical results indicate that such a choice weakens the ability of our TKDE to capture density curvature. For this reason we recommend the use of \hat{h}_0 (both in the numerator and denominator) determined for distribution function estimation, such as the bandwidth selection method proposed by Polansky (1997).

In addition, it was shown in Section 2.1.3 and Section 2.2.2 that the optimal kernel functions used for density and distribution function estimation are the Epanechnikov

density and uniform distribution kernel functions respectively. However, (4.31) suggests use of k and K , the density and distribution function counterparts. Since the normal density and normal distribution function are effective for estimating both densities and distribution functions (see Table 2.2 and Table 2.3), implementation of this kernel function is utilized in this dissertation.

It turns out that the best way to handle the potential explosive behavior in (4.31) is prevention combined with a suitable bandwidth adaptation scheme. This include altering the parametric as well as nonparametric transformations, hence, the proposed solution addresses all levels of transformation. To motivate alteration of the transformation at parametric level, consider standard lognormal input data. It is well known that the logarithmic transformation of this data will produce standard normally distributed data. Hence, a successful transformation, e.g., the shifted Box-Cox transformation (see (3.15)) should resemble the log function. Ideally, one would not wish to alter the shape of this transformation, which raises the question: how can we alter the parametric transformation to prevent the explosive behavior in (4.31), without altering the shape of the transformation? Once we understand the manner in which the parametric transformation influences the derivative (4.31), the answer to the question posed above will be trivial. Again, consider the standard lognormal data and the log transformation function, input data close to 0 are mapped to $-\infty$ while input data in the right tail are retracted. If these transformed data values are smaller than -4 we are in a region where $\hat{F}_{Z_{y_0}}(Z_{y_0}; \hat{h}_0)$ will be close to zero, resulting in the explosive behavior mentioned above. Furthermore, keep in mind that the transformation parameters are estimated using the data, but, for the simulation studies a general grid is required that could exceed the bounds of the data. Hence, even if the minimum input data point is not mapped to a dangerously large negative value, the minimum input grid point might be.

Given this insight, we conclude that the parametric transformation indeed influences the derivative in (4.31) and consequently the TKDE's. Also, it is apparent that the parametric transformation is potentially harmful in regions where the input data are stretched out, i.e., when the transformation is concave at the lower bound and/or convex at the upper bound. It is also important to note that in these regions the derivative of the parametric transformation function is large and when combined with the large

value of the derivative of the nonparametric transformation the resulting TKDE's may tend to infinity in these regions. The remedy proposed in this dissertation is to determine the parametric transformation and then to add a shift quantity to the input data. The transformation is then applied to the shifted input data. In this way the shape of the transformation is preserved. Consider the log transformation to transform standard lognormal data to normality as an example, adding a shift constant to the input data would make any input data point close to zero slightly larger and consequently map these points to smaller negative values without altering the shape of the transformation. Note that Park et al. (1992) proposed adding a constant $c > 0$ to the data when applying the TKDE proposed by Wand et al. (1991), and report that this exercise improves the performance of the method, probably because of smaller transformation derivatives in the potentially harmful tail regions. However, no indication was given how c might be chosen. Subsequently, we will introduce a procedure similar to the idea of Park et al. (1992). This procedure will be described in the stepwise algorithm of the newly proposed TKDE's presented below.

Consider the next level of correction, i.e., at the nonparametric level. Since the nonparametric transformation is employed after an initial parametric transformation, one may argue that the input data at this level possess some level of normality. Hence, the density and distribution function estimates needed in the derivative (4.31) should resemble the normal distribution in some sense, depending on the success of the initial parametric transformation. A *bandwidth adaptation scheme* is also introduced in the density and distribution function estimators based on the transformed data, so that the eventual bandwidths utilized in the tail regions are increased. This adaptation scheme is necessary, since ordinary density estimates suffer from spurious bumps, especially in regions where the derivative of the parametric transformation is large. For convenient notation and without loss of generality, consider the random variables Y_1, \dots, Y_n having density and distribution function, f_Y and F_Y respectively. The estimation procedure proposed by Abramson (1982) qualifies as such a bandwidth adaptation candidate. This procedure is discussed in Section 2.1.6. Recall that

$$\hat{f}(y; b) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i b} k\left(\frac{y - Y_i}{\lambda_i b}\right),$$

where

$$\lambda_i = \left\{ \frac{\tilde{f}(Y_i; b)}{g} \right\}^{-\alpha} \quad \text{and} \quad \log g = \frac{1}{n} \sum_{i=1}^n \log \tilde{f}(Y_i; b).$$

The notation λ_i here should not be confused with the λ -parameter appearing in the transformations of the black box.

In the expression given above g is the geometrical mean of a pilot density estimate $\tilde{f}(Y_i; b)$, with bandwidth b . Abramson (1982) proposed the use of $\alpha = 1/2$ based on theoretical considerations. Although this adaptation could be employed to estimate the density and distribution function present in the derivative of the nonparametric transformation, one might argue that a proper adaptation scheme for the derivative estimation should also incorporate a pilot distribution function estimate. *Based on this argument we developed a new adaptation scheme that includes both pilot density and pilot distribution function estimates, which includes as a special case the adaptation scheme proposed by Abramson (1982).* The newly proposed adaptation scheme is based on the fact that the TKDE can be considered as an ordinary kernel density estimator with variable bandwidth determined according to the reciprocal of the derivative of the transformation utilized. For a discussion on this topic the reader is referred to Section 2.1.6. Based on this observation and the versatility of the compactly supported beta density function, we propose a nonparametric transformation to the beta(α, β) distribution. The adaptation λ_i is then given by the derivative of the nonparametric transformation normalized with its geometric mean, raised to the power $\tilde{\alpha}$ and inverted. Let $be(\cdot, \alpha, \beta)$ and $Be(\cdot, \alpha, \beta)$ be the beta density and distribution function respectively with support $[0, 1]$. For a definition of the beta density the reader is referred to Section 2.1.2, specifically (2.17). From Section 3.2.1 (see (3.10)) it follows that we can define an optimal transformation to a predetermined beta distribution as follows

$$\hat{l}(Y_i) = Be^{-1} [\tilde{F}_Y(Y_i; \hat{h}), \alpha, \beta], \quad i = 1, \dots, n,$$

where \tilde{F}_Y is an ordinary kernel distribution function estimate (see (2.50)) with bandwidth determined according to the method proposed by Polansky (1997) (see Section 2.2.3). The derivative of this transformation is

$$\hat{l}'(Y_i) = \frac{\tilde{f}_Y(Y_i; \hat{h})}{be(Be^{-1} [\tilde{F}_Y(Y_i; \hat{h}), \alpha, \beta], \alpha, \beta)}, \quad i = 1, \dots, n,$$

where \tilde{f}_Y is an ordinary kernel density estimate (see (2.1)). Let

$$\tilde{\lambda}_i = \left\{ \frac{\hat{l}'(Y_i)}{g} \right\}^{-\tilde{\alpha}}, \quad \text{where } \log g = \frac{1}{n} \sum_{i=1}^n \log \hat{l}'(Y_i).$$

Also, let $\tilde{\lambda}_{(1)}, \dots, \tilde{\lambda}_{(n)}$ be the order statistics associated with $\tilde{\lambda}_1, \dots, \tilde{\lambda}_n$. The proposed adaptation scheme is then given by:

$$\lambda_i = \tilde{\lambda}_i + \frac{1 - \tilde{\lambda}_{(1)}}{c_a}, \quad (4.32)$$

where c_a is a positive constant, that determines the smallest bandwidth used. The shift $(1 - \tilde{\lambda}_{(1)})/c_a$ is motivated from the fact that smaller bandwidths lead to an increase in variance of the resulting density and distribution function estimators (see Section 2.1.1 and Section 2.2.1). Hence, to control the possible increase in variance of the adaptive density and distribution function estimators (since some bandwidths will be forced to be smaller), we introduce the shift constant $(1 - \tilde{\lambda}_{(1)})/c_a$. The corresponding adaptive density and distribution function estimators are then given by

$$\hat{f}(y; b) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i b} k \left(\frac{y - Y_i}{\lambda_i b} \right), \quad \text{and} \quad (4.33)$$

$$\hat{F}(y; h) = \frac{1}{n} \sum_{i=1}^n K \left(\frac{y - Y_i}{\lambda_i h} \right). \quad (4.34)$$

For the choice $c_a = 1$, the smallest bandwidth used in the estimates presented in (4.33) and (4.34) will be b and h respectively. Also, note that the procedure proposed by Abramson (1982) uses $c_a = +\infty$, hence, bandwidths smaller than b and h are used in areas where data are abundant. In this dissertation, the choice $c_a = 1$ is utilized. The bandwidths should be chosen optimally for density and distribution function estimation as described in Chapter 2. However, to estimate the derivative in (4.31) we will take $b = h$, as mentioned previously. Note that this adaptation scheme involves both pilot density and pilot distribution function estimation, hence, better performance can be expected when estimating the derivative in (4.31). The sensitivity parameter $\tilde{\alpha}$ will be fixed at $\tilde{\alpha} = 1/2$, which is similar to the choice made by Abramson (1982). Choosing $\alpha = \beta$ for the adaptation in (4.32) seems logical in the current context, since for this choice the beta density is symmetric around $1/2$. For $\alpha = \beta = 1$ the beta density becomes the uniform density, and as $\alpha = \beta$ increases the resulting density takes an increasingly normal like form. From this observation we conclude that $\alpha = \beta = 1$ will cause more

drastic adaptation, while larger $\alpha = \beta$ will result in less adaptation. Hence, larger values of $\alpha = \beta$ will cause a higher p-value for the transformed data while smaller values of $\alpha = \beta$ tend to apply more smoothing in the tail regions and consequently reduces the obtainable p-value. It should also be noted that the new adaptation scheme reduces to the adaptation scheme proposed by Abramson (1982) when $\alpha = \beta = 1$, $c_a = +\infty$ and \hat{b} (bandwidth optimal for density estimation) are used. The adaptation scheme proposed by Abramson (1982) can therefore be seen as similar to the application of an TKDE for a transformation to the uniform distribution. The effect of various values of $\alpha = \beta$ is illustrated in Figure 4.1. We are now armed with tools to estimate derivatives similar to

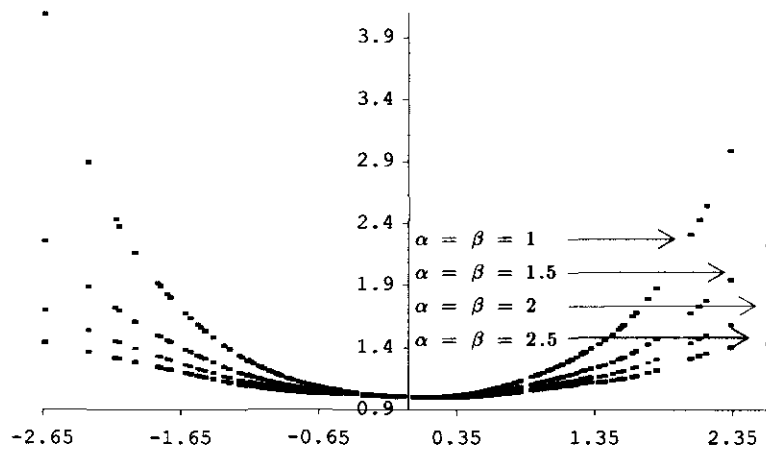


Figure 4.1: Values of λ_i generated by the newly proposed adaptation scheme for certain values of $\alpha = \beta$.

that presented in equation (4.31) successfully.

Next, the stepwise algorithm (with discussion) of the newly proposed optimal semi-parametric TKDE's will be presented. We will deviate slightly from the transformation procedure proposed in Section 3.4 to implement the procedures described above. Let X_1, \dots, X_n be i.i.d. random variables with probability law F_X . Also, let x_1, \dots, x_m be a finely spaced grid where m is the grid size. It should be noted that in the procedure described below all the calculations performed on the data are also performed on the grid. Define δ as a shift constant and let $0 < \varepsilon \leq 1$. Also, we will use the "ease of notation" by suppressing subscripts as described in Section 3.4:

Step 1 Standardize the input data, i.e., let

$$Z_x = \frac{X - \hat{\mu}_x}{\hat{\sigma}_x},$$

where $\hat{\mu}_x = \hat{q}_2$ is the sample median and $\hat{\sigma}_x$ is the robust scale estimator presented in equation (2.22). The standardized grid is then given by $z_x = (x - \hat{\mu}_x)/\hat{\sigma}_x$.

Step 2 Apply the black box of transformations to Z_x , thus

$$Y_0 = g_{\lambda}(Z_x).$$

The black box of transformations includes the Shifted Box-Cox, Yeo-Johnson, Johnson family and Ruppert-Wand transformations. The parameter(s) λ is/are estimated using the profile maximum likelihood, minimum residual and minimum distance methods. For sample sizes, $n < 2000$, the Shapiro-Wilk test statistic will be utilized to select the best transformation - parameter estimation combination. For larger sample sizes an estimate of the scale invariant global roughness measure

$$L(\hat{\lambda}) = \hat{\sigma}_{y_0} R(f''_{Y_0}(\cdot; \hat{\lambda}))^{1/5},$$

is minimized by the best transformation - parameter estimation combination. Kernel estimation of $R(f''_{Y_0}(\cdot; \hat{\lambda}))$ is discussed in Section 2.1.4, see (2.24), with associated bandwidth calculated according to (2.43). The robust scale estimator (2.22) is utilized to estimate σ_{y_0} . The transformed grid is then given by $y_0 = g_{\lambda}(z_x)$.

Step 3 Determine the curvature of the parametric transformation selected, utilizing results from Section 3.3.2. Knowledge of the curvature can be utilized to protect the TKDE against potentially explosive behavior in the tail regions. Let m_x and M_x be the minimum and maximum data-points respectively, i.e.,

$$m_x = \frac{X_{(1)} - \hat{\mu}_x}{\hat{\sigma}_x} \quad \text{and} \quad M_x = \frac{X_{(n)} - \hat{\mu}_x}{\hat{\sigma}_x}.$$

Also, consider the bias corrected density estimate (evaluated in the data-points) according to the adaptation scheme presented in expression (4.32) with $\alpha = \beta = 4$, i.e.,

$$\hat{f}_{bc}(X_i; g, \tilde{a}, \tilde{b}) = \left[\int_{\frac{X_i - \tilde{b}}{g}}^{\frac{X_i - \tilde{a}}{g}} k(t) dt \right]^{-1} \frac{1}{n} \sum_{j=1}^n \frac{1}{\lambda_j g} k\left(\frac{X_i - X_j}{\lambda_j g}\right), \quad (4.35)$$

where \tilde{a} and \tilde{b} are the left and right bounds of the domain of X . Note that for an unbounded domain we have $\tilde{a} = -\infty$ and $\tilde{b} = +\infty$. The bandwidth g is selected according to the method of Sheather and Jones (1991). The choice $\alpha = \beta = 4$ corresponds to a transformation to the beta(4,4) density, which is the easiest density to estimate, using kernel density estimation methods (see Section 2.1.2).

- If $g_{\hat{\lambda}}(\cdot)$ is concave, then, potential explosive behavior in the TKDE's can occur at the left bound of the transformed data, since the derivative of the parametric transformation can be very large in this region. To reduce this effect, choose that δ , $0 \leq \delta \leq 0.5$, which minimizes

$$\sum_{i=1}^n \left[g'_{\hat{\lambda}}(Z_{x,i} + \delta) \hat{f}_Y(g_{\hat{\lambda}}(Z_{x,i} + \delta); \hat{b}, \hat{\lambda}) - \hat{f}_{bc}(X_i; \hat{g}, \tilde{a}, \tilde{b}) \right]^2, \quad (4.36)$$

where \hat{f}_Y is based on no bandwidth adaptation, i.e., $\delta_i = 1$ for all i . The effect of δ is to push the transformed data away from the lower bound.

- If $g_{\hat{\lambda}}(\cdot)$ is convex, then, potential explosive behavior in the TKDE's can occur at the right bound of the transformed data, since the derivative of the parametric transformation can be very large in this region. To reduce this effect, select $-0.5 \leq \delta \leq 0$ to minimize the discrepancy measure given in (4.36). The effect of δ is to pull the transformed data away from the upper bound.
- For the case where $g_{\hat{\lambda}}(\cdot)$ is concave to convex, potential explosive behavior in the TKDE's can occur at both the left and right bounds of the transformed data. In this case, if $|g_{\hat{\lambda}}(m_x)| \geq |g_{\hat{\lambda}}(M_x)|$, minimize the discrepancy measure given in (4.36) for $0 \leq \delta \leq 0.5$; else, minimize the measure for $-0.5 \leq \delta \leq 0$. The effect of δ is to pull or push (δ can be a negative or positive constant) the transformed data away from the bound where the transformation derivative is the largest, hence, the bound where the most damage is caused in the TKDE's.
- For the case where $g_{\hat{\lambda}}(\cdot)$ is convex to concave, select δ exactly as above (concave to convex transformations).

It should be noted that $\delta = 0$ corresponds to the original transformation and that ideally, δ should be selected so that $|\delta|$ is as small as possible. For this

reason we devised the following stop criteria (for the discrepancy measure in (4.36)) in the search for an ideal δ . Let δ_i , $i = 1, \dots, r$, be the i -th δ value considered. It should be noted that we always select $\delta_1 = 0$ and

$$\begin{aligned}\delta_r &= 0.5, \text{ if } 0 \leq \delta_i \leq 0.5, \text{ and} \\ \delta_r &= -0.5, \text{ if } -0.5 \leq \delta_i \leq 0,\end{aligned}$$

for $i = 1, \dots, r$. If $g_{\hat{\lambda}}(\cdot)$ is a concave to convex transformation, we apply Criterion 1, and for all other transformation shapes Criterion 2 is applied.

Criterion 1: Stop the search when $\max \{g'_{\hat{\lambda}}(m_x + \delta_i), g'_{\hat{\lambda}}(M_x + \delta_i)\} \geq \max \{g'_{\hat{\lambda}}(m_x + \delta_{i-1}), g'_{\hat{\lambda}}(M_x + \delta_{i-1})\}$, $i = 2, \dots, r$. This will ensure that no artificial outliers are created in the transformed data.

Criterion 2: Stop the search when $[g_{\hat{\lambda}}(M_x + \delta_i) - g_{\hat{\lambda}}(m_x + \delta_i)] \leq \varepsilon [g_{\hat{\lambda}}(M_x) - g_{\hat{\lambda}}(m_x)]$, $i = 2, \dots, r$. For the simulation study presented in Chapter 5, we used $\varepsilon = 0.8$. This criterion ensures that the range of the shifted transformed data is not too small.

It should be noted that other ways of choosing δ can be invented to improve the performance of the eventual TKDE's. However, this was not pursued any further.

Once δ is determined, the transformed data and grid utilized is given by

$$\begin{aligned}Y_0 &= g_{\hat{\lambda}}(Z_x + \delta) \text{ and} \\ y_0 &= g_{\hat{\lambda}}(z_x + \delta),\end{aligned}\tag{4.37}$$

respectively. To choose α and β for the newly proposed adaptive scheme, we proceed to the following step.

Step 4 The 0-step TKDE is then given by

$$\hat{f}_X(x; \hat{b}_0, \hat{\lambda}) = \left(\frac{dy_0}{dx}\right) \hat{f}_{Y_0}(y_0; \hat{b}_0),\tag{4.38}$$

where $\hat{f}_{Y_0}(y_0; \hat{b}_0)$ is the adaptive KDE (see 4.33) for which the newly proposed adaptation scheme (see 4.32) is utilized. Also, \hat{b}_0 is chosen by the Sheather and

Jones (1991) procedure. To choose an appropriate value of α and β required in expression (4.38), we minimize the following discrepancy measure:

$$\sum_{i=1}^n \left[\hat{f}_X(X_i; \hat{b}_0, \hat{\lambda}) - \hat{f}_{bc}(X_i; \hat{g}, \tilde{a}, \tilde{b}) \right]^2. \quad (4.39)$$

For the simulation study presented in Chapter 5, we restricted the values of α and β to $1 \leq \alpha \leq 1.5$ and $1 \leq \beta \leq 1.5$.

Step 5 The semi-parametric transformed data and grid are defined as

$$\begin{aligned} Y_1 &= \Phi^{-1} \left[\hat{F}_{Y_0}(Y_0; \hat{h}_0) \right] \quad \text{and} \\ y_1 &= \Phi^{-1} \left[\hat{F}_{Y_0}(y_0; \hat{h}_0) \right], \end{aligned} \quad (4.40)$$

respectively, which yields the newly proposed 1-step TKDE

$$\hat{f}_X(x; \hat{b}_1, \hat{h}_0, \hat{\lambda}) = \frac{\hat{f}_{Y_0}(y_0; \hat{h}_0)}{\phi \left(\Phi^{-1} \left[\hat{F}_{Y_0}(y_0; \hat{h}_0) \right] \right)} \left(\frac{dy_0}{dx} \right) \hat{f}_{Y_1}(y_1; \hat{b}_1). \quad (4.41)$$

Note that the adaptation scheme applied for the density and distribution function estimates is similar to that described in Step 4. However, we used $\alpha = \beta = 1$ to estimate the derivative in expression (4.41) and replaced $\hat{f}_X(X_i; \hat{b}_0, \hat{\lambda})$ with $\hat{f}_X(X_i; \hat{b}_1, \hat{h}_0, \hat{\lambda})$ to calculate the values of α and β required for the density estimate $\hat{f}_{Y_1}(y_1; \hat{b}_1)$ in expression (4.41). Also, \hat{h}_0 and \hat{b}_1 , above are chosen according to the procedures of Polansky (1997) and Sheather and Jones (1991) respectively.

Step 6 The semi-parametric transformation can be iterated to obtain an l -step TKDE by repeating steps 4 and 5.

5

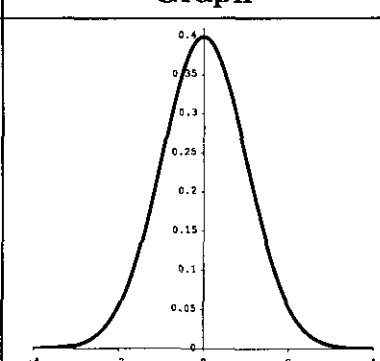
Empirical studies

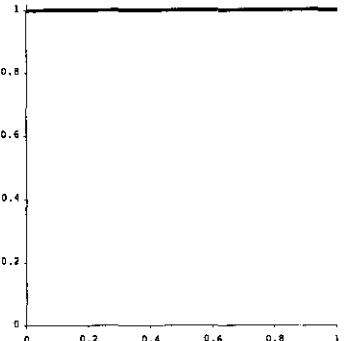
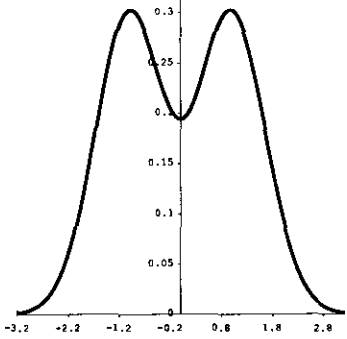
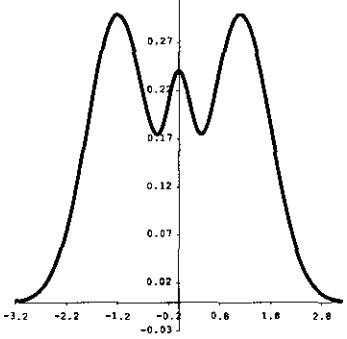
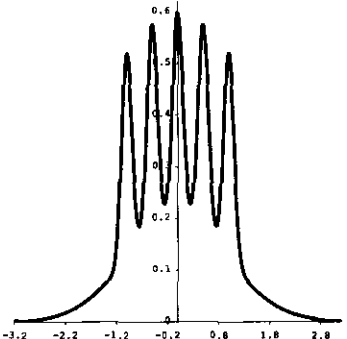
In this chapter we present a Monte Carlo simulation study to evaluate the performance of the newly proposed semi-parametric TKDE's. Real-life applications are also presented. The Fortran code and real-life data can be found on the CD-ROM attached to the back cover of this dissertation.

5.1 Simulation study

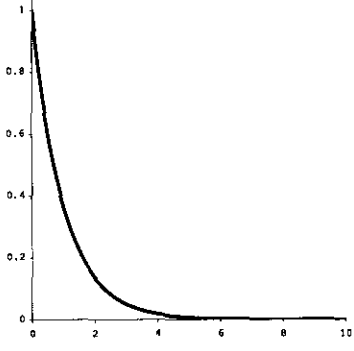
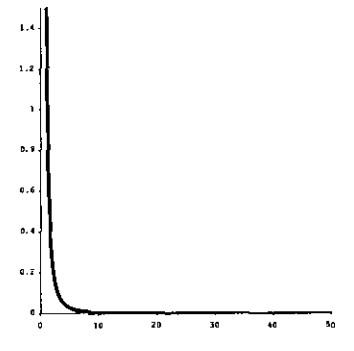
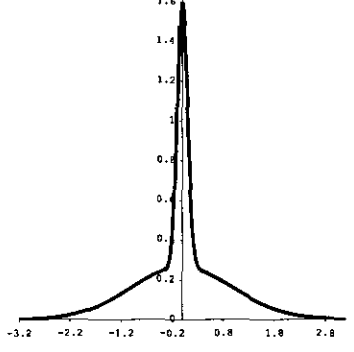
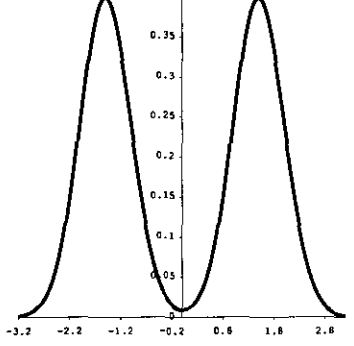
The densities considered in this simulation study are summarized in Table 5.1. The parameter choice considered (where applicable) are indicated in the first column.

Table 5.1: Densities considered in Monte Carlo simulation study

Density	$f_X(x) =$	Graph
1) Standard Normal $\mu = 0$ $\sigma = 1$	$\frac{1}{\sqrt{2\pi\sigma}} e^{-0.5[(x-\mu)/\sigma]^2}$	

Density	$f_X(x) =$	Graph
2) Uniform	$1,$ where $0 \leq x \leq 1.$	
3) Bimodal	$\frac{1}{2}N\left(-1, \left(\frac{2}{3}\right)^2\right)$ $+ \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right).$	
4) Trimodal	$\frac{9}{20}N\left(-\frac{6}{5}, \left(\frac{3}{5}\right)^2\right)$ $+ \frac{9}{20}N\left(\frac{6}{5}, \left(\frac{3}{5}\right)^2\right)$ $+ \frac{1}{10}N\left(0, \left(\frac{1}{4}\right)^2\right).$	
5) Claw	$\frac{1}{2}N(0, 1)$ $+ \sum_{i=0}^4 \frac{1}{10}N\left(\frac{i}{2} - 1, \left(\frac{1}{10}\right)^2\right).$	

Density	$f_X(x) =$	Graph
6) Skewed Bimodal	$\frac{3}{4}N(0, 1) + \frac{1}{4}N\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right).$	
7) Skewed Unimodal	$\begin{aligned} &\frac{1}{5}N(0, 1) \\ &+ \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) \\ &+ \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right). \end{aligned}$	
8) Weibull $\alpha = 1$ $\beta = 1.5$	$\beta\alpha^{-\beta}x^{\beta-1}e^{-(x/\alpha)^\beta},$ <p>where $\alpha > 0, \beta > 0, x \geq 0$.</p>	
9) Lognormal $\mu = 0$ $\sigma = 1$	$\frac{1}{\sqrt{2\pi}x\sigma}e^{-0.5[(\ln x - \mu)/\sigma]^2},$ <p>where $x > 0$.</p>	

Density	$f_X(x) =$	Graph
10) Exponential $\lambda = 1$	$\lambda e^{-\lambda x},$ where $\lambda > 0, x \geq 0.$	
11) Strict Pareto $\alpha = 1.5$	$\alpha x^{-\alpha-1},$ where $\alpha > 0, x \geq 1.$	
12) Kurtotic Unimodal	$\frac{2}{3}N(0, 1)$ $+ \frac{1}{3}N\left(0, \left(\frac{1}{10}\right)^2\right).$	
13) Separated Bimodal	$\frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$ $+ \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right).$	

The normal mixture densities were introduced by Marron and Wand (1992) and is considered to be extremely difficult to estimate using the ordinary KDE. The standard normal density is included to test the performance of the semi-parametric TKDE's in cases where little or no transformation is required. The uniform density is included since this density has a low kurtosis and has a bounded support. Hence, the usual kernel density estimate suffers from boundary bias effects. The bimodal, trimodal, claw and skewed bimodal densities are included since these densities contain more than one mode. The separated bimodal density is included as an example where all proposed *parametric* transformations fail drastically in transforming this data to normality. The skewed bimodal, weibull, standard exponential, standard lognormal and strict pareto densities are skewed to the left or right with potentially long tails. In addition, most of these densities have a bounded support. Hence, for these densities the performance of the semi-parametric TKDE's will be tested in the tail regions where spurious bumps can occur. Also, boundary bias occurs for the exponential and strict-pareto densities. The kurtotic unimodal density is included since this density has a high kurtosis, i.e., a high peak and heavy tails. Here we will test the ability of the semi-parametric TKDE's to estimate the density in the tail regions without spurious bumps, whilst capturing the peak of the density.

From the discussion presented above it should be clear that the densities selected covers a wide range of density forms and more importantly, the semi-parametric TKDE's will be subjected to data, where both boundary bias and spurious bumps in the tails can occur. The competitive density estimators are the:

- ordinary KDE (see (2.1)) denoted by ODE,
- TKDE proposed by Wand et al. (1991) (see Section 4.1) denoted by W-M-R,
- TKDE proposed by Yang and Marron (1999) (see Section 4.1) denoted by M-Y,
- adaptive KDE proposed by Abramson (1982) (see (2.49)) denoted by ADAP.

It should be noted that the TKDE procedures proposed by Wand et al. (1991) and Yang and Marron (1999) also suffer from the potential explosive behavior observed for the newly proposed semi-parametric TKDE's (see Section 4.2). For this reason we also added a constant shift δ (see (4.37)) to the input data. The constant shift δ is based on the procedure used in Step 4 of our stepwise procedure, introduced in Section 4.2.

However, no bandwidth adaptation was performed for these two density estimators. In order to assess the influence of iteration on the semi-parametric TKDE we consider zero, one and two iterations of the newly proposed semi-parametric TKDE. The procedure used to implement this method can be found in Section 4.2. Hence, our candidates will be the:

- semi-parametric TKDE (see Section 4.2) without iteration denoted by SEMI_0,
- semi-parametric TKDE with one iteration denoted by SEMI_1,
- semi-parametric TKDE with two iterations denoted by SEMI_2.

The abbreviations used for the transformations from the black box are summarized in Table 5.2. The abbreviations used for the transformation parameter estimation techniques

Table 5.2: Abbreviations for the transformations from the black box

Transformation	Abbreviation
Shifted Box-Cox	SBC
Yeo-Johnson	Y-J
Johnson family ($\gamma = 1, J = +1$)	JJ(+1)
Johnson family ($\gamma = 1, J = -1$)	JJ(-1)
Johnson family ($\gamma = 2$)	JU
Johnson family ($\gamma = 3$)	JB
Ruppert-Wand	R-W
Identity	NONE

are summarized in Table 5.3. For all the densities (except the claw density) considered, the Monte Carlo study was performed for sample sizes $n = 100$, $n = 200$ and $n = 500$. Sample sizes considered for the claw density were $n = 500$, $n = 800$ and $n = 1000$. The Monte Carlo repetition number used was $MC = 200$. A general fixed grid was constructed between the minimum and maximum data values observed after pooling all the Monte Carlo samples. For each density estimator an average value (over all the Monte Carlo samples) was obtained at each grid point, enabling us to assess the performance of the estimators with regard to bias graphically. However, the ISE (integrated squared error)

Table 5.3: Abbreviations for the parameter estimation techniques

Estimation technique	Abbreviation
Profile maximum likelihood	ML
Minimum residual	MR
Minimum distance	MD

was calculated between the minimum and maximum data value of each Monte Carlo sample. Hence, let X_1, \dots, X_n be i.i.d. random variables with associated order statistics $X_{(1)}, \dots, X_{(n)}$. Also, let $\hat{f}_{c,i}(x)$ be any of the candidate density estimates evaluated for the i -th Monte Carlo sample, $i = 1, \dots, MC$. We then define:

$$\widehat{ISE} [\hat{f}_{c,i}(\cdot)] = \int_{X_{(1)}}^{X_{(n)}} (\hat{f}_{c,i}(x) - f(x))^2 dx, \quad i = 1, \dots, MC, \quad (5.1)$$

which will be calculated for each Monte Carlo repetition. The Monte Carlo estimate of the MISE (mean integrated squared error) of the candidate density estimate \hat{f}_c is then calculated as follows:

$$M\widehat{ISE} [\hat{f}_c(\cdot)] = \frac{1}{MC} \sum_{i=1}^{MC} \widehat{ISE} [\hat{f}_{c,i}(\cdot)]. \quad (5.2)$$

The Monte Carlo estimate of SE (standard error) is calculated as follows:

$$\widehat{SE} = \frac{1}{\sqrt{MC}} \sqrt{\frac{1}{MC-1} \sum_{i=1}^{MC} \left[\widehat{ISE} [\hat{f}_{c,i}(\cdot)] - \frac{1}{MC} \left(\sum_{i=1}^{MC} \widehat{ISE} [\hat{f}_{c,i}(\cdot)] \right) \right]^2}. \quad (5.3)$$

For each Monte Carlo trial the Shapiro-Wilk p-value was recorded for the transformed data obtained from the selected parametric transformation from the black box of transformations. Recall that the transformation selection procedure is described in the stepwise algorithm presented in Section 4.2. Let p_1, \dots, p_{MC} denote these p-values. The average Monte Carlo p-value was calculated according to

$$\bar{p}_{MC} = \frac{1}{MC} \sum_{i=1}^{MC} p_i, \quad (5.4)$$

with associated standard error calculated according to

$$\sqrt{\frac{\bar{p}_{MC} (1 - \bar{p}_{MC})}{MC}}. \quad (5.5)$$

For each of the densities considered, the following output will be presented:

- The Monte Carlo MISE and SE for each of the densities considered, calculated according to (5.2) and (5.3). See Tables 5.4, 5.9, 5.14, 5.19, 5.24, 5.29, 5.34, 5.39, 5.44, 5.49, 5.54, 5.59 and 5.64.
- The average Monte Carlo Shapiro-Wilk p-value and the associated SE for the parametric transformed data. See Tables 5.5, 5.10, 5.15, 5.20, 5.25, 5.30, 5.35, 5.40, 5.45, 5.50, 5.55, 5.60 and 5.65.
- The percentage of times that each parametric transformation was selected according to the transformation selection procedure. See Tables 5.6, 5.11, 5.16, 5.21, 5.26, 5.31, 5.36, 5.41, 5.46, 5.51, 5.56, 5.61 and 5.66.
- The percentage of times that each parameter estimation technique was selected. See Tables 5.7, 5.12, 5.17, 5.22, 5.27, 5.32, 5.37, 5.42, 5.47, 5.52, 5.57, 5.62 and 5.67.
- Selective graphical output of the estimated densities. See Tables 5.8, 5.13, 5.18, 5.23, 5.28, 5.33, 5.38, 5.43, 5.48, 5.53, 5.58, 5.63 and 5.68.

Remarks:

- Each population density is indicated with a thick grey line in the graphical output.
- Figure 3 from Table 5.58 is a plot of all the (Monte Carlo average) density estimates, except the ordinary and adaptive KDE's, for which spurious bumps in the tail regions were evident.
- Several conclusions derived from the Monte Carlo studies are presented in Section 5.1.14.

5.1.1 Normal

Table 5.4: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	6.32	0.34	3.60	0.19	1.93	0.08
SEML0	6.98	0.37	4.05	0.21	2.21	0.09
SEML1	8.74	0.42	5.47	0.24	3.29	0.11
SEML2	9.91	0.46	6.40	0.26	3.89	0.13
ADAP	7.40	0.42	4.15	0.23	2.20	0.10
M-Y	6.90	0.38	3.78	0.20	1.90	0.08
W-M-R	6.42	0.35	3.67	0.19	1.94	0.08

Table 5.5: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.677	0.033	0.655	0.034	0.655	0.034

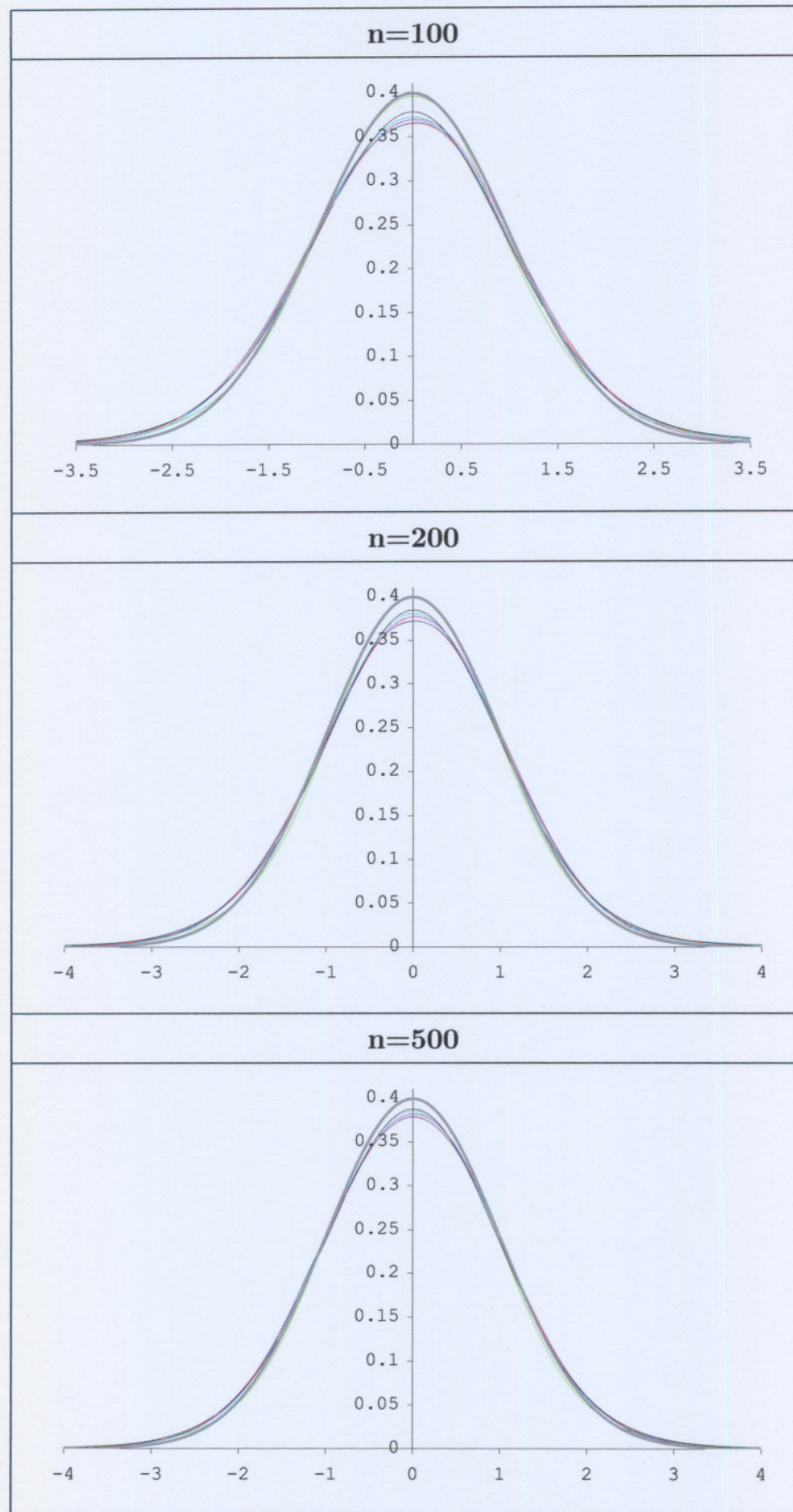
Table 5.6: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	5	2	2
JJ(-1)	14	16	10
JU	0	0	0
JB	39	47	56
Y-J	20	18	14
R-W	2	0	0
SBC	20	17	18
NONE	0	0	0

Table 5.7: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	52	48	39
MR	33	30	21
MD	15	22	40

Table 5.8: Density estimates



5.1.2 Uniform

Table 5.9: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	37.91	1.33	29.60	0.76	21.08	0.43
SEML0	28.06	1.80	15.92	0.84	9.90	0.45
SEML1	34.99	1.89	22.20	0.83	14.27	0.48
SEML2	39.96	2.00	26.18	0.88	16.73	0.48
ADAP	54.44	1.87	41.85	1.07	29.36	0.57
M-Y	30.00	1.88	16.32	0.77	9.18	0.32
W-M-R	39.29	1.40	30.37	0.79	21.32	0.43

Table 5.10: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.159	0.026	0.140	0.025	0.056	0.016

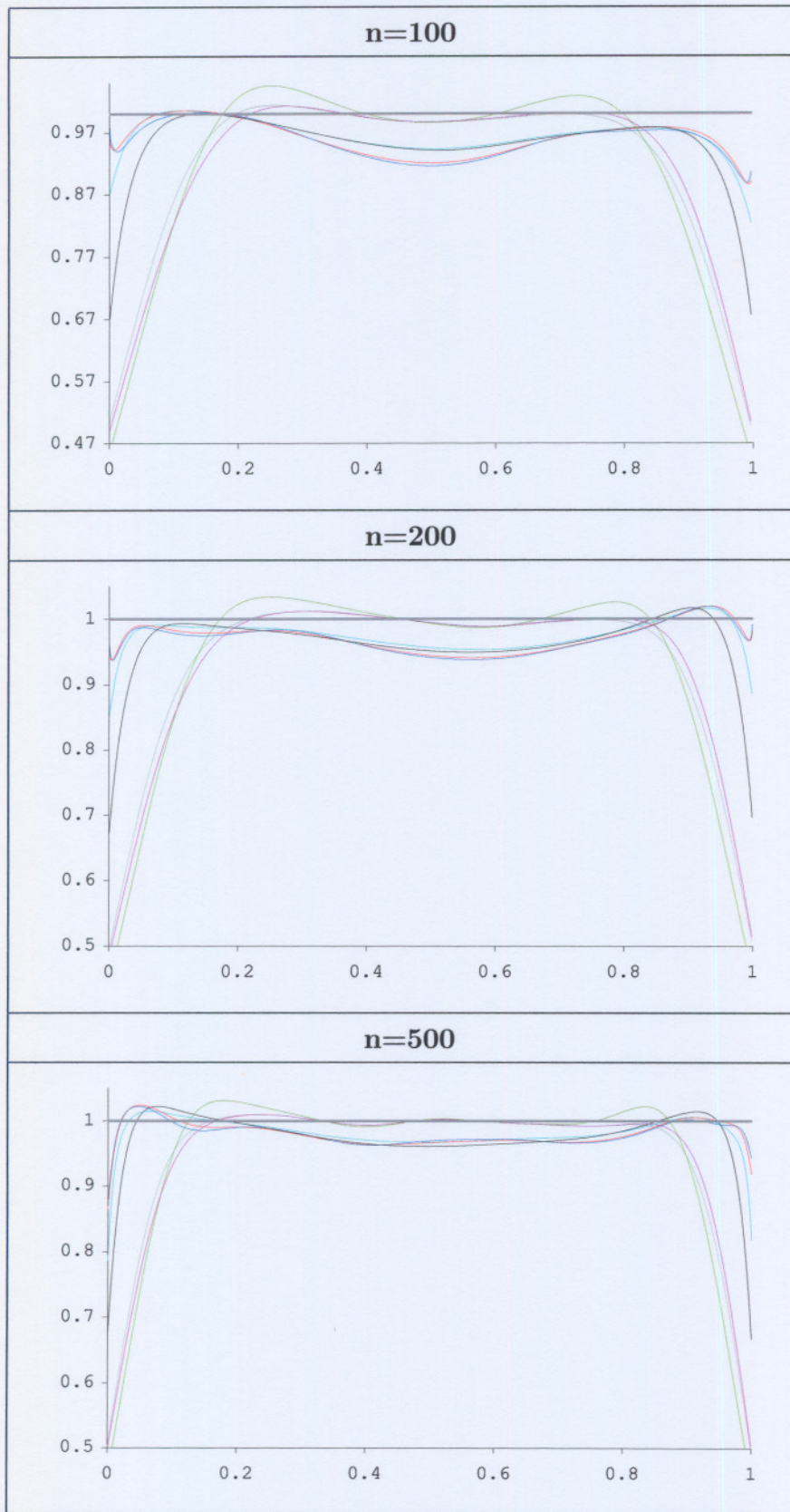
Table 5.11: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	0	0	0
JJ(-1)	5	0	0
JU	0	0	0
JB	88	96	93
Y-J	0	0	0
R-W	0	0	0
SBC	7	4	0
NONE	0	0	7

Table 5.12: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	73	83	87
MR	9	5	10
MD	18	12	3

Table 5.13: Density estimates



5.1.3 Bimodal

Table 5.14: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	7.85	0.32	4.60	0.19	2.40	0.09
SEML0	8.51	0.37	4.76	0.20	2.35	0.09
SEML1	8.38	0.38	5.20	0.20	2.94	0.11
SEML2	8.93	0.41	5.97	0.22	3.42	0.12
ADAP	8.59	0.42	5.14	0.23	2.51	0.11
M-Y	9.09	0.38	5.01	0.21	2.54	0.10
W-M-R	8.22	0.35	4.71	0.20	2.41	0.09

Table 5.15: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.149	0.025	0.077	0.019	0.041	0.014

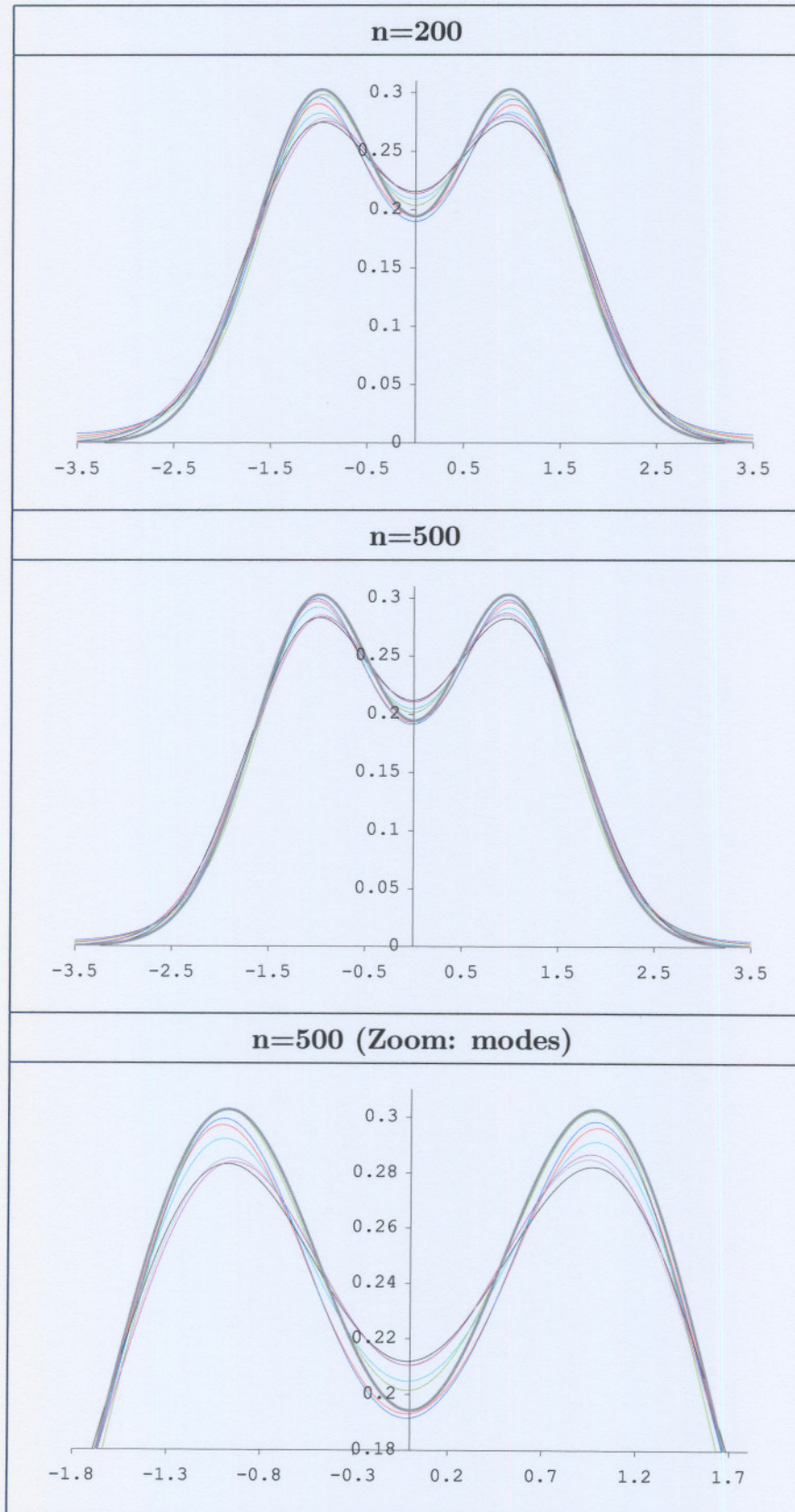
Table 5.16: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	0	0	0
JJ(-1)	7	1	0
JU	0	0	0
JB	85	96	99
Y-J	1	0	0
R-W	0	0	0
SBC	7	3	1
NONE	0	0	0

Table 5.17: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	83	88	90
MR	7	3	2
MD	10	9	8

Table 5.18: Density estimates



5.1.4 Trimodal

Table 5.19: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	9.13	0.30	6.07	0.21	3.26	0.10
SEML ₀	9.43	0.35	6.20	0.23	3.22	0.09
SEML ₁	9.14	0.33	6.14	0.23	3.32	0.10
SEML ₂	9.58	0.34	6.66	0.24	3.68	0.11
ADAP	9.99	0.39	6.20	0.25	3.03	0.10
M-Y	9.89	0.36	6.51	0.23	3.50	0.10
W-M-R	9.32	0.32	6.16	0.22	3.28	0.10

Table 5.20: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.073	0.018	0.047	0.015	0.002	0.003

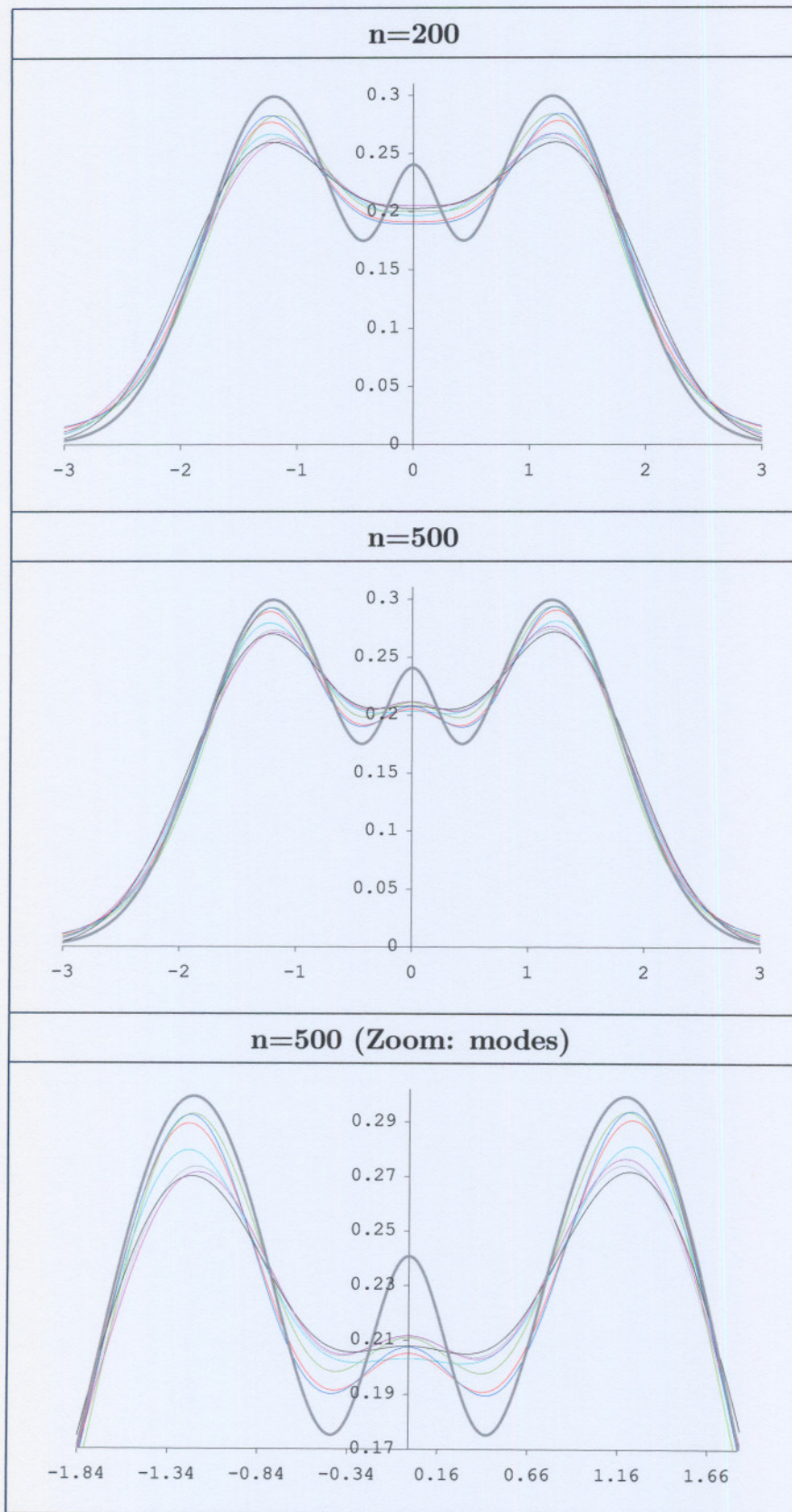
Table 5.21: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	0	0	0
JJ(-1)	5	0	0
JU	0	0	0
JB	87	97	94
Y-J	2	2	1
R-W	0	0	0
SBC	6	1	0
NONE	0	0	5

Table 5.22: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	86	89	92
MR	7	3	5
MD	7	8	3

Table 5.23: Density estimates



5.1.5 Claw

Table 5.24: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	42.74	0.25	37.57	0.29	34.64	0.30
SEML0	42.95	0.23	37.86	0.26	34.78	0.27
SEML1	34.32	0.35	23.07	0.35	17.78	0.29
SEML2	29.79	0.38	17.39	0.31	12.53	0.25
ADAP	41.10	0.34	33.28	0.38	28.94	0.37
M-Y	42.99	0.25	37.95	0.28	34.98	0.29
W-M-R	42.63	0.25	37.27	0.29	34.31	0.30

Table 5.25: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.447	0.050	0.359	0.048	0.284	0.045

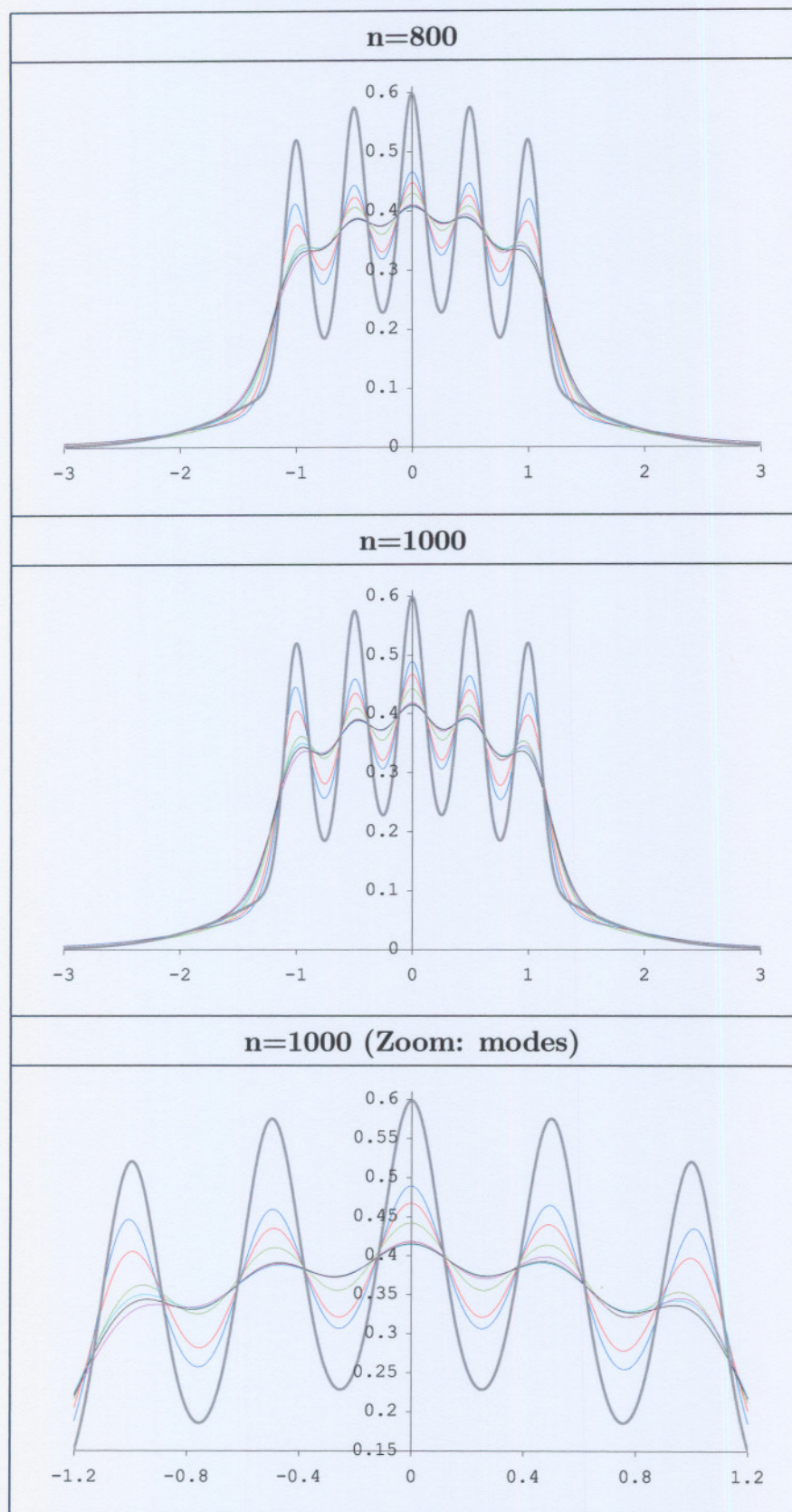
Table 5.26: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	0	0	0
JJ(-1)	10	14	5
JU	0	0	0
JB	67	67	81
Y-J	7	7	4
R-W	0	0	0
SBC	16	12	10
NONE	0	0	0

Table 5.27: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	14	11	5
MR	16	17	9
MD	70	72	86

Table 5.28: Density estimates



5.1.6 Skewed bimodal

Table 5.29: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	10.09	0.31	6.35	0.22	3.44	0.12
SEML ₀	9.79	0.34	5.85	0.23	3.00	0.11
SEML ₁	10.15	0.38	6.11	0.25	3.39	0.12
SEML ₂	10.68	0.42	6.71	0.27	3.88	0.13
ADAP	10.86	0.43	6.38	0.27	3.20	0.13
M-Y	9.83	0.35	5.73	0.22	2.98	0.11
W-M-R	9.45	0.33	5.69	0.22	3.00	0.11

Table 5.30: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.171	0.027	0.080	0.019	0.015	0.009

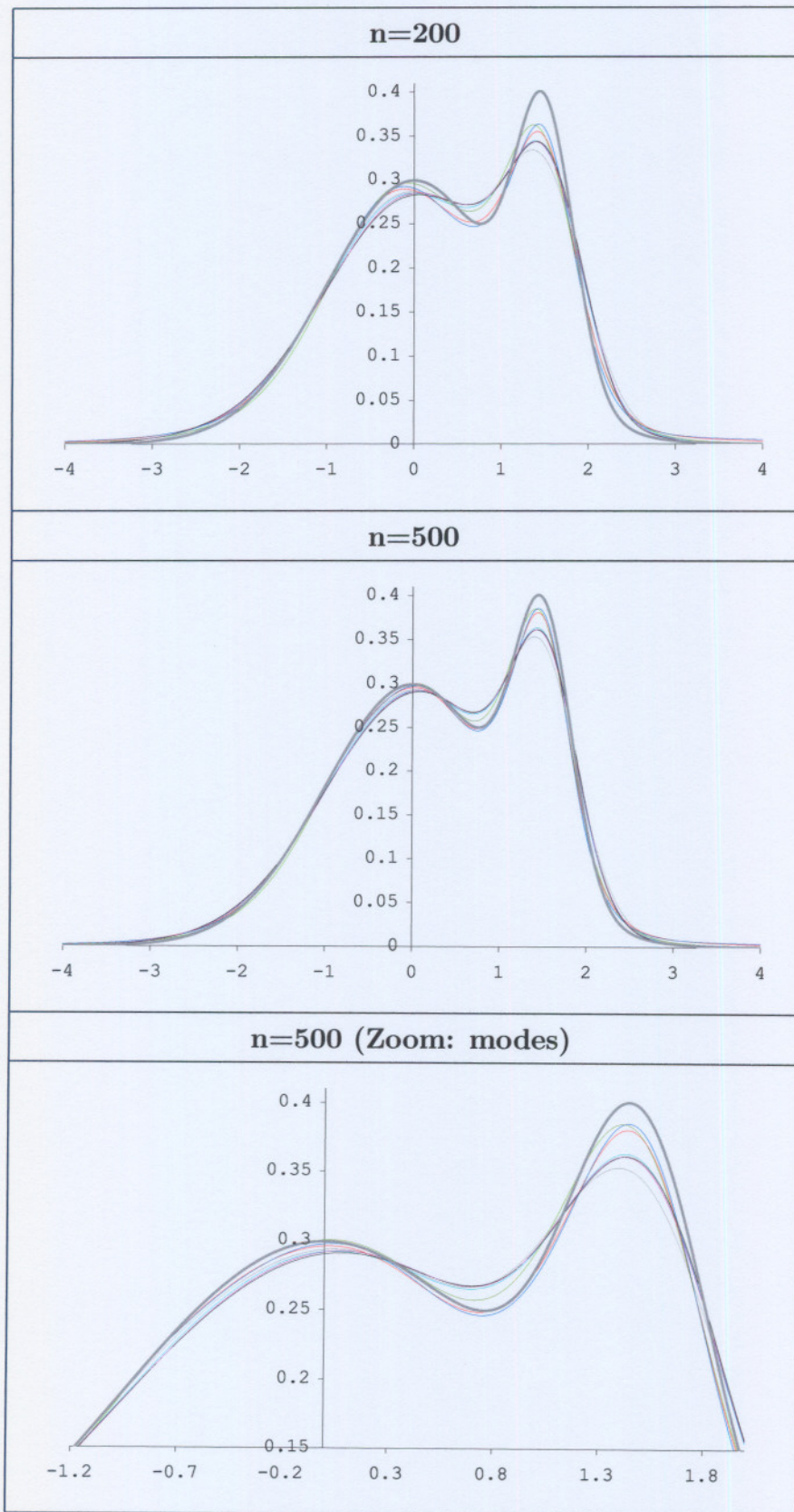
Table 5.31: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	0	0	0
JJ(-1)	50	58	81
JU	0	0	0
JB	43	39	16
Y-J	5	3	3
R-W	0	0	0
SBC	2	0	0
NONE	0	0	0

Table 5.32: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	71	70	76
MR	14	12	5
MD	15	18	19

Table 5.33: Density estimates



5.1.7 Skewed unimodal

Table 5.34: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	9.28	0.45	5.18	0.22	2.66	0.11
SEML0	8.91	0.47	5.21	0.23	2.66	0.12
SEML1	10.70	0.51	6.97	0.28	4.09	0.16
SEML2	12.19	0.55	8.19	0.32	4.91	0.18
ADAP	10.03	0.56	5.95	0.29	2.89	0.13
M-Y	9.97	0.51	5.47	0.24	2.77	0.12
W-M-R	8.65	0.44	4.75	0.21	2.50	0.12

Table 5.35: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.654	0.034	0.655	0.034	0.633	0.034

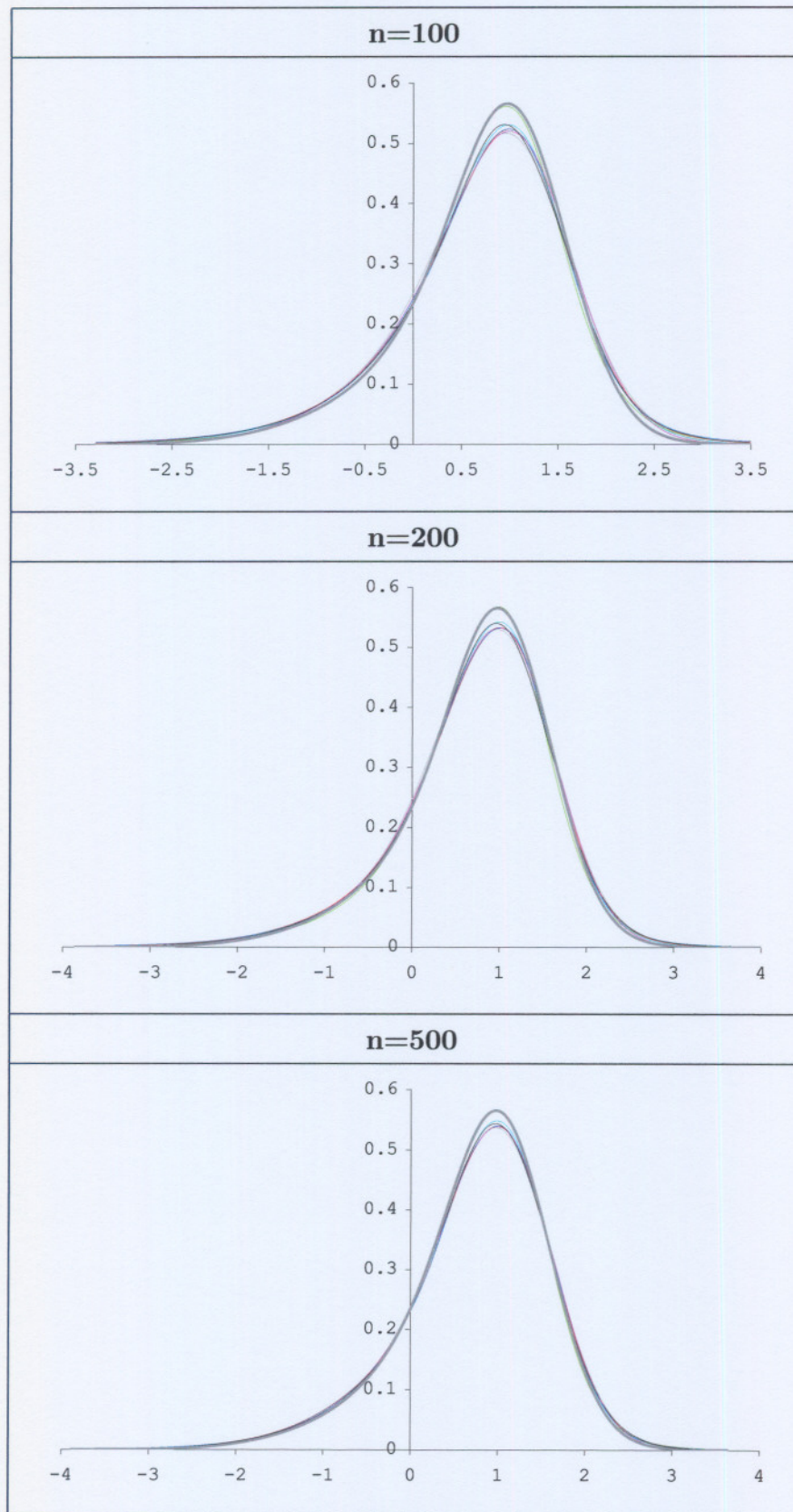
Table 5.36: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	0	0	0
JJ(-1)	47	41	52
JU	0	0	0
JB	1	0	0
Y-J	45	57	48
R-W	1	0	0
SBC	6	2	0
NONE	0	0	0

Table 5.37: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	43	47	32
MR	40	37	29
MD	17	16	39

Table 5.38: Density estimates



5.1.8 Weibull

Table 5.39: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	14.00	0.67	8.69	0.36	4.61	0.18
SEML0	11.52	0.62	8.09	0.40	4.03	0.17
SEML1	14.77	0.70	11.36	0.49	6.25	0.22
SEML2	17.34	0.76	13.56	0.55	7.53	0.25
ADAP	17.70	0.82	11.97	0.50	6.70	0.25
M-Y	11.46	0.71	7.25	0.39	3.28	0.15
W-M-R	11.46	0.63	7.73	0.41	3.47	0.15

Table 5.40: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.429	0.035	0.315	0.033	0.205	0.029

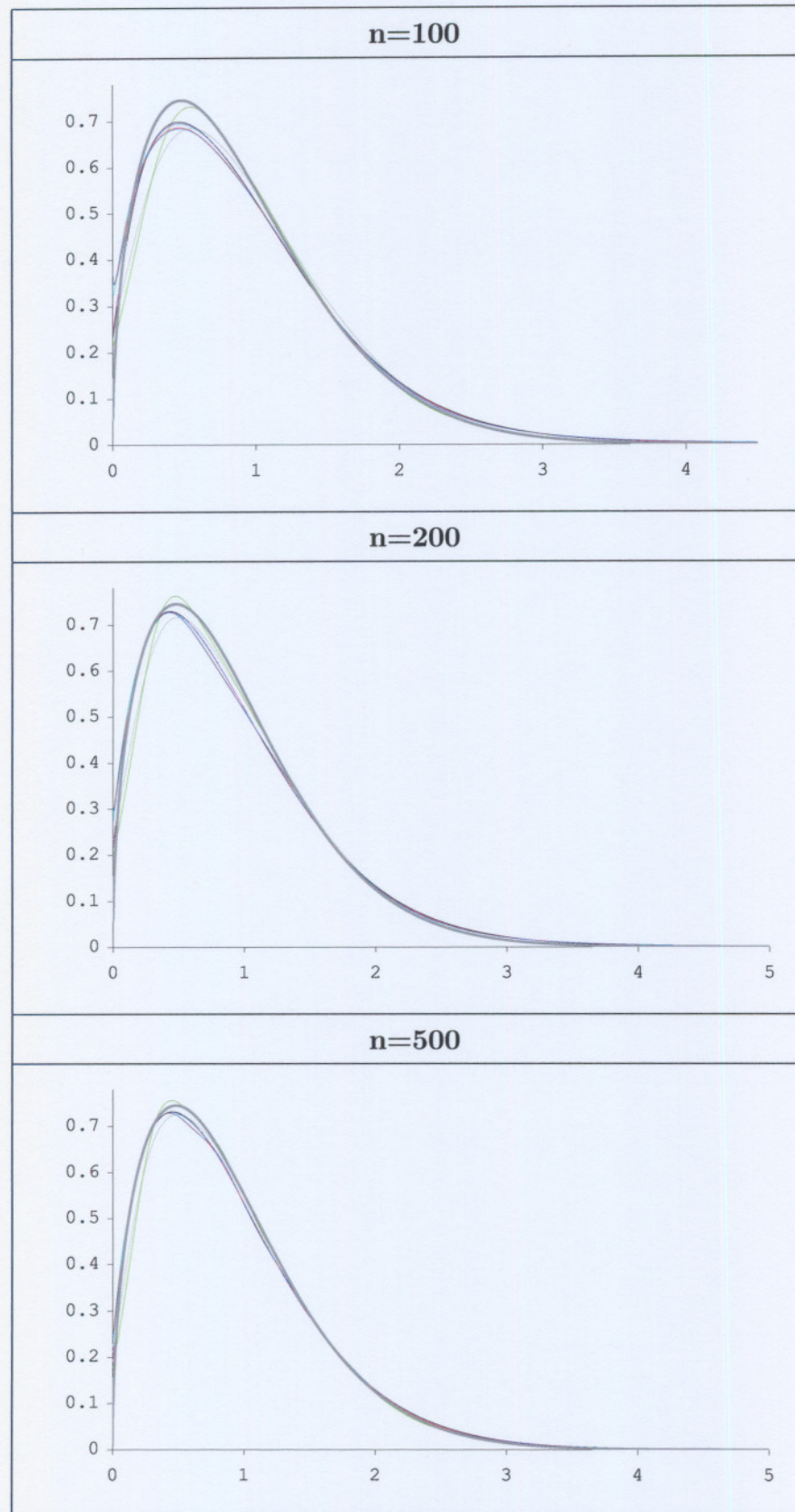
Table 5.41: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	11	4	1
JJ(-1)	0	0	0
JU	0	0	0
JB	0	0	0
Y-J	0	0	0
R-W	0	0	0
SBC	89	96	99
NONE	0	0	0

Table 5.42: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	5	2	0
MR	76	74	85
MD	19	24	15

Table 5.43: Density estimates



5.1.9 Lognormal

Table 5.44: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	19.32	0.76	11.35	0.38	6.01	0.18
SEML0	9.44	0.53	5.95	0.29	3.08	0.13
SEML1	11.44	0.55	7.65	0.34	4.47	0.17
SEML2	13.20	0.61	8.97	0.37	5.44	0.19
ADAP	18.19	0.79	10.70	0.39	5.86	0.19
M-Y	8.83	0.49	5.67	0.30	2.82	0.13
W-M-R	9.31	0.50	5.85	0.36	2.67	0.13

Table 5.45: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.618	0.034	0.584	0.035	0.602	0.035

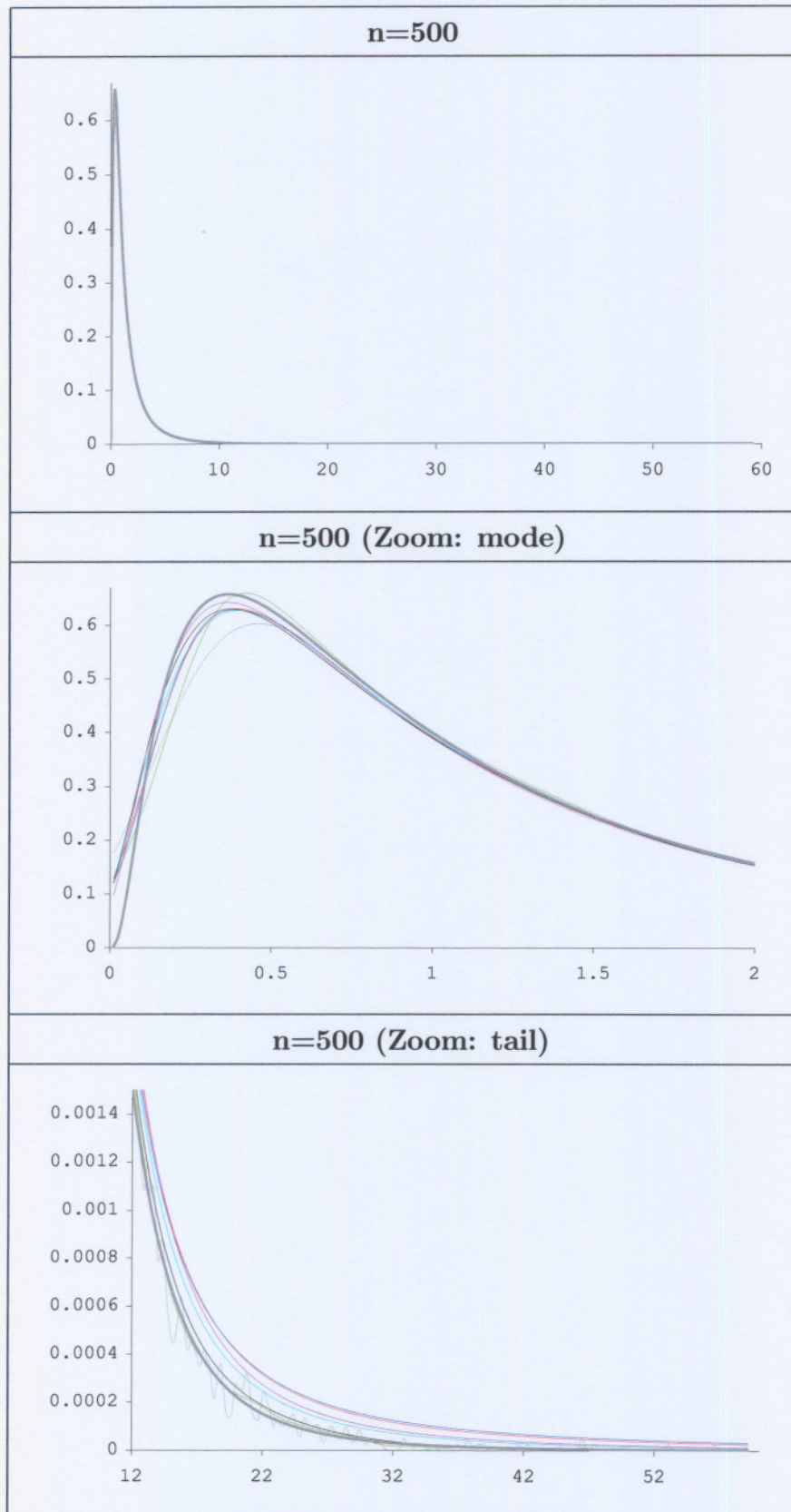
Table 5.46: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	45	39	27
JJ(-1)	0	0	0
JU	0	0	0
JB	0	0	0
Y-J	0	0	0
R-W	0	0	0
SBC	55	61	73
NONE	0	0	0

Table 5.47: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	32	26	16
MR	44	49	55
MD	24	25	29

Table 5.48: Density estimates



5.1.10 Exponential

Table 5.49: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	38.28	1.06	29.10	0.78	20.15	0.39
SEML0	11.93	0.79	7.46	0.40	3.71	0.17
SEML1	15.14	0.90	10.05	0.47	5.80	0.23
SEML2	17.97	0.99	12.07	0.54	7.06	0.26
ADAP	37.66	1.15	27.91	0.80	19.16	0.38
M-Y	14.51	0.98	9.90	0.48	5.92	0.20
W-M-R	15.58	0.93	10.37	0.53	6.01	0.23

Table 5.50: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.418	0.035	0.320	0.033	0.195	0.028

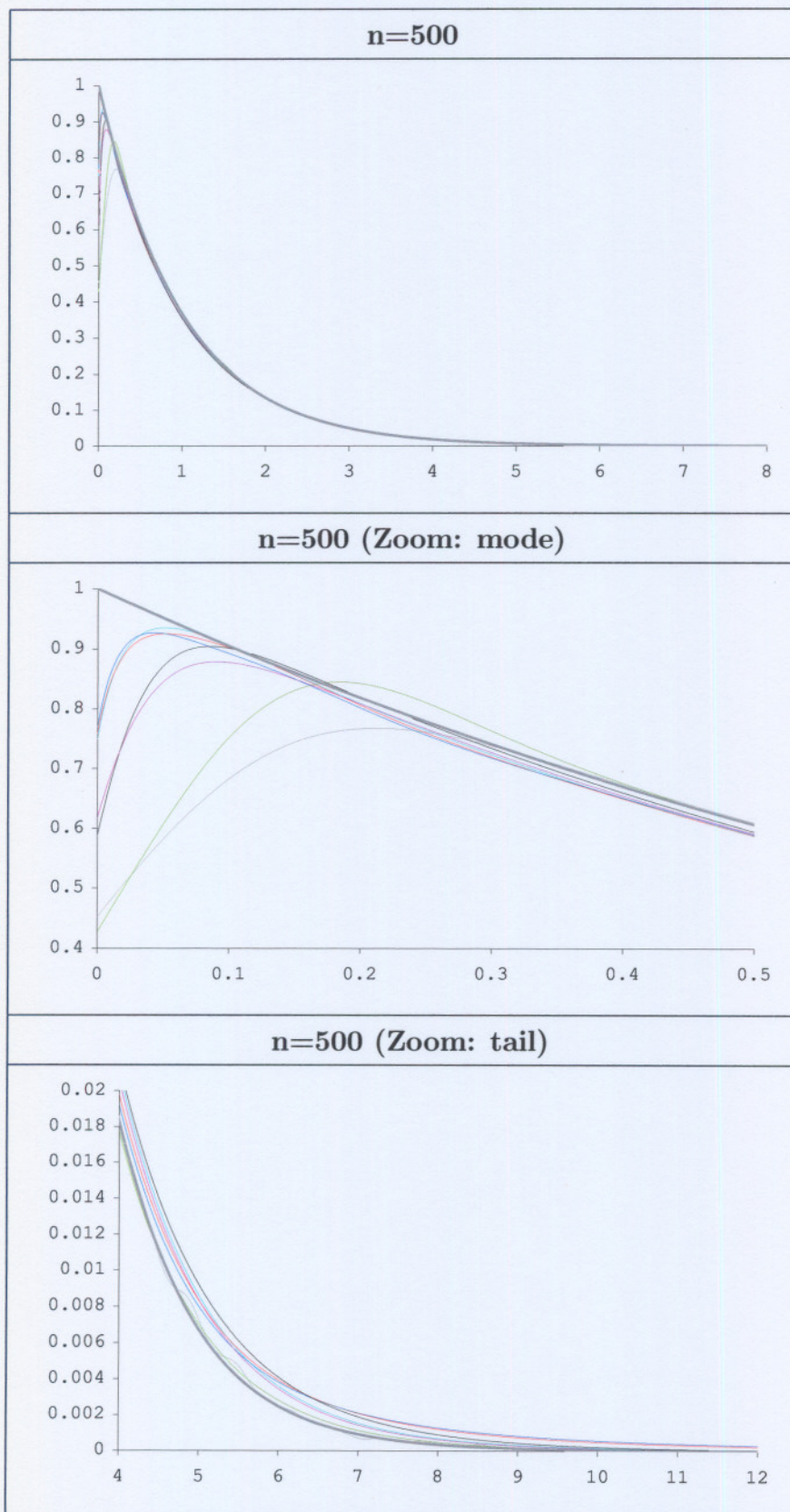
Table 5.51: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	8	3	0
JJ(-1)	0	0	0
JU	0	0	0
JB	0	0	0
Y-J	0	0	0
R-W	0	0	0
SBC	92	97	100
NONE	0	0	0

Table 5.52: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	4	0	0
MR	73	79	83
MD	23	21	17

Table 5.53: Density estimates



5.1.11 Strict-Pareto

Table 5.54: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	81.12	1.82	93.70	1.00	257.97	2.16
SEML0	28.30	1.27	32.15	1.43	85.78	3.09
SEML1	30.69	1.37	34.38	1.57	85.55	3.60
SEML2	33.23	1.50	35.72	1.77	83.30	4.28
ADAP	68.93	1.48	91.02	1.36	254.32	2.81
M-Y	30.90	1.18	38.36	1.34	107.22	3.01
W-M-R	31.43	1.41	40.98	1.62	128.47	3.60

Table 5.55: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.643	0.034	0.652	0.034	0.588	0.035

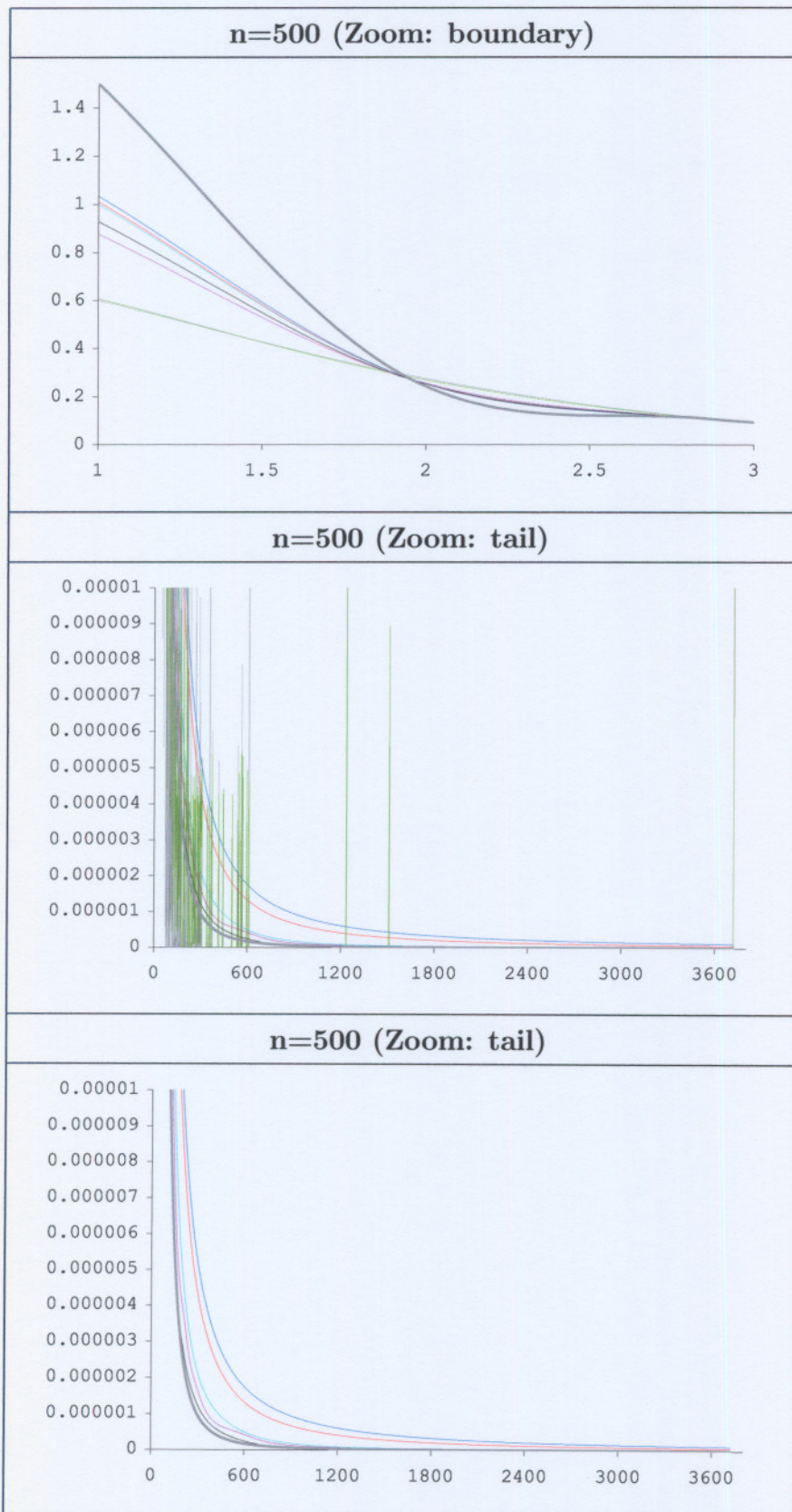
Table 5.56: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	56	43	33
JJ(-1)	0	0	0
JU	0	0	0
JB	0	0	0
Y-J	0	0	0
R-W	0	0	0
SBC	44	57	67
NONE	0	0	0

Table 5.57: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	43	36	32
MR	33	34	25
MD	24	30	43

Table 5.58: Density estimates



5.1.12 Kurtotic unimodal

Table 5.59: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	66.49	2.67	36.89	1.27	16.87	0.56
SEML0	56.92	2.62	27.85	1.16	10.69	0.44
SEML1	47.16	2.46	22.02	1.03	9.95	0.38
SEML2	40.83	2.18	19.20	0.89	9.80	0.35
ADAP	40.19	2.00	20.18	0.79	10.60	0.34
M-Y	54.99	2.53	30.78	1.19	13.78	0.60
W-M-R	66.99	2.69	36.84	1.27	16.91	0.56

Table 5.60: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.147	0.025	0.042	0.014	0.000	0.001

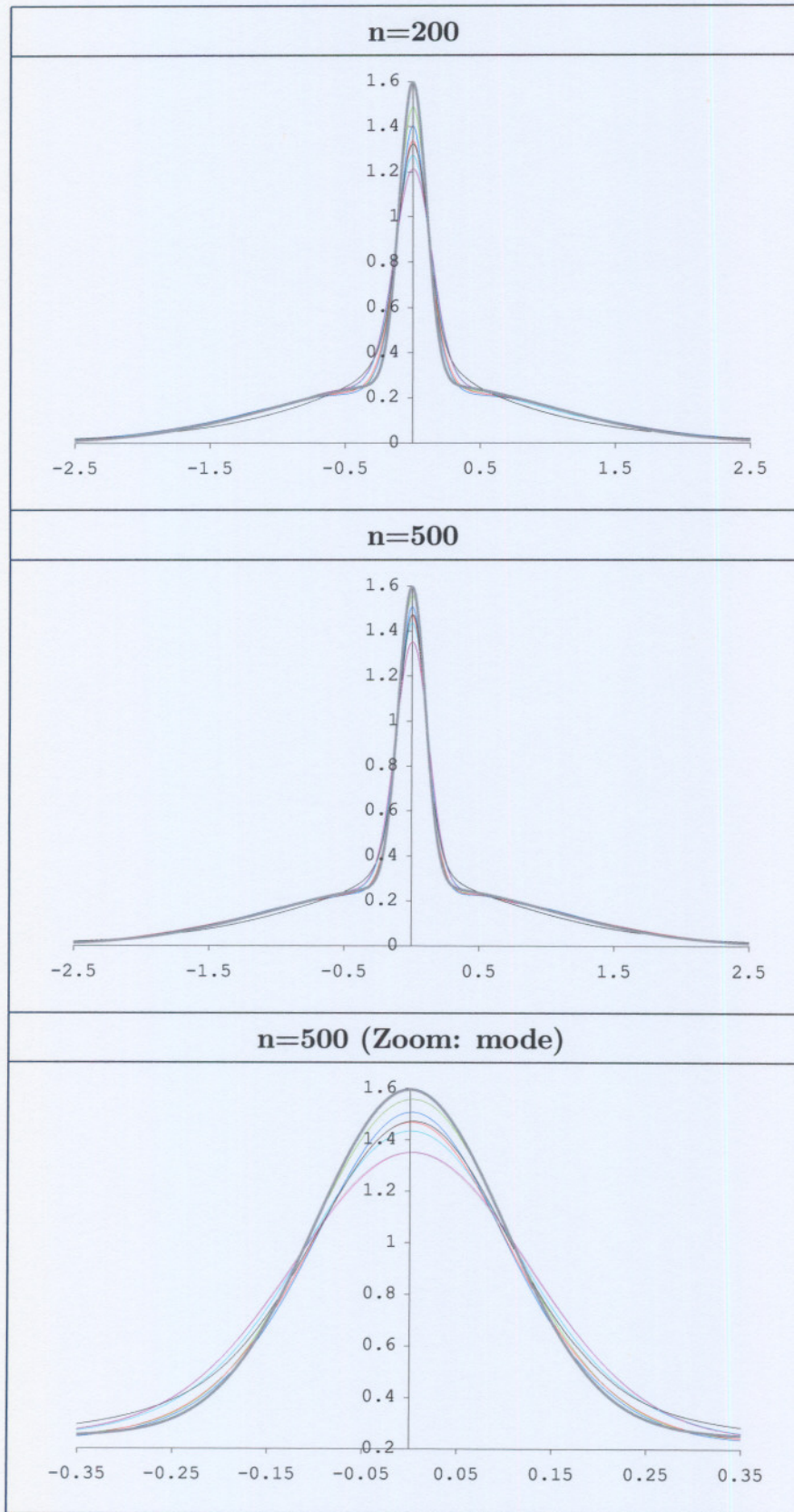
Table 5.61: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	1	1	0
JJ(-1)	5	0	0
JU	0	0	0
JB	0	0	0
Y-J	8	2	0
R-W	85	97	100
SBC	1	0	0
NONE	0	0	0

Table 5.62: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	93	97	100
MR	5	2	0
MD	2	1	0

Table 5.63: Density estimates



5.1.13 Separated bimodal

Table 5.64: Mean Integrated Squared Error ($\times 10^3$)

Method	n=100		n=200		n=500	
	MISE	SE	MISE	SE	MISE	SE
ODE	11.39	0.41	7.06	0.24	3.46	0.12
SEML0	15.95	0.71	6.75	0.22	3.42	0.12
SEML1	12.42	0.54	7.46	0.22	4.55	0.14
SEML2	12.72	0.52	8.08	0.24	4.99	0.15
ADAP	10.20	0.44	6.23	0.22	3.44	0.12
M-Y	16.20	0.62	8.23	0.31	3.68	0.13
W-M-R	15.16	0.74	7.23	0.25	3.50	0.12

Table 5.65: Shapiro-Wilk p-value

n=100		n=200		n=500	
p-value	SE	p-value	SE	p-value	SE
0.000	0.000	0.000	0.000	0.000	0.000

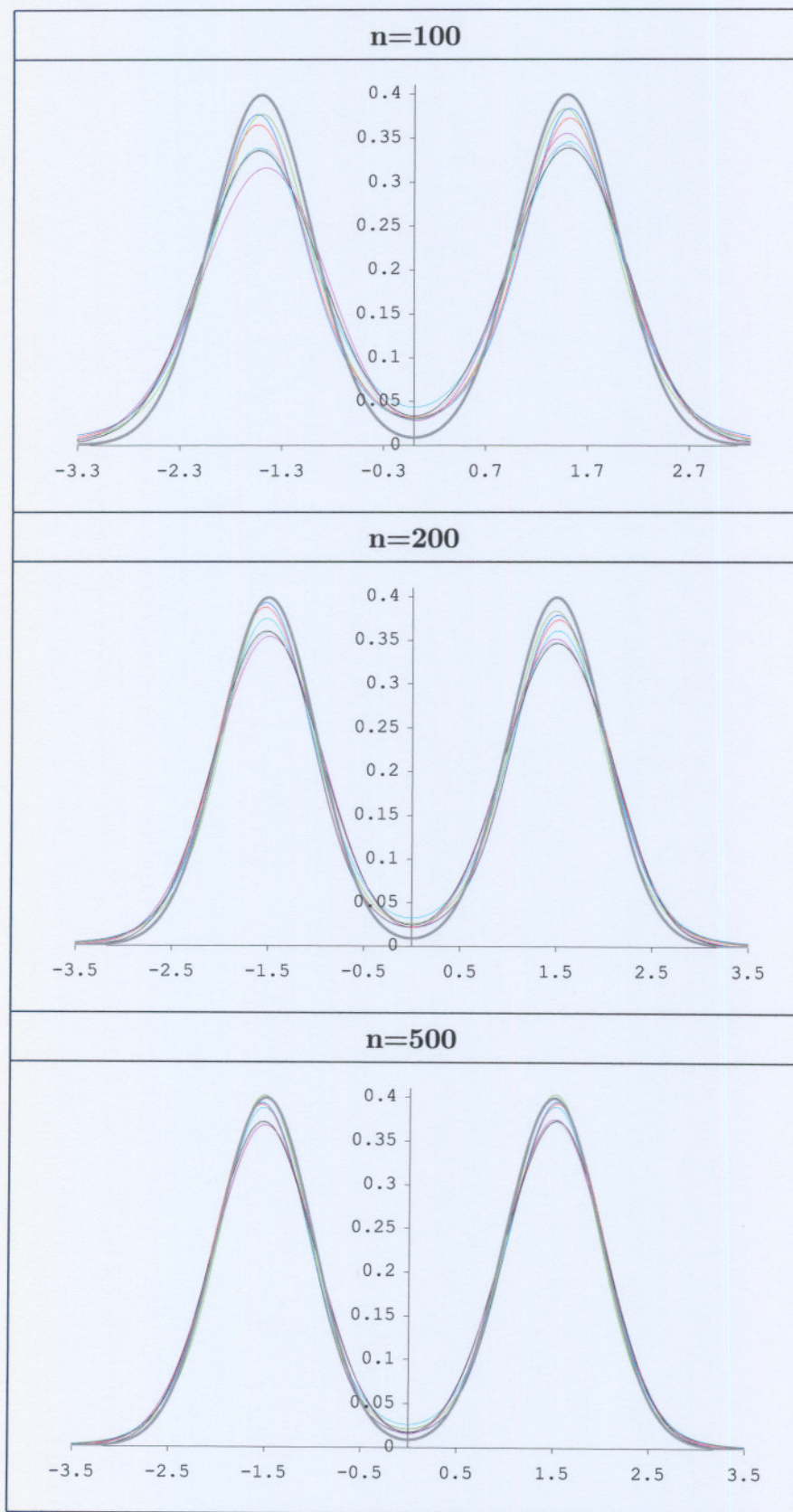
Table 5.66: Transformation selected

Transformation	% Selected		
	n=100	n=200	n=500
JJ(+1)	2	0	0
JJ(-1)	17	0	0
JU	0	0	0
JB	54	1	0
Y-J	5	0	0
R-W	0	0	0
SBC	19	0	0
NONE	3	99	100

Table 5.67: Parameter estimation

Estimation Method	% Selected		
	n=100	n=200	n=500
ML	59	0	0
MR	25	99	100
MD	16	1	0

Table 5.68: Density estimates



5.1.14 Conclusions

- The 0-step TKDE performs generally well when applied to data obtained from densities *without excessive curvature* such as the normal, uniform, skewed unimodal, weibull, lognormal, exponential and pareto densities. *We recommend the 0-step TKDE for application to unimodal and low-kurtosis densities.*
- The 1-step TKDE is suitable for *capturing density curvature*, hence detecting, for example, density modes. *We recommend this estimate for application to densities such as the bimodal, trimodal, claw, skewed bimodal, kurtotic unimodal and separated bimodal.*
- The performance of the 2-step TKDE is similar to that of the 1-step TKDE with the advantage that the 2-step procedure captures density curvature more profoundly.
- *Spurious bumps in the tail regions are significantly removed by all the newly proposed procedures.* The reader is referred to the graphical output of the following densities: skewed unimodal, lognormal, exponential and pareto. *It should be noted that the 0-step procedure seems to be most effective at estimating density tails.*
- *Boundary bias is addressed automatically.* The reader is referred to the graphical output of the following densities for visual confirmation: uniform, exponential and pareto.
- The somewhat larger MISE values observed for the 1-step and 2-step TKDE's for the bimodal, trimodal, skewed bimodal and separated bimodal densities are unexpected. However, empirical investigation confirms that these relatively larger MISE values are caused by only a few samples where the density peaks were excessively over estimated and the density valleys under estimated by the 1-step and 2-step TKDE's. To rectify this phenomenon the value c_a given in expression (4.32) should depend on the data. An open research problem is to construct an effective data-dependent choice of c_a . This will not be pursued in this dissertation. Also, more advanced procedures can be invented to select the values of α and β required to implement the newly proposed adaptation scheme given in expression (4.32). For densities with sharp peaks and/or valleys, for example the claw and kurtotic unimodal densities, the occurrence of this MISE performance is absent. *Nevertheless,*

the exceptional ability of the 1-step and 2-step TKDE's to capture density curvature is evident from the graphical output presented.

- The Shapiro-Wilk p-value indicates that the parametric transformation to normality of data from densities with curvature, i.e., the bimodal, trimodal and skewed bimodal is not successful. Data from the kurtotic unimodal (high kurtosis) and uniform (low kurtosis) densities were also not transformed successfully, using a parametric transformation. The parametric transformation to normality of data from the normal, skewed unimodal, weibull, lognormal, exponential and strict-pareto densities performed exceptionally well, yielding moderate to high p-values. It should also be noted that the parametric transformation failed drastically for the separated bimodal density. Surprisingly, the parametric transformation performed well for the claw density. The performance of the parametric transformation to normality is easily explained by inspection of the shape of these transformation functions. For a more detailed discussion of the above-mentioned issue, the reader is referred to Section 3.3.2. *It should be noted that the performance of the optimal transformation function (see 3.10) can assume all the possible shapes of a parametric transformation and any combination of these shapes, which results in high p-values when testing normality of the transformed data.* The reader is referred to Table 3.6 for a more comprehensive illustration of this remark.
- The transformation selection procedure (i.e., select the transformation that produces the highest Shapiro-Wilk p-value when testing for normality) performs well for all the densities considered. The dominant transformation functions selected for each density are summarized in Table 5.69. The reader is referred to Section 3.3.2 for a detailed discussion concerning the applicability of the parametric transformations considered. The message conveyed in Table 5.69 is that *the proposed transformation selection procedure frequently selects the correct transformation.*
- The profile maximum likelihood parameter estimation technique (ML) was selected predominantly for the densities containing some measure of curvature, i.e., bimodal, trimodal and skewed bimodal densities. Again, the claw density was an exception where the minimum distance (MD) procedure was preferred. In addition, the profile maximum likelihood parameter estimation technique (ML) also favours densities with high kurtosis, for example the kurtotic unimodal density and densities with

Table 5.69: Dominant transformation functions selected

Transformation function	Densities
SBC	exponential, weibull
SBC, JJ(+1)	lognormal, strict-pareto
JJ(-1), Y-J	skewed unimodal
SBC, Y-J, JB	normal
JB	uniform, bimodal, trimodal, claw
JB, JJ(-1)	skewed bimodal
R-W	kurtotic unimodal
NONE	separated bimodal

low kurtosis, for example the uniform density. For all the other densities considered, i.e., the unimodal densities, the minimum residual (MR) and minimum distance (MD) parameter estimation techniques were more dominant, with the MR technique selected more often. It is interesting to note that in all the cases were the profile maximum likelihood parameter estimation technique was dominant, low Shapiro-Wilk p-values were returned when testing for normality. *Hence, we conclude that the newly formulated minimum residual and minimum distance techniques contribute to better transformations to normality (according to the Shapiro-Wilk p-value), when a parametric transformation is applicable.*

- The Yang and Marron (1999) density estimator performs well in most cases. This density estimator can, however, be considered in the class of 0-step TKDE's.
- The Wand et al. (1991) density estimator performs well in cases where the density considered is unimodal skewed to the left or right. In all other cases this estimator has similar performance as that of the ordinary kernel density estimator. The Wand et al. (1991) density estimator can also be considered in the class of 0-step TKDE's.
- Perhaps the most competitive kernel density estimator is the adaptive kernel density estimator proposed by Abramson (1982). However, empirical studies suggest that this estimator sometimes seriously over estimates density peaks and under estimates density valleys.

5.2 Applications to real data

Throughout this section we consider the 0-step, 1-step and 2-step TKDE's (see Section 4.2). The same colour scheme used in the simulation study is utilized. Hence, our candidates are the following estimators:

- semi-parametric TKDE without iteration denoted by [SEMI.0](#).
- semi-parametric TKDE with one iteration denoted by [SEMI.1](#).
- semi-parametric TKDE with two iterations denoted by [SEMI.2](#).

5.2.1 Example 1: British income data

This data consist of 7201 British incomes, see Wand (1997), for the year 1975, and were divided by their sample average, yielding the observations X_1, \dots, X_n . Let

$$Z_{x,i} = \frac{X_i - 0.9188}{0.551}, \quad i = 1, \dots, 7201.$$

For this sample the scale invariant global roughness measure

$$L(\hat{\lambda}) = \hat{\sigma}_{y_0} R \left(f''_{Y_0}(\cdot; \hat{\lambda}) \right)^{1/5},$$

was minimized by the best transformation - parameter estimation combination. The shifted Box-Cox (SBC) transformation was selected with the minimum residual (MR) parameter estimation technique. The estimated parameter values are $\lambda_1 = 1.7459$ and $\lambda_2 = 0$, rendering the log transformation. The constant shift parameter δ (see 4.37) is 0.185. Table 5.70 displays the transformation function and resulting transformation function derivative for iterations 0, 1 and 2. The effect of the constant shift parameter, δ , is visible in both the transformation function and transformation function derivative, since the input data are mapped to a smaller domain and the derivative at the lower bound (potential explosive behaviour of the TKDE can occur at the lower bound) is smaller. The reader is referred to Section 4.2 for a more detailed discussion concerning the potential explosive behaviour of the TKDE.

Table 5.70: Transformation functions and transformation function derivatives

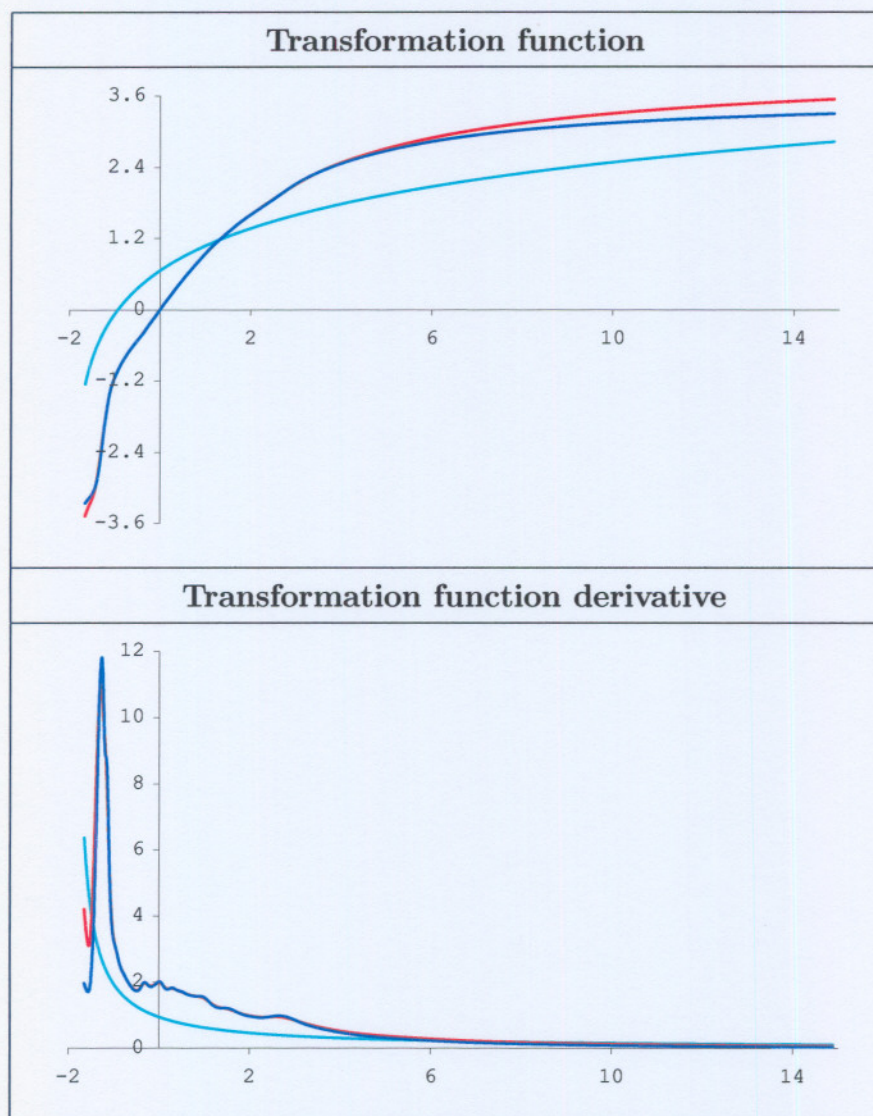


Table 5.71 displays the density estimates of the 0-step ([SEMI_0](#)), 1-step ([SEMI_1](#)) and 2-step ([SEMI_2](#)) semi-parametric TKDE's. The density estimates resemble the lognormal density, however, two modes are detected near the left bound. It is clear that the 1-step and 2-step procedures capture too much density curvature, while all three estimates perform well in the tail region. Hence, the 0-step semi-parametric TKDE should be preferred, which is displayed in Table 5.72.

Table 5.71: Density estimates

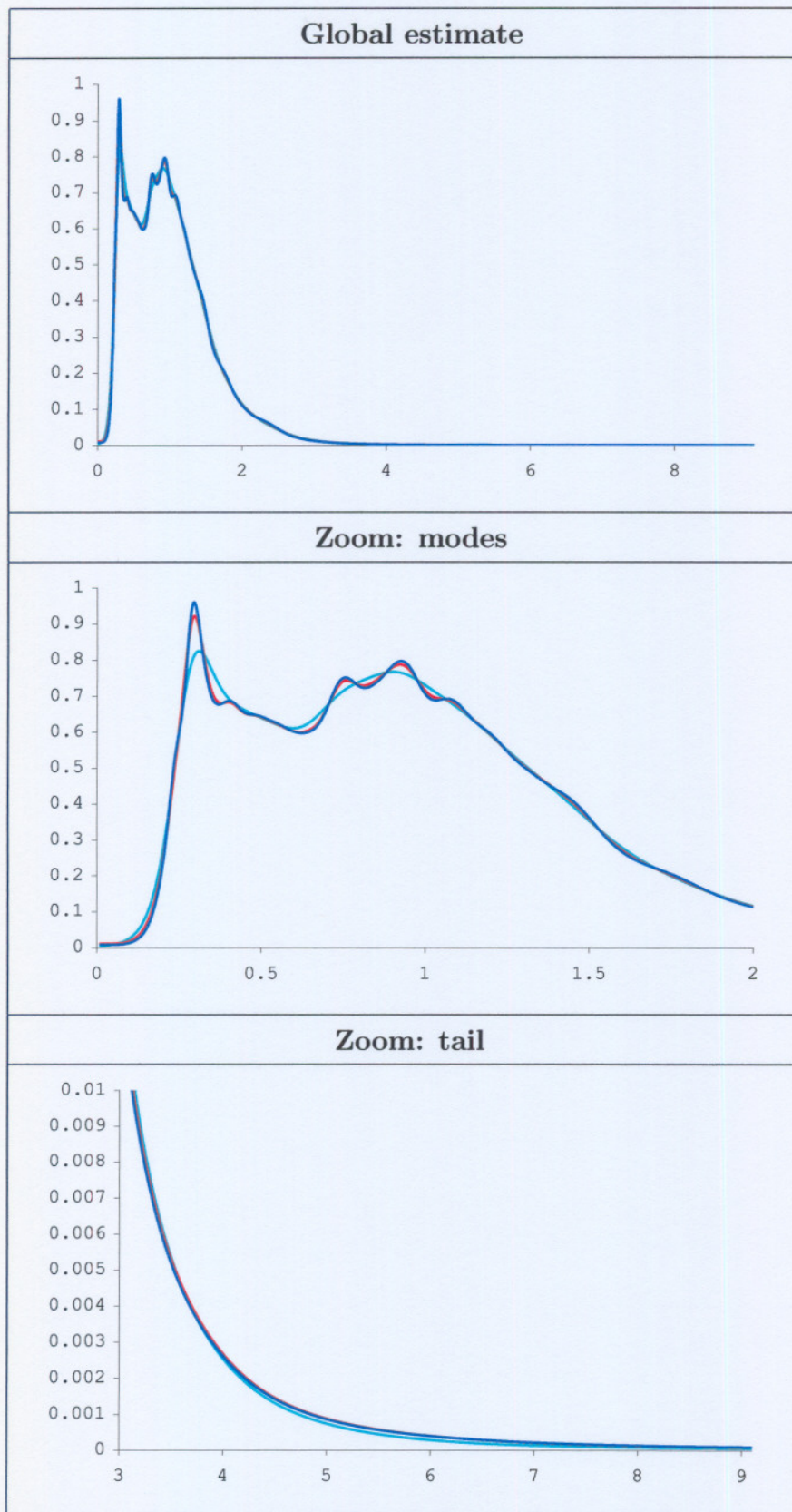
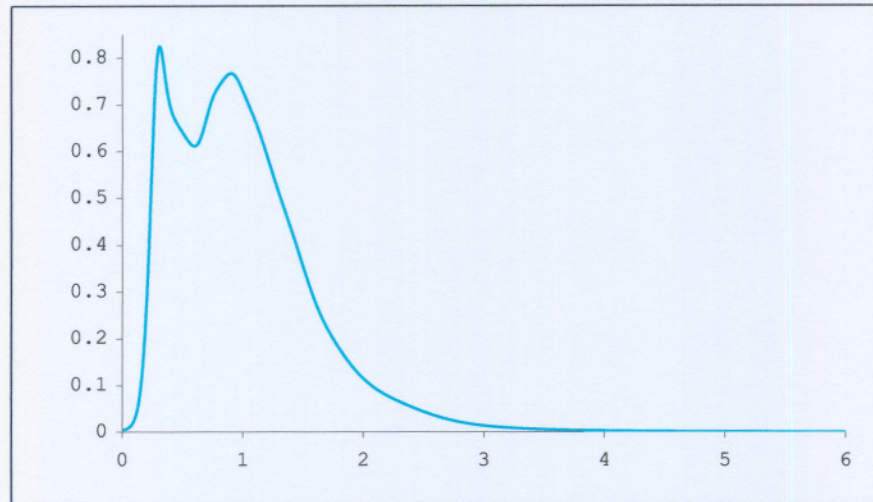


Table 5.72: 0-Step (SEML0)



The 0-step semi-parametric TKDE clearly shows a sharp bimodal structure in the data, which was also found by Wand (1997) and de Beer and Swanepoel (1999).

5.2.2 Example 2: Astrophysical data

This data were obtained from an Astrophysical experiment by considering all pulsar phases above 50 MeV for the Geminga pulsar (see Mayer-Hasselwander (1994)). The data consisting of 5018 phases were extracted from the public domain Phase I of the EGRET experiment on Compton Gamma Ray Observatory. The data were standardized yielding

$$Z_{x,i} = \frac{X_i - 0.5487}{0.3026}, \quad i = 1, \dots, 5018.$$

The scale invariant global roughness measure

$$L(\hat{\lambda}) = \hat{\sigma}_{y_0} R \left(f''_{Y_0}(\cdot; \hat{\lambda}) \right)^{1/5},$$

was minimized by the best transformation - parameter estimation combination. The Johnson (1949) family of transformations with $\gamma = 3$ (JB) was selected with the minimum residual (MR) parameter estimation technique. The estimated parameter is $c = 0.5448$. The constant shift parameter δ (see 4.37) is 0.165. Table 5.73 displays the transformation function and resulting transformation function derivative for iterations 0, 1 and 2. It should be clear that for the first and second iteration steps, the transformations change shape more often, from convex to concave and vice versa, than the 0-step transformation. These changes in shape are clearly visible from inspection of the transformation function

derivatives. The effect of these shape changes is that the resulting 1-step and 2-step semi-parametric TKDE's will capture the density curvature more profoundly.

Table 5.73: Transformation functions and transformation function derivatives

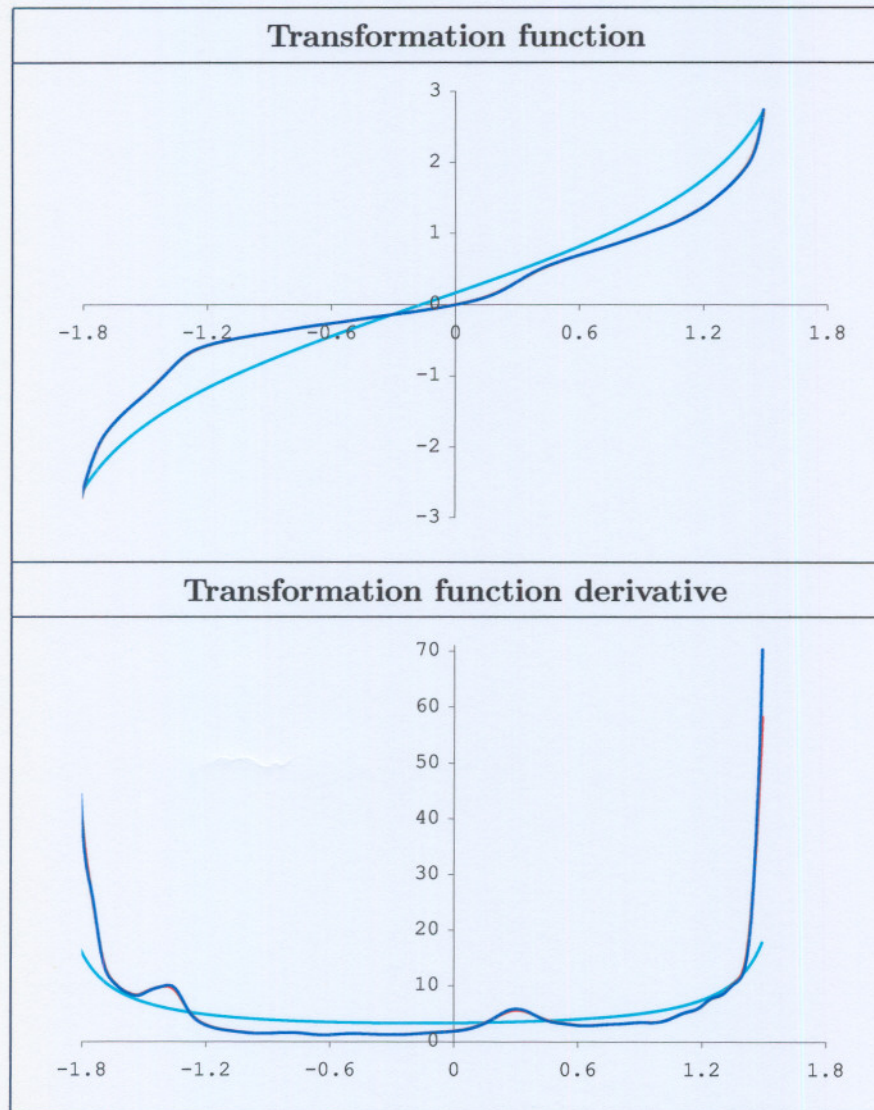


Table 5.74 displays the density estimates of the 0-step ([SEML_0](#)), 1-step ([SEML_1](#)) and 2-step ([SEML_2](#)) semi-parametric TKDE's. The 0-step semi-parametric TKDE appears to over smooth the density. It should be noted that the 1-step and 2-step estimators appear to be similar, which can be ascribed to the fact that a large sample was used. Hence, the 1-step semi-parametric TKDE should be preferred, which is displayed in Table 5.75.

Table 5.74: Density estimates

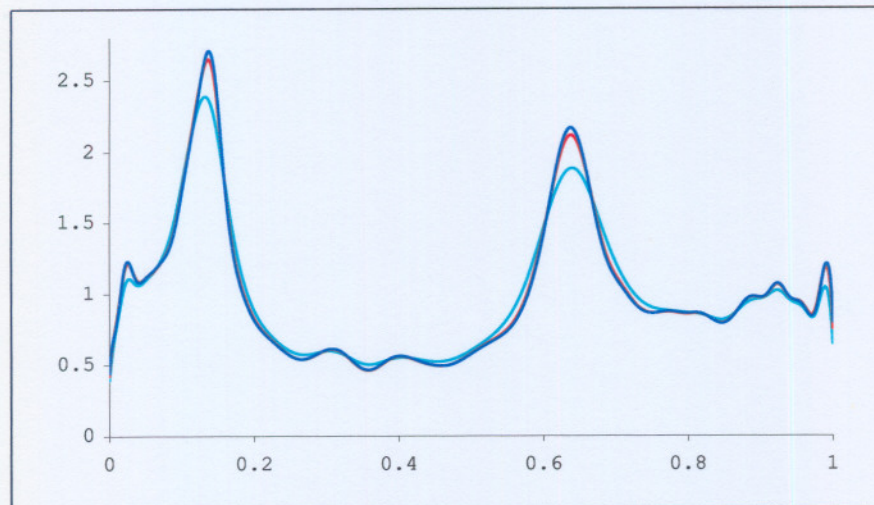
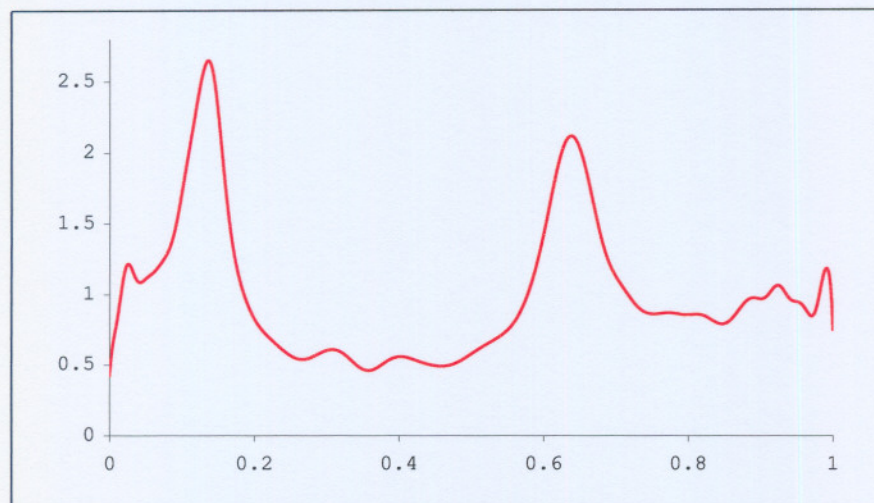


Table 5.75: 1-Step (SEML1)



The 1-step parametric TKDE complies to all the profiles of the Geminga pulsar described by de Jager (1994). De Beer & Swanepoel (1999) derived similar conclusions.

5.2.3 Example 3: Buffalo snowfall data

This data set is the well-known Buffalo snowfall data which consist of 63 observations (see Scott (1992), p.279). Much controversy exists in the literature regarding the distribution of this data. According to Scott (1992, p.109) some researchers argued that this data appear to be trimodal, while others suggested a unimodal distribution. The data were standardized yielding

$$Z_{x,i} = \frac{X_i - 79.6}{23.72}, \quad i = 1, \dots, 63.$$

Since $n \leq 2000$, the Shapiro-Wilk p-value for normality was maximized by the best transformation - parameter estimation combination. The Johnson (1949) family of transformations with $\gamma = 3$ (JB) was selected with the profile maximum likelihood (ML) parameter estimation technique. The estimated parameter is $c = 0.3545$. The Shapiro-Wilk p-value for normality is 0.832 and the constant shift parameter δ (see 4.37) is 0.165. Table 5.76 displays the transformation function and resulting transformation function derivative for iterations 0, 1 and 2.

Table 5.76: Transformation functions and transformation function derivatives

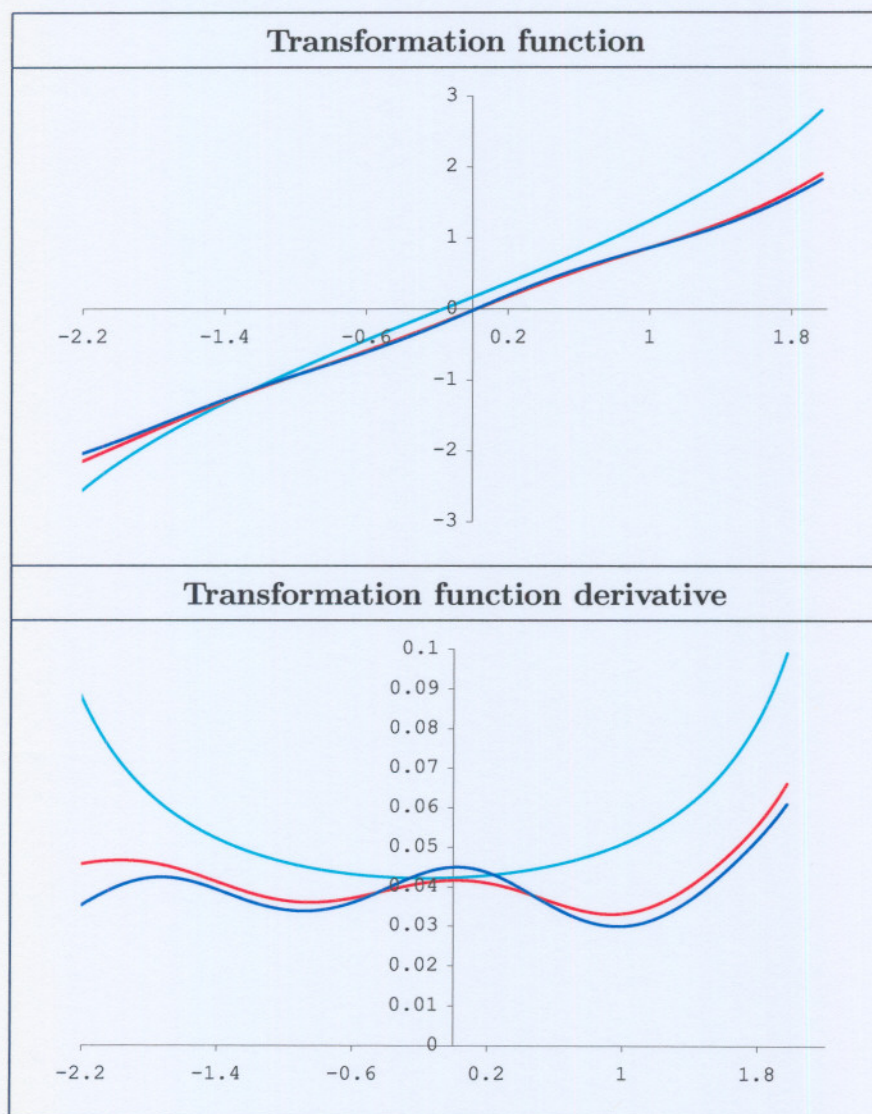


Table 5.77 displays the density estimates of the 0-step (SEML0), 1-step (SEML1) and 2-step (SEML2) semi-parametric TKDE's. The 0-step semi-parametric TKDE appears

to be unimodal, while the 1-step and 2-step TKDE's appear to be trimodal. For this example we prefer the 2-step semi-parametric TKDE, which is displayed in Table 5.78.

Table 5.77: Density estimates

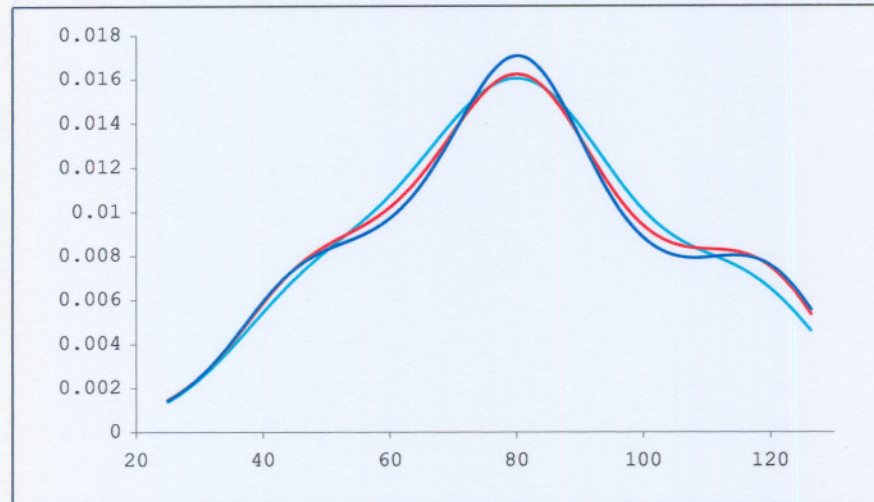
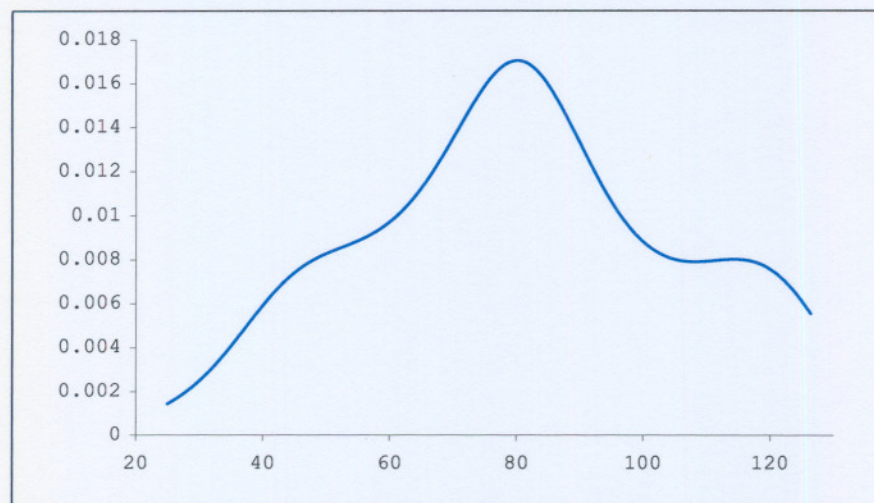


Table 5.78: 2-Step (SEMI_2)



The 2-step semi-parametric TKDE clearly suggests a trimodal density. This conclusion is in contrast to the finding of other researchers who suggested a unimodal density. Moreover, it is also in contrast to the conclusion reached by de Beer and Swanepoel (1999) that the underlying density of the data is bimodal. *Hence, we conclude that the newly proposed semi-parametric TKDE has the ability to capture density features more efficiently.* It should also be noted that the sample size used to calculate the density estimate given in Table 5.78 is relatively small and, if increased, we believe that the trimodal nature of the population density would become even more visible.

Bibliography

- Abramson, I. S. (1982). On bandwidth variation in kernel estimates - a square root law, *The Annals of Statistics*, **10**: 1217–1223.
- Alberts, T. and Karunamuni, R. J. (2003). A semiparametric method of boundary correction for kernel density estimation, *Statistics & Probability Letters*, **61**: 287–298.
- Aldershof, B. (1991). *Estimation of integrated squared density derivatives*, PhD thesis, University of North Carolina, Chapel Hill.
- Altman, N. and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation, *Journal of Statistical Planning and Inference*, **46**: 195–214.
- Atkinson, A. C. (1985). *Plots, Transformations and Regression*, Clarendon Press, Oxford University Press.
- Atkinson, A. C., Pericchi, L. R. and Smith, R. L. (1991). Grouped likelihood for the shifted power transformation, *Journal of the Royal Statistical Society, Series B*, **53**: 473–482.
- Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method, *Biometrika*, **68**: 326–328.
- Beirlant, J., Teugels, J. L. and Vynckier, P. (1996). *Practical Analysis of Extreme Values*, University Press, Leuven.
- Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited, *Journal of the American Statistical Association*, **76**: 296–311.
- Bolancé, C., Guillen, M. and Nielsen, J. P. (2003). Kernel density estimation of actuarial loss functions, *Insurance: Mathematics and Economics*, **32**: 19–36.

- Bowman, A., Hall, P. and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions, *Biometrika*, **85**: 799–808.
- Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71**: 353–360.
- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, **26**: 211–252.
- Burdige, J. B., Magee, L. and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable, *Journal of the American Statistical Association*, **83**: 123–127.
- Burnham, K. P., Anderson, D. R. and Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations, *Wildlife Monograph*, **72**.
- Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Computational Statistics & Data Analysis*, **17**: 153–176.
- Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin, *Journal of the Royal Statistical Society, Series B*, **45**: 394–403.
- Chiu, S.-T. (1991a). Bandwidth selection for kernel density estimation, *The Annals of Statistics*, **19**: 1883–1905.
- Chiu, S.-T. (1991b). The effect of discretization error on bandwidth selection for kernel density estimation, *Biometrika*, **78**: 436–441.
- Chiu, S.-T. (1992). An automatic bandwidth selector for kernel density estimation, *Biometrika*, **79**: 771–782.
- Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation, *Statistica Sinica*, **6**: 129–145.
- Chu, I.-S. (1995). Bootstrap smoothing parameter selection for distribution function estimation, *Mathematical Japonica*, **41**: 189–197.

- Cline, D. B. H. and Hart, J. D. (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives, *Statistics*, **22**: 69–84.
- Cowling, A. and Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation, *Journal of the Royal Statistical Society, Series B*, **58**: 551–563.
- D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit techniques*, Marcel Dekker, Inc, New York.
- Davison, A. C. and Hall, P. (1997). On kernel density estimation without bumps in the tail, Unpublished Manuscript.
- de Beer, C. F. and Swanepoel, J. W. H. (1999). Simple and effective number-of-bins circumference selectors for a histogram, *Statistics and Computing*, **9**: 27–35.
- de Jager, O. C. (1994). On periodicity tests and flux limit calculations for gamma-ray pulsars, *The Astrophysical Journal* **436**: 239–248.
- Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*, Wiley, New York.
- Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling extremal events for Insurance and Finance*, Springer, Berlin.
- Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions, in T. Gasser and M. Rosenblatt (eds), *Smoothing techniques for curve estimation*, Springer-Verlag, Heidelberg, pp. 23–68.
- Gaudard, M. and Karson, M. (2000). On estimating the Box-Cox transformation to normality, *Communications in Statistics - Simulation and Computation*, **29**: 559–582.
- Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives, *Statistics & Probability Letters*, **6**: 109–115.
- Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation, *Probability Theory and Related Fields*, **92**: 1–20.

- Hössjer, O. and Ruppert, D. (1995). Asymptotics for the transformation kernel density estimator, *The Annals of Statistics*, **23**: 1198–1222.
- Hill, B. M. (1975). A simple general approach to inference about the tail of the distribution, *The Annals of Statistics*, **3**: 1163–1174.
- Hjort, N. L. (1996). Bayesian approaches to non- and semiparametric density estimation, in J. M. Berger, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics*, Vol. 5, Oxford Press University, New York, pp. 223–253.
- Hodges, J. L. and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors to the t-test, *The Annals of Mathematical Statistics*, **13**: 324–335.
- Hurvich, C. M. and Simonoff, J. S. (1998). Smoothing parameter selection in nonparametric regression using an improved AIC criterion, *Journal of the Royal Statistical Society, Series B*,.
- Janssen, P., Marron, J. S., Veraverbeke, N. and Sarle, W. (1995). Scale measures for bandwidth selection, *Journal of Nonparametric Statistics*, **5**: 359–380.
- John, J. A. and Draper, N. R. (1980). An alternative family of transformations, *Journal of the Royal Statistical Society, Series C*, **29**: 190–197.
- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation, *Biometrika*, **36**: 149–176.
- Jones, M. C. (1990). The performance of kernel density functions in kernel distribution function estimation, *Statistics & Probability Letters*, **9**: 129–132.
- Jones, M. C. (1993). Simple boundary correction for kernel density estimation, *Statistics and Computing*, **3**: 135–146.
- Jones, M. C. and Foster, P. J. (1993). Generalized jackknifing and higher order kernels, *Journal of Nonparametric Statistics*, **3**: 81–94.
- Jones, M. C. and Foster, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation, *Statistica Sinica*, **6**: 1005–1013.
- Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics*, **11**: 337–379.

- Koekemoer, G. (1999). *A comparative study of non-parametric density estimators*, Master's thesis, P.U. for C.H.E, Potchefstroom.
- Loader, C. R. (1995). Old faithful erupts: bandwidth selection reviewed, Unpublished Manuscript.
- Loots, H. (1995). *Nonparametric estimation of the antimode and the minimum of density function*, PhD thesis, Potchefstroom University for Christian Higher Education.
- Manley, B. F. (1976). Exponential data transformations, *The Statistician*, **25**: 37–42.
- Markovitch, N. M. and Krieger, U. R. (2000). Nonparametric estimation of long-tailed density functions and its application to the analysis of World Wide Web traffic, *Performance Evaluation*, **42**: 205–222.
- Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation, *Journal of the Royal Statistical Society, Series B*, **56**: 653–671.
- Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error, *The Annals of Statistics*, **20**: 712–736.
- Mayer-Hasselwander (1994). High energy gamma radiation from gemina observed by egret, *The Astrophysical Journal*, **421**: 276–283.
- Müller, H.-G. (1985). Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators, *Statistical Decisions*, **2**: 193–206.
- Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints, *Biometrika*, **78**: 521–530.
- Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors, *Journal of the American Statistical Association*, **85**: 66–72.
- Park, B. U. and Marron, J. S. (1992). On the use of pilot estimators in bandwidth selection, *Journal of Nonparametric Statistics*, **1**: 231–240.
- Park, B. U. and Turlach, B. A. (1992). Practical performance of several data driven bandwidth selectors, *Computational Statistics*, **7**: 251–270.

- Park, B. U., Chung, S. S. and Seog, K. H. (1992). An empirical investigation of the shifted power transformation method in density estimation, *Computational Statistics & Data Analysis*, **14**: 183–191.
- Parzen, E. (1962). On the estimation of a probability density function and the mode, *The Annals of Mathematical Statistics*, **33**: 1065–1076.
- Polansky, A. M. (1997). Bandwidth selection for kernel distribution functions, Unpublished Manuscript.
- Reiss, R. D. (1981). Nonparametric estimation of smooth distribution functions, *Scandinavian Journal of Statistics*, **8**: 116–119.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, **27**: 832–837.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, **9**: 65–78.
- Ruppert, D. and Cline, D. B. H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations, *The Annals of Statistics*, **22**: 185–210.
- Ruppert, D. and Wand, M. P. (1992). Correcting for kurtosis in density estimation, *Australian Journal of Statistics*, **34**: 19–29.
- Sakia, R. M. (1992). The Box-Cox transformation technique: a review, *The Statistician*, **41**: 169–178.
- Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions, *Journal of Statistical Planning and Inference*, **35**: 65–75.
- Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities, *Communications in Statistics - Theory and Methods*, **14**: 1123–1136.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York.
- Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association*, **82**: 1131–1146.

- Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality, *Biometrika*, **52**: 591–611.
- Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, **53**: 683–690.
- Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimators, *Journal of the American Statistical Association*, **85**: 410–416.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Staniswalis, J. G. (1989). Local bandwidth selection for kernel estimates, *Journal of the American Statistical Association*, **84**: 284–288.
- Swanepoel, J. W. H. (1987). Optimal kernels when estimating nonsmooth densities, *Communications in Statistics - Theory and Methods*, **16**: 1835–1848.
- Swanepoel, J. W. H. (1988). Mean integrated squared error properties and optimal kernels when estimating a distribution function, *Communications in Statistics - Theory and Methods*, **17**: 3785–3799.
- Tan, W. D., Gan, F. F. and Chang, T. C. (2004). Using normal quantile plot to select an appropriate transformation to achieve normality, *Computational Statistics & Data Analysis*, **45**: 609–619.
- Terrell, G. R. (1990). The maximal smoothing principle in density estimation, *Journal of the American Statistical Association*, **85**: 470–477.
- Titterton, D. M. (1985). Comment on ‘Estimating parameters in continuous univariate distributions’, *Journal of the Royal Statistical Society, Series B*, **47**: 115–116.
- Tukey (1957). The comparative anatomy of transformations, *The Annals of Mathematical Statistics*, **28**: 602–632.
- van Graan, F. C. (1982). *Nie-parametriese beraming van verdelingsfunksies*, Master’s thesis, P.U. for C.H.E, Potchefstroom.

- Wand, M. P. (1997). Data-based choice of histogram bin width, *The American Statistician*, **51**: 59–64.
- Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, London.
- Wand, M. P. and Schucany, W. R. (1990). Gaussian-based kernels, *The Canadian Journal of Statistics*, **18**: 197–204.
- Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformations in density estimation, *Journal of the American Statistical Association*, **86**: 343–361.
- Yang, L. and Marron, J. S. (1999). Iterated transformation kernel density estimation, *Journal of the American Statistical Association*, **94**: 580–589.
- Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry, *Biometrika*, **87**: 954–959.
- Zhang, S. and Karunamuni, R. J. (1998). On kernel density estimation near endpoints, *Journal of Statistical Planning and Inference*, **70**: 301–316.
- Zhang, S., Karunamuni, R. J. and Jones, M. C. (1999). An improved estimator of the density function at the boundary, *Journal of the American Statistical Association*, **94**: 1231–1241.