Selecting tuning parameters in minimum distance estimators

Thesis

Selecting tuning parameters in minimum distance estimators

Jane Warwick, BSc

Department of Statistics
Faculty of Mathematics and Computing
The Open University

Submitted for the degree of Doctor of Philosophy
August 2001

EX12

THE OPEN UNIVERSITY

RESEARCH SCHOOL

Library Authorisation Form

20 FEB 2002

Please return this form to the Research School with the two bound copies of your thesis to be deposited with the University Library. All candidates should complete parts one and two of the form. Part three only applies to PhD candidates.

Part One: Candidates Details

Name: JANE WARWICK PI: R2013251

Degree: PhD

Thesis title: SELECTING TUNING PARAMETERS IN MINIMUM DISTANCE ESTIMATORS

Part Two: Open University Library Authorisation

I confirm that I am willing for my thesis to be made available to readers by the Open University Library, and that it may be photocopied, subject to the discretion of the Librarian.

Signed: J. Warwick Date: 16:02:02

Part Three: British Library Authorisation [PhD candidates only]

If you want a copy of your PhD thesis to be available on loan to the British Library Thesis Service as and when it is requested, you must sign a British Library Doctoral Thesis Agreement Form. Please return it to the Research School with this form. The British Library will publicise the details of your thesis and may request a copy on loan from the University Library. Information on the presentation of the thesis is given in the Agreement Form.

Please note the British Library have requested that theses should be printed on one side only to enable them to produce a clear microfilm. The Open University Library sends the fully bound copy of theses to the British Library.

The University has agreed that your participation in the British Library Thesis Service should be voluntary. Please tick either (a) or (b) to indicate your intentions.

☑ I am willing for the Open University to loan the British Library a copy of my thesis.
A signed Agreement Form is attached

☐ I do not wish the Open University to loan the British Library a copy of my thesis.

Signed: J. Warwick Date: 16:02:02

Abstract

Many minimum distance estimators have the potential to provide parameter estimates which are both robust and efficient and yet, despite these highly desirable theoretical properties, they are rarely used in practice. This is because the performance of these estimators is rarely guaranteed *per se* but obtained by placing a suitable value on some tuning parameter. Hence there is a risk involved in implementing these methods because if the value chosen for the tuning parameter is inappropriate for the data to which the method is applied, the resulting estimators may not have the desired theoretical properties and could even perform less well than one of the simpler, more widely used alternatives. There are currently no data-based methods available for deciding what value one should place on these tuning parameters hence the primary aim of this research is to develop an objective way of selecting values for the tuning parameters in minimum distance estimators so that the full potential of these estimators might be realised.

This new method was initially developed to optimise the performance of the density power divergence estimator, which was proposed by Basu, Harris, Hjort and Jones [3]. The results were very promising so the method was then applied to two other minimum distance estimators and the results compared.

Contents

Acknowledgements

For the most part, working towards my PhD has felt more like an obstacle race than education. Family responsibilities have often conflicted with my research and maintaining the balance between the two has been the greatest challenge. None of this work would have been possible without suitable care for my children so my first task must be to thank Chrissie Kingman and Sarah Baddock, Acorn Pre-school Nursery, Acorns Day Nursery, Aylesbury Vale Play Association and The Buzz Clubs. I also wish to thank John and Dorothy Berryman, Richard Warwick, Elizabeth and Clive Whitemore, Sandra Berryman and Philippa Evett for helping with the children on so many occassions. Most of the credit for childcare however must go to the best nanny in Milton Keynes, Miss Hazel Baker, without whose reliability, loyalty and sense-of-humour I would surely have died of stress many months ago. My supervisor, Professor Chris Jones, also deserves praise for providing excellent academic guidance and Professor John Mason for his much needed personal support and good advice. Finally, I must thank my husband Simon and our daughters Katie and Anna for enduring the last four years with such good grace.

Chapter 1

Introduction

The purpose of many statistical analyses is to learn more about a particular population. This generally involves taking a sample from the population of interest and extrapolating findings based on this sample to the whole population. Thus the classical parametric estimation problem arises from the need to develop statistical models for populations from sampled data. The distribution of the population is generally referred to as the true distribution and any distribution derived from sample data as the model even though, in reality, both are models because the true distribution is simply the statistical representation of some physical or biological phenomenon. In order for a model developed from sampled data to be a useful tool for extrapolation the family of distributions for the model must be chosen appropriately and the values placed on the parameters within that model must be close

to those of this true distribution. Since the parameter estimates and, to a lesser extent, the choice of model family are based on the sample data it is imperative that this sample is representative of the population from which it was taken.

Maximum likelihood is often used to estimate the unknown parameters in models because it provides asymptotically unbiased estimators which have the lowest possible asymptotic variance (equal to the Cramér-Rao lower bound) when the model is true (i.e. whilst there is no contamination in the data and the model family is chosen correctly). This optimality does not necessarily hold under departures from the model, however, and the danger in using maximum likelihood estimators in such situations is easily demonstrated in Figure 1.1 (page 4). Here the true density of the data, denoted g, is the $N(0,1)$ distribution and shown in black. This density is often referred to as the target distribution because in most practical situations it is the parameters of the true density which we wish to estimate. The distribution of sampled data is not usually of direct interest and can be viewed as an estimate of the distribution of g. Several studies have found that up to 10% of the values in a typical data set might be incorrect [1] as a result of recording errors, measurement errors and rounding. Errors can arise in a variety of ways but are often difficult to identify or quantify in practice because the underlying true density is not generally known.

Therefore it is important to investigate how such errors affect the parameter estimates and models obtained. First a random sample of size 100 was taken from this distribution and the location parameter in the $N(\theta, 1)$ family of models was estimated using maximum likelihood. The resulting model (the red curve in Figure 1.1, p.4) is very close to the target distribution which confirms that for clean data such as this maximum likelihood estimation leads to very satisfactory results. By adding a number of contamination points with the value 10 to the $N(0, 1)$ samples (to represent errors in the data) revised estimates of the location parameter can be obtained, again using maximum likelihood estimation. These models are also shown in Figure 1.1 and demonstrate that as the percentage of contamination points increases the model obtained using maximum likelihood estimation shifts further away from the true density g. This highlights the fact that the maximum likelihood estimator is very sensitive to errors in the data so that models fitted in this way, and the predictions or opinions based upon them, could be misleading.

A robust estimation method is therefore one which leads to models which are close to the true density even though the distribution of the data might not all be. It is an unfortunate feature of many robust estimation methods that they do not perform as well as maximum likelihood when there is no contamination in the data. In practice, since the true density is un-

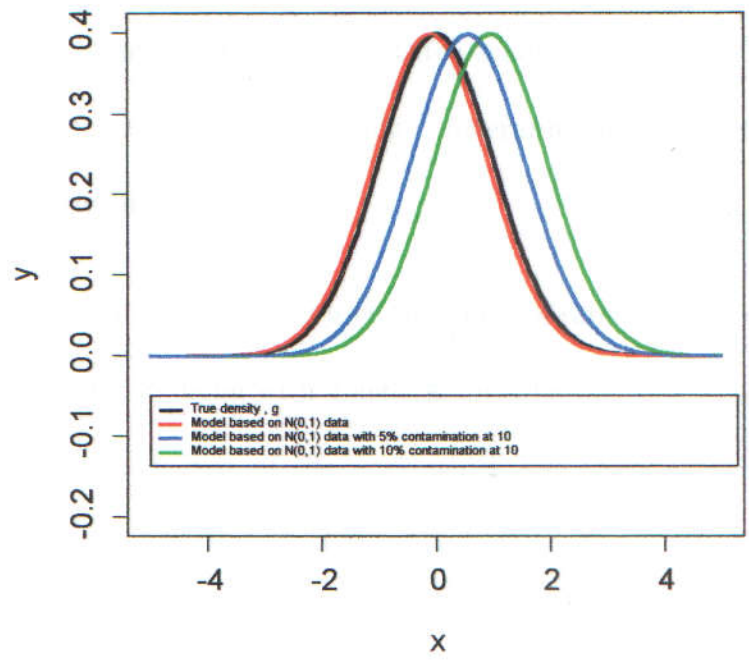

Figure 1.1: Illustration of how contamination or errors in the data can lead to poor parameter estimates.

known, deciding between a robust method and maximum likelihood is a balancing act and the cost of making the wrong choice can be very high. Minimum distance estimators [3],[27],[38] are of interest because they are a class of methods which have the desirable properties of being both robust and efficient. Nonetheless, these methods are not widely used because these theoretical benefits are often difficult to attain in practice. The aim of this research has therefore been to study three examples from this class of robust estimators with a view to developing a general method for optimising their practical performance.

A summary of the most widely used measures of robustness and efficiency is given in Chapter 2, along with a general introduction to minimum distance estimators and a review of the work which has already been carried out to optimise the performance of this class of estimator. Each of the three minimum distance estimators considered here is introduced in Chapter 3 and a new approach to optimising their performance will be applied to each in Chapters 4 to 6. The results are summarised in Chapter 7 so that the effectiveness of this new approach can be fully assessed and conclusions drawn in Chapter 8. To make the thesis easier to read theoretical details, such as the derivation of the asymptotic properties of each estimator, are outlined in Appendices A to D and referenced where appropriate.

Chapter 2

Robustness, efficiency and minimum distance estimators

Robustness theory has developed from the pioneering work by Huber and Hampel in the late 1960's and early 1970's. Huber's many papers on the minimax approach are summarised in his book on robust statistics [16] and Hampel's infinitesimal approach (based on influence functions) in another [11]. This chapter brings together the key ideas from both these schools, alongside more recent developments, needed to explain and justify the new methods suggested later in this thesis.

2.1 Functional representation of estimators

When studying the theoretical properties of estimators it is often advantageous to regard them as functionals. The mathematical justification for this representation is given in Section 2.1a of Hampel *et al* [11] and a more intuitive explanation in Staudte and Sheather [35] on page 12. In simple terms the feature which is to be estimated, location for example, is represented by the functional $T(F)$ where T is a general term for the estimand and F is the distribution to which T is applied. Thus, using F_n to denote the empirical distribution function and G to denote the true distribution function, the population mean $\int x\, g\,(x)\, dx$ can be denoted as $T(G)$ and the sample mean as $T\,(F_n)$. This separation of the estimand from the data greatly simplifies the task of investigating the effect of small changes in the data on parameter estimates and allows theory which is common to many different estimators to be expressed in general terms. This means that it does not need to be re-derived for each specific method.

This notation also gives a neat representation of the variance of an estimator as follows

$$var\left(\sqrt{n}\,(T(F_n) - T(G))\right) \equiv V\,(T, G)$$

which leads to

$$var\left(T(F_n)\right) = \frac{1}{n}V\left(T, G\right).$$

2.2 Modelling contaminated data

The theoretical behaviour of an estimator under departures from the model can be explored by using a mixture model to represent real data. This mixture density (also known as the gross-error model) was introduced by Huber [15] and is

$$g_\varepsilon\left(x\right) = (1-\varepsilon)g\left(x\right) + \varepsilon h\left(x\right)$$

where $0 \leq \varepsilon \leq 1$ so that $100(1-\varepsilon)\%$ of the observations are from the true distribution g and the remaining $100\varepsilon\%$ are from some other density distribution h. The choice of density for h determines the type of contamination which will be obtained. Setting $h\left(x\right) = \delta_\xi\left(x\right)$ where δ is the Dirac delta function enables contamination points with a particular value ξ to be generated as can symmetric heavy tailed data by setting $h\left(x\right) = t_k\left(x\right)$, the Student's t density, with degrees of freedom $k = 2, 3$ or 4. The Student's t distribution with $k = 5$ is not appropriate for this purpose it is very similar to the standard normal and the random samples obtained would not have the desired outliers in the tails. The converse is true when $k = 2$ in which case the resulting random samples would be highly skewed and therefore

also unsuitable. The percentage of contamination points can be varied by changing the value of ε and so gives a very flexible model which can be adjusted to mimic many of the different types of data observed in practice.

2.3 Investigating the effect of contamination on estimators

The functional representation of estimators and use of mixture models for contamination are extremely valuable tools for investigating the robustness of estimators. As described in Section 2.1 (p.7), an estimator can be written in general terms as $T(F)$, where F is a probability density or distribution function. The true value of a location parameter, for example, is then written as $T(g) = \theta$ where g is the true density function. Setting $F = g_\varepsilon$ gives $T(g_\varepsilon) = \theta_\varepsilon$ which is the estimator when the method is applied to contaminated data. Thus the effect of contamination on an estimator can be denoted simply as $T(g_\varepsilon) - T(g)$ which is the bias in using θ_ε to estimate the true location parameter θ_* (the location parameter associated with g). This functional expression for bias has many uses in robustness theory as will be seen in later sections.

2.4 Measuring efficiency

Maximum likelihood is often used in parametric estimation problems because it provides asymptotically unbiased estimators which have the lowest possible asymptotic variance (equal to the Cramer-Rao lower bound) when there is no contamination in the data and the model family is chosen correctly. Thus maximum likelihood is the optimal choice of estimation method when the model is true and there is no data contamination and it therefore provides a benchmark for judging the efficacy of other estimation methods. Using F to represent any probability distribution function the efficacy of an estimator $T(F_n)$ is known as the relative efficiency and is defined as follows

$$\begin{aligned} RE_{T,ML} &= \frac{V(ML,G)}{V(T,G)} \\ &\propto (V(T,G))^{-1} \end{aligned}$$

where $V(ML,G)$ is the variance of the maximum likelihood estimator and $V(T,G)$ is proportional to the variance of $T(F_n)$.

Since it is rarely possible to express the variance of an estimator explicitly the asymptotic variance, which is a large sample approximation to the true variance, may be used instead. In this case the efficacy of an estimator $T(F_n)$ is measured by the asymptotic relative efficiency which is defined as

follows

$$\begin{aligned} ARE_{T,ML} &= \frac{v\left(ML,G\right)}{v\left(T,G\right)} \\ &\propto \left(v\left(T,G\right)\right)^{-1} \end{aligned}$$

where $v(ML,G)$ is the asymptotic variance of the maximum likelihood estimator and $v\left(T,G\right)$ is proportional to the asymptotic variance of $T\left(F_n\right)$.

Maximum likelihood estimators are optimal and therefore fully (or 100%) efficient using either measure of efficiency. The asymptotic relative efficiency of some other estimators may also be 100% despite their performance being sub-optimal in the finite case.

For many parametric estimators the ARE does not depend on θ and is therefore a measure of the relative sample size required by ML in order to estimate θ with the same accuracy as T. Thus if T is 85% efficient one would need to increase the sample size by 18% for T to perform as well as maximum likelihood.

2.5 Measuring Robustness

2.5.1 Breakdown point

As demonstrated in Section 2.2 (p.8) the density g_ε is used to investigate the effect of contaminated data on estimators. As the proportion of con-

taminants in the data increases one would expect the parameter estimates to become less reliable so one way of assessing robustness is to consider how large ε can get before the validity of the parameter estimates is undermined. This value of ε is known as the breakdown point and can be interpreted as the degree of contamination which the estimator can tolerate. A highly robust estimation method might therefore have a breakdown point of around 40% whereas the maximum likelihood estimator, which cannot tolerate contaminated data, has a breakdown point of 0%. Details of the breakdown points of the three minimum distance estimators considered here are given in Sections 3.1, 3.2 and 3.3 of Chapter 3 on pages 34, 47 and 64 respectively.

2.5.2 Influence function and related measures

Another aspect of robustness is to consider how the magnitude of the contamination point might affect the estimator. Once again a mixture model is used for contamination, $g_\varepsilon(x) = (1-\varepsilon)\, g(x) + \varepsilon \partial_\xi(x)$ and the effect of contamination on the estimator T is given by $T(g_\varepsilon) - T(g)$. The relative influence of each of $\varepsilon\%$ contamination points is therefore $\frac{T(g_\varepsilon)-T(g)}{\varepsilon}$ and the infinitesimal behaviour of an estimator is found by taking the limit of this expression as $\varepsilon \to 0$. This function, known as the influence function or influence curve, describes the relative influence of a contamination point at ξ on the estimator and thus represents an estimator's sensitivity

to small changes in the data. The influence function is therefore denoted $IF(\xi) = \lim_{\varepsilon \to 0} \left(\frac{T(g_\varepsilon) - T(g)}{\varepsilon} \right)$. This approach, first suggested by Hampel [10], is an extremely useful tool because it gives a simple graphical representation of robustness which can be easily interpreted. The influence functions for the mean and Cramér-von Mises estimator [13] (the $\psi_\theta = f_\theta$ case of the weighted Cramér-von Mises estimator which is described in Subsection 2.9.3, p.23) are shown in Figure 2.1 (p.14) to illustrate some of the features of interest. The mean is an example of an estimator which has an unbounded influence function. This means that as the distance between the true mean (zero in this case) and the contamination point increases so does the effect on the estimator, without limit. In contrast, the influence function for the Cramér-von Mises estimator is bounded above and below at approximately ± 2 which means that large contamination points have much less influence on this estimator than on the mean. Thus if the influence function for an estimator is bounded it is generally thought to be robust and vice versa.

There are other aspects of influence functions which also need to be considered when assessing the robustness of an estimator. The position of the bounding affects robustness because it represents the maximum influence that a contamination point could ever exert on an estimator. This value is

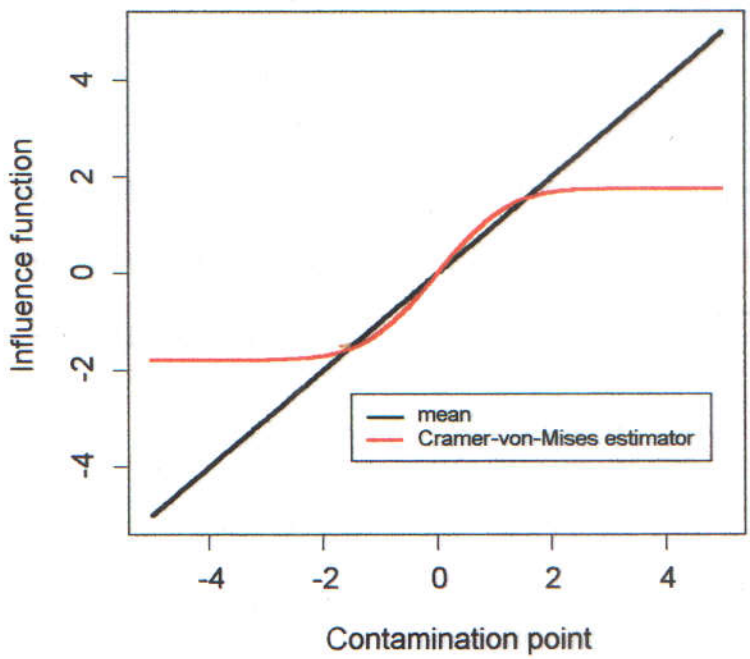

Figure 2.1: Influence function for the mean and Cramér-von-Mises estimators.

known as the gross error sensitivity and defined as

$$\gamma^* = \sup_{\xi} |IF(\xi)|$$

where the supremum is taken over all ξ where $IF(\xi)$ exists. If γ^* is finite then T is said to be *B-robust* or bias robust [28]

The general shape of the influence function may also be of interest because this aspect may also affect robustness. Several estimators have influence functions which redescend to zero which means that large contamination points have very little influence and as a consequence these estimators are highly robust. When g is symmetric the rejection point for a redescending estimator can therefore be defined as follows

$$\rho^* = \inf \{r > 0; IF(\xi) = 0 \text{ when } |\xi| > r\}$$

If no such r exists then $\rho^* = \infty$.

The effect of changes in ε on the asymptotic variance of an estimator can be studied via the change-of-variance function [11], [29] which is defined as

$$CVF(\xi) = \lim_{\varepsilon \to 0} \left(\frac{w(T(g_\varepsilon)) - w(T(g))}{\varepsilon} \right)$$

where $w(T(F)) = \ln[v(T, F)]$, the logarithm of the asymptotic variance of $T(F)$.

A positive value for the CVF means increased variability of the estimator whilst negative values imply increased accuracy. Therefore, for maximum

variance robustness, the CVF function should be bounded above but need not be bounded from below. The change-of-variance sensitivity is the supremum of the standardised CVF as follows

$$\kappa^* = \sup_{\xi} \{CVF(\xi)/v(T,G)\}$$

Note that if a delta function with a positive factor occurs in the CVF then κ^* is defined as $+\infty$. An estimator with finite κ^* is known as variance or V-robust.

2.6 Relationship between robustness and efficiency

The relationship between efficiency, robustness and influence functions, detailed in Lindsay [19], can be derived from the Taylor series approximation to the estimator under contamination which is

$$T(g_\varepsilon) = T(g) + \varepsilon \left.\frac{dT(g_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} + \frac{\varepsilon^2}{2} \left.\frac{d^2T(g_\varepsilon)}{d\varepsilon^2}\right|_{\varepsilon=0} + \ldots$$

The influence function can therefore be viewed as a first order approximation to the asymptotic bias because

$$\begin{aligned} T(g_\varepsilon) - T(g) &\simeq \varepsilon \left.\frac{dT(g_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0} \\ &= \varepsilon IF(\xi). \end{aligned}$$

The influence function is also related to efficiency via von Mises expansion [20] so for any probability distribution function F, subject to certain regularity conditions being fulfilled, the asymptotic variance of an estimator can be obtained as

$$v(T,F) = \int IF^2(x)\, dF(x). \tag{2.1}$$

where $IF^2(x)$ is the squared influence function of the estimator $T(F)$.

A simple example of how the influence function can be used to obtain an estimate of the standard error of $T(F_n)$ is given in Staudte and Sheather [35] (Chapter 3, p.80). For the $N(\theta, \sigma^2)$ model the maximum likelihood estimator of location is the sample arithmetic mean, denoted $T(F_n) = \overline{X}_n$, which has the influence function $IF(\xi) = \xi - \overline{X}_n$. Since $var(T(F_n)) = \frac{1}{n}v(T,G)$, an estimate of $var(T(F_n))$ can be obtained as follows

$$\begin{aligned} var(T(F_n)) &= \frac{1}{n}\int IF^2(x)\, dG(x) \\ &\simeq \frac{1}{n^2}\sum_{i=1}^{n} IF^2(X_i) \\ &\simeq \frac{1}{n^2}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \\ &\simeq \frac{\widehat{\sigma}^2}{n} \end{aligned}$$

where n is the sample size and F_n is the empirical distribution function. Thus although robustness and efficiency might appear to be entirely different properties they are not unrelated. The precise nature of their interac-

tion is difficult to identify, particularly because there is no single measure of robustness, but it exists nonetheless and it seems foolish to assess the performance of an estimator in terms of one of these aspects without due consideration of the other.

2.7 Joint measures of robustness and efficiency

Since the performance of an estimator is usually assessed in terms of both its robustness and efficiency, it is surprising that no joint measure of these features has been suggested in the literature. One possibility, however, is to use the mean squared error (MSE) of the estimator $T(F_n)$ since $E(T(F_n) - T(G))^2 = (E(T(F_n)) - T(G))^2 + V(T(F_n))$. The first of these two components is the bias which is a measure of robustness and the second would give efficiency if there were no contamination in the data. Since the mean and variance of few minimum distance estimators can be expressed explicitly it is generally more appropriate to use the asymptotic mean squared error ($AMSE$) instead. Thus robustness is measured by the asymptotic bias and efficiency by the asymptotic variance (when $f_\theta = g$). The $AMSE$ therefore summarises the key properties of an estimator and might be extremely useful in assessing their overall performance. Surprisingly, whilst this mea-

sure is often seen in the simulation sections of papers where it is used to assess the performance of robust estimators on simulated data its' use as a theoretical joint measure of robustness and efficiency is often overlooked.

2.8 Minimum distance estimators

Minimum distance estimators are obtained by minimising the difference between the true density g and the model f_θ according to some distance measure. The distances have the general form

$$D\left(f_\theta, g\right) = \int d\left(f_\theta\left(x\right), g\left(x\right)\right) dx$$

where d is some distance function. The majority of these distance functions are based on probability density or distribution functions but there are other possibilities such as distances between characteristic functions [7] or quantile functions [18]. Some examples of density based measures are the L_2 distance $\int \left(f_\theta\left(x\right) - g\left(x\right)\right)^2 dx$ [30] and the Hellinger distance $\int \left(\sqrt{f_\theta\left(x\right)} - \sqrt{g\left(x\right)}\right)^2 dx$ [6] while the Cramér-von Mises distance $\int \left(F_\theta\left(x\right) - G\left(x\right)\right)^2 dx$ [13] is a well known distribution based measure. The minimum distance estimator, $\widehat{\theta}$, is the value of θ which minimises $D\left(f_\theta, \widehat{g}\right)$ where $\widehat{g} = g_n$ or a kernel density estimate of g and, since few minimum

distance estimators can be expressed explicitly, it is generally located using numerical integration and optimisation procedures.

The minimum distance approach to parametric estimation was developed by Wolfowitz [38] in a series of papers during the early 1950's but very little research followed on from this until the late 1970's. Given the amount of computation needed to solve the estimating equations it seems quite likely that the lack of interest during this period was due to these practical problems rather than any doubts concerning the theoretical benefits of these estimators. Computational difficulties may also account for the fact that many of the early minimum distance estimators were based on the distribution function because this meant that the empirical distribution function could be used to estimate the true distribution of the data and sums of order statistics could replace the integrals. A review of this early literature was carried out by Parr [25] in 1981 and shows how interest arose in several minimum distance estimators, in particular in variants of the Cramér-von Mises, Kolmogorov-Smirnoff and Neyman's chi-squared estimators. The introduction of the Hellinger distance estimator by Beran [6] in 1977 sparked further interest because this paper demonstrated that robustness and full asymptotic efficiency could be achieved in the same estimator. This feature of simultaneous robustness and efficiency is the key attraction of this class of estimators and much recent research has concentrated on finding

new minimum distance estimators with these properties. There are also a significant number of theoretical papers which determine the regularity conditions necessary to ensure the existence, consistency and convergence of the estimators. These methods are rarely used in practice, however, perhaps because the gulf between their theory and application has not yet been bridged. The following section examines some of the work which has been done in this area and highlights why this has not established how best to apply these methods in practice.

2.9 Some methods suggested for optimising the performance of minimum distance estimators

2.9.1 Optimally bounding the influence function

A new class of estimator, called M-estimators, was identified by Huber [15] in 1964. Since many minimum distance estimators, the Cramér-von Mises estimators for example, fall within this class, methods which have been suggested for optimising the performance of M-estimators may also be suitable for minimum distance estimators. M-estimators are generalised maximum likelihood estimators and defined as solutions to

$$\int \psi\left(x, T\left(F\right)\right) dF\left(x\right) = 0$$

The maximum likelihood estimator, for example, can be obtained by choosing $\psi = x - \theta$ when $f = N(\theta, 1)$ and there are many other estimators which share this general form and therefore also share the same general expression for their influence function and asymptotic variance. The influence function and asymptotic variance of an M-estimator are as follows:

$$IF\left(x; \psi, F\right) = \frac{\psi\left(x\right)}{\int \psi'(x)\, dF\left(x\right)}$$

and

$$var\left(x; \psi, F\right) = \frac{\int \psi^2\left(x\right) dF\left(x\right)}{\left(\int \psi'(x)\, dF\left(x\right)\right)^2}.$$

In order to construct M-estimators which are both robust and efficient Hampel *et al* [11] suggested placing an upper bound on their gross error sensitivity to make them robust and then maximising their efficiency subject to that constraint to optimise their performance. Constructing an estimator in this way is known as optimally bounding the gross error sensitivity but, because this approach modifies the estimator to make it into a Huber type function, the term "optimal Huberising" may also be used. A Huber type

function, $h_k(x)$, is defined as follows

$$h_k(x) = \begin{cases} -k, & x < -k \\ h(x), & -k \leq x \leq +k \\ +k, & x > +k \end{cases} .$$

where $h(x)$ is any function of x and k is a constant [16].

Unfortunately there is no objective method, as yet, for deciding what the bound on the gross error sensitivity should be and so this method is of limited practical use.

2.9.2 Optimally bounding the change-of-variance function

A similar approach was used by Rousseeuw [28] who placed an upper bound on the change of variance sensitivity κ^* and then found the T which minimises $v(T, F)$ subject to this constraint. This method suffers from the same practical difficulties as optimally bounding the influence function because, once again, it is not known what the value of the bound should be.

2.9.3 Weighted Cramér-von Mises estimators

Parr and De Wet [26] proposed weighting a Cramér-von Mises estimator by some function $\psi_\theta(x)$ and derived a general formula for the influence function

of these estimators. The weighted Cramér-von Mises distance is therefore

$$\int \left(F_\theta\left(x\right) - G\left(x\right)\right)^2 \psi_\theta\left(x\right) dx$$

with

$$IF\left(\xi\right) = \frac{\int \left(\Delta_\xi\left(x\right) - F_\theta\left(x\right)\right) \psi_\theta\left(x\right) f_\theta\left(x\right) \frac{dF_\theta}{d\theta} dx}{\int \left(\frac{dF_\theta}{d\theta}\right)^2 \psi_\theta\left(x\right) f_\theta\left(x\right) dx}.$$

This influence function is bounded so long as the weight function ψ_θ is bounded but since there are an infinite number of such weight functions it is necessary to impose another constraint in order to find the best choice for ψ_θ. By utilising the relationship between the influence function and the asymptotic variance (2.1) it is possible to deduce which ψ_θ will lead to the greatest efficiency and thereby optimise the performance of this family of estimators. The authors present the optimal choice of weight function for data from several distributions where the location parameter is unknown. For $N(\theta, 1)$ data, for example, efficiency is maximised by choosing $\psi_\theta = f_\theta^{-2}$ and for t data with k degrees of freedom $\psi_\theta = \left(k - (x-\theta)^2\right)\left(k + (x-\theta)^2\right)^{k-1}$. A major weakness in this approach is that it relies on such a crude measure of robustness. As illustrated in Section 2.5.2 many other features of the influence function, such as the gross-error sensitivity, may be used to quantify robustness but these aspects are not taken into account. Thus, this approach would prefer a weight function which leads to 95% efficiency with a gross error sensitivity of 4 over another which gives 94% efficiency with a

gross error sensitivity of 2 despite the fact that, on balance, the latter might be the most pragmatic choice.

This approach was also used by Öztürk and Hettmansperger to derive the optimal weight functions for generalised weighted Cramér-von Mises distances [22]. The weights considered were all functions of F_θ^p and affect robustness by either increasing or decreasing the influence of observations in the tails of the model. Several new distribution based distance measures of this type have been developed [14],[21],[23],[24] in which the robustness and efficiency of the estimators is controlled, to some extent, by this additional parameter p. The criterion function proposed in [24], for example, is $C_F(\theta; p) = \int [G^p(x) - F_\theta^p(x)]^2 dx$ and the optimal value of p for a particular type of data was found via simulation. Unfortunately, since the distribution of the data is rarely known, deciding what value to place on p can be problematic and as a consequence these methods are difficult to apply in practice.

2.9.4 Residual Adjustment Function estimators

In 1994 Lindsay [19] proposed an alternative to the influence function for describing the robustness of the Hellinger distance estimator. This led to the identification of a new family of estimators whose estimating equations

are of the form

$$\sum A\left(\partial\left(x\right)\right)\frac{df_{\theta}\left(x\right)}{d\theta}=0$$

where $\partial\left(x\right)=\left[g\left(x\right)-f_{\theta}\left(x\right)\right]/f_{\theta}\left(x\right)$ and $A\left(\partial\right)$ is an increasing twice differentiable function on $[-1,\infty)$ with $A\left(0\right)=0$ and $A'\left(0\right)=1$.

These estimators are all first-order efficient and have the same influence function as the maximum likelihood estimator. They have different robustness and second-order efficiency properties however, as determined by $A\left(\partial\right)$ which is known as the residual adjustment function (RAF). By setting

$$A\left(\partial\right)=\frac{\left(1+\partial\right)^{\lambda+1}-1}{\lambda+1}$$

one can obtain a variety of distance measures including maximum likelihood ($\lambda=0$), Hellinger distance ($\lambda=-\frac{1}{2}$) and Neyman's chi-squared ($\lambda=-2$). The RAF's for these three estimators, shown in Figure 2.2, illustrate how the shape of these functions determines the robustness and efficiency of the resulting estimators. Large positive values of ∂ represent outliers in the data so the behaviour of the RAF's as ∂ increases indicates the robustness of these estimators to large outliers and furthermore the shape of the RAF when ∂ is close or equal to 0 explains whether these estimators will be efficient. Clearly Neyman's chi-squared is the most robust of these methods because $A\left(\partial\right)$ is least affected by large ∂ but it is also the least like maximum

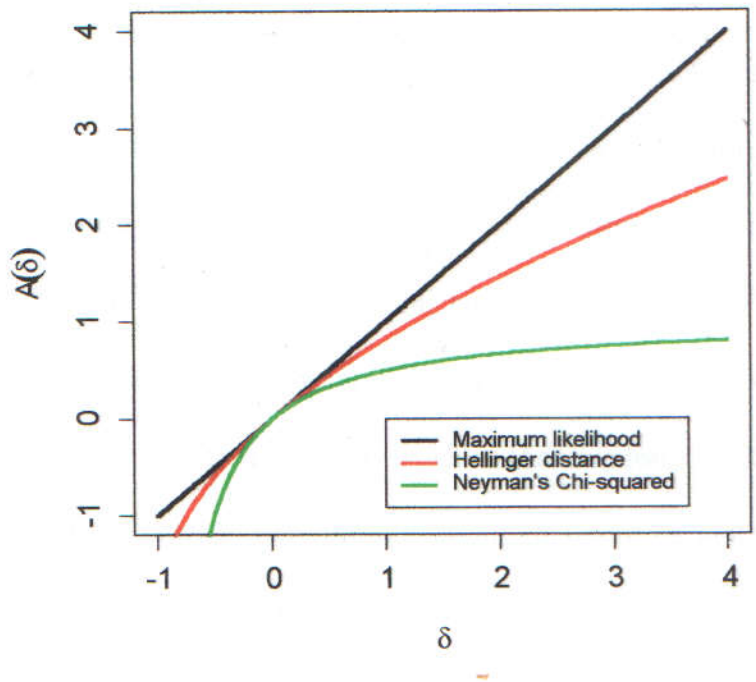

Figure 2.2: Residual adjustment functions for maximum likelihood, Hellinger distance and Neyman's chi-squared estimators.

likelihood when ∂ is close to 0 and therefore the least efficient. Thus it is the curvature of these RAF's, determined by $A''(0)$,which controls the trade-off between robustness and efficiency for these estimators.

A major drawback with these RAF estimators, however, is that they often utilise kernel density estimation to provide smooth estimates of g. Although much work has been done to establish the general conditions which the bandwidth of the kernel density estimate (denoted h) must satisfy to ensure the invariance and consistency of the resulting estimators there is

little guidance available for deciding what value to place on the bandwidth in practice. Several different methods of bandwidth selection have been utilised in the literature, for example, when studying the Hellinger distance estimator in 1986 Tamura and Boos [36] chose h to be the minimiser of the asymptotic integrated mean squared error of the estimator g under normality [32] whereas in 1989 Simpson [34] used $h = (35e)^{\frac{1}{5}} \left(\frac{\pi}{8}\right)^{\frac{1}{10}} \sigma n^{-\frac{1}{5}}$ which had been proposed by Devroye and Györfi [9]. More recently, second generation bandwidth selection methods (such as that suggested by Sheather and Jones [31]) have become popular and might therefore be more appropriate choices but, as with their predecessors, these methods were developed for use in density estimation and their effect on the performance of parametric estimators has not been fully assessed. Basu and Lindsay [5] suggested applying the same smoothing to the model and data so that the consistency and asymptotic normality of the estimators are independent of h but again this approach does not suggest specific values for the bandwidth. One way of avoiding these problems is to restrict the use of these methods to discrete models so that the empirical density function could estimate g thus making the kernel density estimate redundant but this matter does need to be addressed in the continuous case if the full potential of these extremely promising estimators is to be exploited.

Furthermore, Lindsay also discovered that by modifying the weights in both

the Chi-squared and Hellinger distances he was able to generate two new families of distance measure with a much wider range of robustness and efficiency properties. This means that, in theory at least, these distances could be tailored to meet one's needs exactly but the problem of bandwidth selection remains and is now further complicated by the introduction of the additional weighting parameters.

Consequently, although the study of RAF estimators has undoubtedly led to the development of several new estimators which have extremely desirable theoretical properties, because of the problems regarding the bandwidth these methods are not yet practical alternatives to the distribution based methods, such as the Cramér-von-Mises estimator, which are currently in use. (Distribution based estimation methods are generally easier to implement than density based methods because the empirical distribution function which is used to estimate G does not require smoothing.)

2.9.5 Density power divergences

Another class of minimum distance estimator is obtained by minimising a density power divergence in which the power parameter directly controls the robustness and efficiency of the resulting estimators. An example is the Cressie-Read family of divergences $I^{\lambda}\left(f_{\theta}, g\right) = \frac{1}{\lambda(\lambda+1)} \int f_{\theta}\left(x\right) \left(\left(\frac{f_{\theta}(x)}{g(x)} \right)^{\lambda} - 1 \right) dx$

which was introduced in 1984 [8]. This family includes several RAF estimators such as Pearson's Chi-squared ($\lambda = 1$), Neyman's Chi-squared ($\lambda = -2$) and the Hellinger distance $\left(\lambda = \frac{1}{2}\right)$ so by choosing suitable values for λ estimators with a wide range of asymptotic properties can be obtained. Although this flexibility is useful in theory it does make the method tricky to implement in practice because the choice of λ is so vital to its performance. If the λ chosen is inappropriate for the data the resulting estimators may not have the desired asymptotic properties but the distribution of the data is usually unknown so it is not clear how the decision regarding the choice of λ should be made. Furthermore, in practice many of these estimators require a smooth density estimate and so the problem of bandwidth selection occurs yet again.

In 1998 Basu, Harris, Hjort and Jones [4] proposed a density power divergence (described in detail in Section 3.1, p.34) which doesn't need smoothing because the form of the estimating equation is such that the empirical density function f_n can be used to estimate the true density g. However, as with the Cressie-Read estimators, the performance of this estimator depends on a power parameter which is unknown and yet must be appropriate for the data to which the method is applied.

2.10 Barriers to the wider use of minimum distance estimators

Several new families of minimum distance estimator have been introduced as a result of attempts to optimise robustness and efficiency but unfortunately it is a common feature of many that their highly desirable theoretical properties are difficult to attain in practice. Their performance often depends on unknown parameters which may be thought of as tuning parameters because they, either directly or indirectly, determine the robustness and efficiency of the resulting estimators. The RAF estimators, for example, are dependent on the choice of bandwidth for the kernel density estimate and the density power divergences on the choice of power. In the literature little attention has been paid to these matters with most papers suggesting suitable values for data from a particular distribution without acknowledging that in practice this distribution is rarely known nor considering the sensitivity of estimators to this uncertainty.

2.11 Aim and rationale for this research

The main goal of this research is therefore to develop a method for selecting appropriate values for tuning parameters so that the full potential of

this class of methods can be realised. In order to do this one must decide on some basis for determining the optimum value for a parameter and it seems reasonable to suggest that this should be the one which optimises the performance of the estimator. As indicated in Section 2.7, the asymptotic mean squared error ($AMSE$) is an appropriate joint measure of robustness and efficiency so by minimising an expression for the $AMSE$ of the estimator, which is a function of the tuning parameter, it is hoped that the optimal value for the tuning parameter might be obtained. The feasibility and effectiveness of this new approach will be investigated by applying the method to simulated data using a variety of different distance functions. Three minimum distance estimators in particular will be considered here, each one chosen because of its robustness and efficiency from a different class, with just one tuning parameter to estimate, so that the general applicability of the method might be established. First the density power divergence which was introduced by Basu, Harris, Hjort and Jones [4] will be studied as an example of a minimum distance estimator with an unknown power parameter, then the Hellinger distance [6] because it requires smoothing and finally Öztürk and Hettmansperger's criterion function [23] which is interesting because it is distribution function based and has an unknown power parameter. These basic methods will be described in Chapter 3 before considering their $AMSE$ functions in order to optimise the choice

of tuning parameter in Chapters 4 to 6.

Chapter 3

Introduction to the three minimum distance estimators

3.1 Minimum density power divergence

The family of density power divergences was introduced by Basu, Harris, Hjort and Jones [4] and is defined as

$$d_\alpha(g, f_\theta) = \int \left[f_\theta^{\alpha+1}(x) - \left(1 + \frac{1}{\alpha}\right) g(x) f_\theta^\alpha(x) + \frac{1}{\alpha} g^{\alpha+1}(x) \right] dx \text{ for } \alpha > 0 \tag{3.1}$$

where f_θ is the model density, g is the true density.

The above expression is undefined for $\alpha = 0$ and is therefore redefined, in

this special case, as

$$d_0(g, f_\theta) = \lim_{\alpha \to 0} d_\alpha(g, f_\theta).$$

Applying L'Hôpital's rule and utilising the fact that $\lim_{h \to 0} \frac{(t^h - 1)}{h} = \log t$ leads to

$$d_0(g, f_\theta) = \int g(x) \ln(g(x)/f_\theta(x)) dx$$

which is the Kullback-Leibler divergence.

This distance measure, which for convenience will be referred to as BHHJ, was developed from the L_2 distance estimator which is the minimiser of $\int (f_\theta(x) - g(x))^2 dx$. The L_2 distance is commonly used in smoothing applications [33] and its properties as a method for parametric estimation were reported recently by Scott [30] who found the L_2 estimators to be highly robust but very inefficient. The power parameter was introduced by Basu, Harris, Hjort and Jones in (3.1) to remedy this deficiency by controlling the robustness and efficiency of the resulting estimators. For example, when $\alpha = 0$ the distance reduces to the Kullback-Leibler divergence which is equivalent to using maximum likelihood estimation and therefore fully efficient. When $\alpha = 1$ the estimating equation reduces to that for the L_2 distance and the estimators produced will be highly robust. By choosing α somewhere between 0 and 1 it is hoped that a trade-off between these two

extremes might be attained and provide robust estimators with an acceptable degree of efficiency. Choosing $\alpha > 1$ does little to improve robustness and leads to greatly reduced efficiency so it is the range $0 \leq \alpha \leq 1$ which is of interest.

If g were known, the BHHJ estimator would be the value of θ which minimises the distance function (3.1) for a given value of α and is therefore denoted θ_α. It is obtained by differentiating $d_\alpha(g, f)$ by θ and setting equal to zero to give

$$0 = \int \left[f_{\theta_\alpha}^{\alpha+1}(x) u_{\theta_\alpha}(x) - f_{\theta_\alpha}^{\alpha}(x) u_{\theta_\alpha}(x)\, g(x) \right] dx$$

where $u_\theta = \frac{\partial \log f_\theta}{\partial \theta}$ is the score function.

Note that when $\alpha = 0$ the estimating equation is $0 = \int g(x) u_\theta(x) dx$ and the method is therefore equivalent to maximum likelihood.

The integral over g is the expected value of $f_\theta^\alpha(x) u_\theta(x)$ and can therefore be estimated by taking the average of the function over the data. This means that g does not need to be estimated directly and so the issue of smoothing does not arise. Hence the data-based estimating equation does not involve g and is as follows

$$0 = \int f_{\widehat{\theta}_\alpha}^{\alpha+1}(x) u_{\widehat{\theta}_\alpha}(x)\, dx - \frac{1}{n} \sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) u_{\widehat{\theta}_\alpha}(X_i) \tag{3.2}$$

where $\widehat{\theta}_\alpha$ is the BHHJ estimate of θ obtained from data.

3.1.1 Asymptotic properties

The asymptotic properties of BHHJ estimators were derived by Basu, Harris, Hjort and Jones and presented in outline only [4]. Full details of the derivation, which uses the Taylor series approximation to the estimating equation (3.2), are therefore given in Appendix A.1 on page 209. The asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}_\alpha - \theta_\alpha\right)$ is shown to be *Normal* with mean 0 and variance $J^{-1}KJ^{-1}$ where

$$K = \int f_{\theta_\alpha}^{2\alpha}(x) u_{\theta_\alpha}(x)\, u_{\theta_\alpha}^T(x)\, g(x)dx - \left[\int f_{\theta_\alpha}^{\alpha}(x) u_{\theta_\alpha}(x)\, g(x)dx\right]\left[\int f_{\theta_\alpha}^{\alpha}(x) u_{\theta_\alpha}(x)\, g(x)dx\right]^T$$

$$\text{and } J = \int u_{\theta_\alpha}(x)\, u_{\theta_\alpha}^T(x)\, f_{\theta_\alpha}^{\alpha+1}(x)dx + \int \left(i_{\theta_\alpha}(x) - \alpha u_{\theta_\alpha}(x)\, u_{\theta_\alpha}^T(x)\right)\left(g(x) - f_{\theta_\alpha}(x)\right) f_{\theta_\alpha}^{\alpha}(x)dx$$

and θ_α and $\widehat{\theta}_\alpha$ are the true and data-based BHHJ estimates of θ, u_θ is the score function and $i_\theta = -\partial\{u_\theta\}/\partial\theta$ is the observed Fisher information of the model.

3.1.2 Robustness

The robustness properties of $d_\alpha(g, f)$ can be investigated by considering the influence function

$$IF(\xi) = J^{-1}\left(u_\theta(\xi) f_\theta^\alpha(\xi) - \int f_\theta^\alpha(x) u_\theta(x) g(x) dx\right)$$

where u_θ is the score function, g is a probability density function and J is as defined in the previous section. The derivation of this function is detailed in Basu, Harris, Hjort and Jones [4] and outlined in Appendix B.1 on page 237.

When $f_\theta = g = N(\theta, 1)$ this reduces to

$$IF(\xi) = (\xi - \theta)\phi_\theta^\alpha(\xi)(2\pi)^{\frac{\alpha}{2}}(1+\alpha)^{\frac{3}{2}} \tag{3.3}$$

which is plotted for several values of α in Figure 3.1 (p.39). The key feature is that the influence function is bounded for all $\alpha > 0$ and should therefore be robust for all $\alpha > 0$. Furthermore, for any $\alpha > 0$ the influence curve redescends towards zero with the speed of descent varying according to the value of α. For example when $\alpha = 1$ and $f_\theta = g = N(0, 1)$ data points which are outside the range ± 3 have virtually no effect on the parameter estimates whereas when $\alpha = 0.5$ points need to be greater than 4 to be downweighted to the same degree. Thus the robustness properties of this method depend on α.

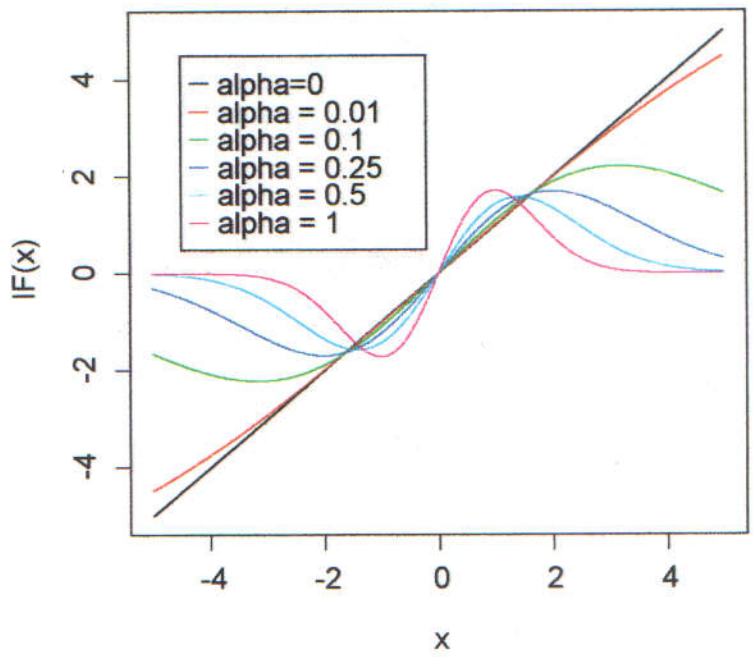

Figure 3.1: Influence function for the BHHJ estimator when $f_\theta = N(0,1)$.

Replacing the true density g with the empirical density function f_n gives a data-based estimate of the distance function

$$\widehat{d}_\alpha(f_n, f_\theta) = \int \left[f_\theta^{\alpha+1}(x) - \left(1 + \frac{1}{\alpha}\right) f_n(x) f_\theta^\alpha(x) + \frac{1}{\alpha} f_n^{\alpha+1}(x) \right] dx. \quad (3.4)$$

Since the BHHJ estimates are obtained by minimising this data-based estimate of the distance function, and not $d_\alpha(g, f_\theta)$, it seems sensible to consider the data-based estimate of the theoretical influence function

$$\widehat{IF}(\xi) = \widehat{J}^{-1} \left(u_{\widehat{\theta}_\alpha}(\xi) f_{\widehat{\theta}_\alpha}^\alpha(\xi) - \frac{1}{n} \sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^\alpha(X_i) u_{\widehat{\theta}_\alpha}(X_i) \right) \quad (3.5)$$

where

$$\widehat{J} = \frac{1}{n}\sum_{i-1}^{n}\left(i_{\widehat{\theta}_\alpha}(X_i) - \alpha u_{\widehat{\theta}_\alpha}(X_i)\, u^T_{\widehat{\theta}_\alpha}(X_i)\right) f^{\alpha}_{\widehat{\theta}_\alpha}(X_i)$$
$$+ \int \left((1+\alpha)\, u_{\widehat{\theta}_\alpha}(x)\, u^T_{\widehat{\theta}_\alpha}(x) - i_{\widehat{\theta}_\alpha}(x)\right) f^{\alpha+1}_{\widehat{\theta}_\alpha}(x)dx.$$

When $f_{\widehat{\theta}_\alpha} = N(\widehat{\theta}_\alpha, 1)$ this simplifies to

$$\widehat{IF}(\xi) = \frac{\left(\xi - \widehat{\theta}_\alpha\right)\phi^{\alpha}_{\widehat{\theta}_\alpha}(\xi) - \frac{1}{n}\sum_{i=1}^{n}\phi^{\alpha}_{\widehat{\theta}_\alpha}(X_i)\left(X_i - \widehat{\theta}_\alpha\right)}{\frac{1}{n}\sum_{i=1}^{n}\left[(1-\alpha(X_i-\widehat{\theta}_\alpha)^2) f^{\alpha}_{\widehat{\theta}_\alpha}(X_i)\right] - \alpha(2\pi)^{-\frac{\alpha}{2}}(1+\alpha)^{-\frac{3}{2}}} \quad (3.6)$$

The theoretical and data-based influence functions (equations 3.3 and 3.6) when $\alpha = 0.1$ are plotted for a data set drawn from $N(\theta, 1)$ with 10% contamination at 10 in Figure 3.2 (p.41). The two curves differ slightly, as one would expect, but they have the same general shape and clearly show that robustness to outliers is a feature offered by both $d_\alpha(g, f_\theta)$ and $\widehat{d}_\alpha(f_n, f_\theta)$.

The key to further understanding the robustness of BHHJ is to consider how $\widehat{d}_\alpha(f_n, f_\theta)$ is affected by the value given to α and the degree of contamination in the data. The effect of the value of α on the distance function $\widehat{d}_\alpha(f_n, f_\theta)$ is illustrated in Figure 3.3 (p.42) which shows how, for a sample of $N(0, 1)$ data with 10% contamination at 10, the position of the global minimum remains close to $\theta = 0$ for any $\alpha > 0$. This suggests that for this particular data set $\alpha = 0.1$ would be sufficient to provide robust estimators.

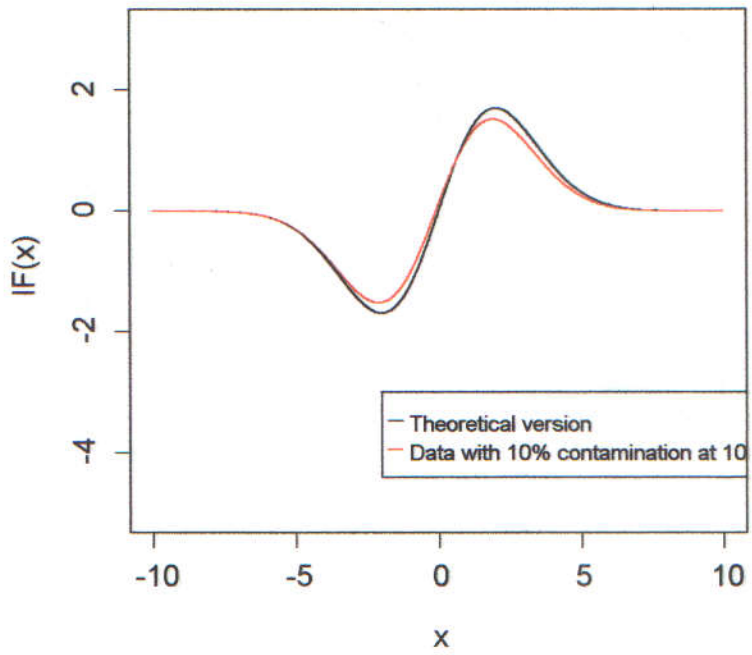

Figure 3.2: Comparison between the theoretical and data-based versions of the influence function for the BHHJ estimator when $\alpha = 0.1$ and $f_\theta = N(0, 1)$.

Figure 3.4 (p.43) shows how the global minimum of the estimated distance function shifts from approximately zero to the contamination point as the degree of contamination increases. This illustrates the method breaking down but it should be noted that breakdown does not necessarily result in the parameter estimates jumping from one region to another, as in this case. When $\alpha = 0$ the breakdown point is 0 and there are no sudden changes in the location of the global minimum so the parameter estimates move steadily

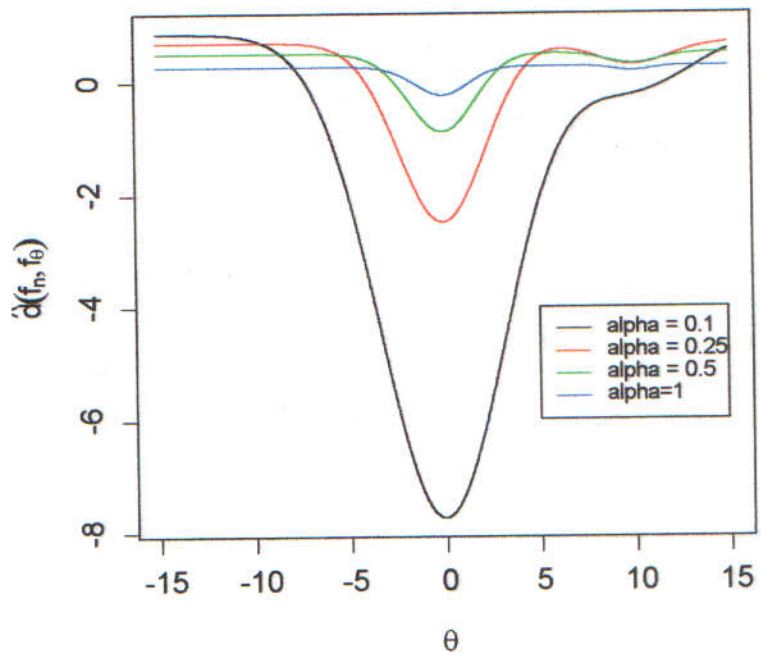

Figure 3.3: $\hat{d}_\alpha(f_n, f_\theta)$ for $N(0,1)$ data with 10% contamination at 10.

away from the target value as ε increases, as illustrated in Figure 1.1, p.4. For $\alpha > 0$ breakdown occurs in a similar way to that shown in Figure 3.4, with the breakdown point varying between 50% when $\alpha = 0.1$ and 40% when $\alpha = 1$. Thus, the breakdown point does not depend solely on the percentage of contamination in the data; the value of α also plays a role.

By changing the location of the contamination point whilst keeping the percentage of contaminants fixed another aspect of the robustness of the BHHJ estimator, closely related to the influence function, can be demonstrated.

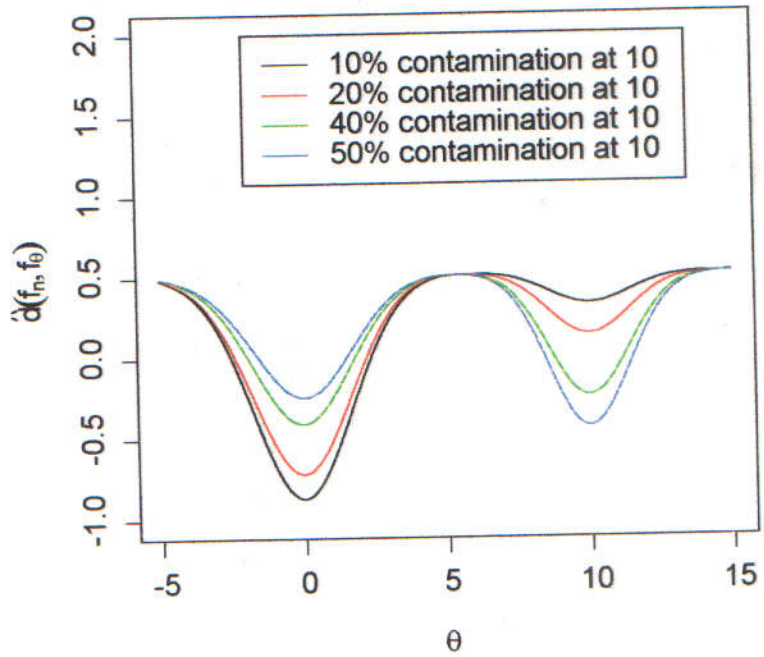

Figure 3.4: $\hat{d}(f_n, f_\theta)$ with $\alpha = 0.5$ for $N(0,1)$ data with differing degrees of contamininination at 10.

The global minimum of $\widehat{d}_\alpha(f_n, f_\theta)$, shown for a variety of contamination points in Figure 3.5 (p.44) barely moves from the target value as the contamination point moves away from the mean and therefore confirms that the estimation method can be extremely robust for some α.

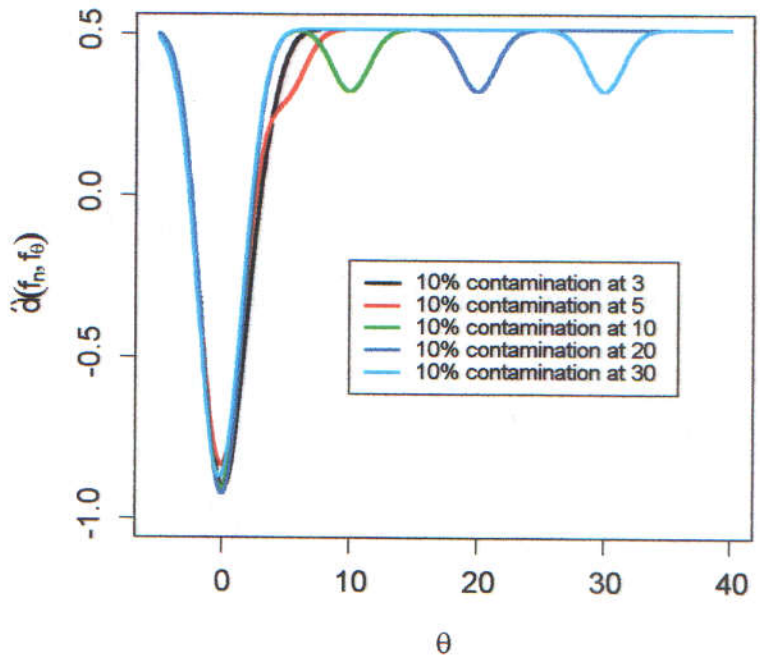

Figure 3.5: $\hat{d}(f_n, f_\theta)$ with $\alpha = 0.5$ for $N(0,1)$ data with 10% contamination at various points.

3.1.3 Locating the estimates of θ_α

The minimiser of $\widehat{d}_\alpha(f_\theta, g)$ was found using a quasi-newton optimisation procedure. In the one-parameter problem this function has two possible minima. When the percentage of contamination is small (generally $\varepsilon < 0.2$) the global minimum is close to the target value but, as the degree of contamination increases to around 20% or 30%, the global minimum may shift to the contamination point. As mentioned in the previous section, the exact point at which this breakdown occurs depends on both the value of α and the distribution of the data f_n. The two-parameter case is slightly more complicated in that the global minimum can now shift for either or both parameters, independently of each other. In order to be certain that it was the global minimum that was found in each simulation the search area was split into several regions, each one containing only one possible minimum. The search procedure was then applied to each region in turn and the global minimum found by comparing the results.

3.1.4 Simulations

The method was applied to several simulated data sets to confirm that this estimation method is indeed robust and also to illustrate that this robustness can be controlled by α. The results of these simulations for several values of

α are shown in Table 3.1 (p.47). These data sets were generated by taking 100 random samples of size 100 from each distribution (with the random seed reset for each sample). The BHHJ method was then applied to each sample with α set as 0, 0.01, 0.1 and 1 in turn so for each distribution and value of α 100 parameter estimates were obtained. The average of the parameter estimates for each distribution and α is given in column 3 of Table 3.1 and their mean squared error in column 4. Putting $\alpha = 1$ produces highly robust results for the contaminated data but performs less well for $N(0,1)$ or t_2 data. When $\alpha = 0$ the large mean squared errors for estimates from the contaminated data illustrate clearly the maximum likelihood estimator's lack robustness. The mean squared errors summarise both robustness and efficiency and therefore follow a different pattern to the average $\widehat{\theta}_\alpha$'s but again clearly suggest that the degree of robustness and efficiency achieved depends on both the choice of α and the distribution of the data.

In conclusion, for $0 < \alpha < 1$ the BHHJ estimators are less efficient than those of maximum likelihood and less robust than L_2 so if either of these properties alone is desired there is nothing to be gained by using this method. The main attraction of these estimators is that they offer a compromise between these two properties and might be both robust and efficient. In practice the performance of this method will depend on whether a

Table 3.1: Simulation results for the BHHJ estimator when $f_\theta = N(\theta, 1)$

Distribution	α	**Average** $\widehat{\theta}_\alpha$	**Mean Squared Error** $(\widehat{\theta}_\alpha)$
$\phi(x)$	0	**-0.014**	**0.011**
	0.01	-0.007	0.011
	0.1	-0.007	0.011
	1	-0.010	0.016
$0.9\ \phi(x) + 0.1\ \Delta_{10}(x)$	0	**1.030**	**1.200**
	0.01	0.664	0.491
	0.1	0.015	0.011
	1	-0.001	0.019
$0.8\ \phi(x) + 0.2\ \Delta_{10}(x)$	0	**2.040**	**4.300**
	0.01	1.514	2.447
	0.1	0.017	0.015
	1	-0.012	0.023
t_2	0	**0.060**	**0.280**
	0.01	-0.008	0.037
	0.1	0.009	0.024
	1	0.019	0.031

Key: $\phi(x) = N(0, 1)$ and $\Delta_{10}(x)$ is a contamination data point equal to 10.

suitable value for α can be found and this will not be easy given that the distribution from which the data is sampled and the degree of contamination is rarely known. It is vital therefore that some reliable way of choosing α from data is devised so that the theoretical benefits of using this estimation method may be realised.

3.2 Hellinger distance

The Hellinger distance (HD) was first used for continuous models by Beran [6] and is defined as

$$HD(\theta) = \int \left(f_\theta^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x) \right)^2 dx \tag{3.7}$$

where f_θ is the model density and g is the true density.

If g were known, the HD estimate of θ would be the value of θ which minimises this distance and is obtained by differentiating equation (3.7) with respect to θ and setting it equal to 0. This leads to the estimating equation

$$0 = -\int g^{\frac{1}{2}}(x) f_\theta^{\frac{1}{2}}(x) u_\theta(x)\, dx \tag{3.8}$$

where u_θ is the score function.

Since the true density of the data, g, is not known it must be estimated from the data. Using a kernel density estimate of g, denoted by $\widehat{g}_n$, to replace g in equation (3.8) and subscripting the parameter estimates by h to show their dependence on the bandwidth used, gives a revised estimating equation as follows

$$0 = -\int \widehat{g}_n^{\frac{1}{2}}(x) f_{\widehat{\theta}_h}^{\frac{1}{2}}(x) u_{\widehat{\theta}_h}(x)\, dx. \tag{3.9}$$

The kernel density estimation procedure will be described in detail later in Section 5.1.1, p.117.

3.2.1 Asymptotic properties

The asymptotic properties of HD estimators are derived using a Taylor series approximation to the estimating equation (3.9). Full details of this are given in Appendix A.2 which shows that (in common with the BHHJ estimator in Section 3.1, p.34) the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_h - \theta)$ is *Normal* with mean 0 and variance $J^{-1}KJ^{-1}$. However, in this case, J and K are re-defined as follows,

$$J = \int \left[f_\theta^{-\frac{3}{2}}(x) \left[\frac{df_\theta}{d\theta}\right] \left[\frac{df_\theta}{d\theta}\right]^T - 2 f_\theta^{-\frac{1}{2}}(x) \frac{d^2 f_\theta}{d\theta^2} \right] g^{\frac{1}{2}}(x)\, dx$$

and

$$K = \int \left[\frac{df_\theta}{d\theta}\right] \left[\frac{df_\theta}{d\theta}\right]^T f_\theta^{-1}(x)\, dx - \psi\psi^T$$

and $\psi = \int \left[\frac{df_\theta}{d\theta}\right] f_\theta^{-\frac{1}{2}}(x)\, g^{\frac{1}{2}}(x)\, dx$. Here, $\frac{d^2 f_\theta}{d\theta^2}$ is the matrix of second derivatives of f_θ with respect to θ.

An important feature of HD estimators is that they are asymptotically equivalent to maximum likelihood estimators at the model and therefore fully efficient. They also share the same influence function and yet, in contrast to the maximum likelihood estimator, they are robust.

When $f_\theta = g$ the asymptotic variance of the HD estimator reduces to $\left[\int \left[\frac{df_\theta}{d\theta}\right] \left[\frac{df_\theta}{d\theta}\right]^T f_\theta^{-1}(x)\, dx\right]^{-1}$ which is the same as the variance of the maxi-

mum likelihood estimator and thus confirms that HD estimators are asymptotically fully efficient.

3.2.2 Robustness

The influence function for the Hellinger Distance, $IF(\xi) = \frac{d\theta_\varepsilon}{d\varepsilon} = (\xi - \theta)$, is the same as that for the maximum likelihood estimator and is unbounded. Details of this derivation were given by Beran [6] and are outlined in Appendix B.2 on page 239. An unbounded influence function generally indicates that an estimator will not be robust and yet simulation studies by Simpson [34], Tamura and Boos [36] and Basu and Lindsay [5] indicate that the HD estimator is very robust indeed. This unexpected robustness is explained by Lindsay [19] who showed that the influence function for the HD estimator is a first-order approximation to bias whereas for maximum likelihood the influence function is exact. In this paper Lindsay demonstrated that the next term in this approximation is of a similar size to the first, but opposite in sign. Therefore, taking a second-order approximation to bias, the second term balances the first and thus the effect of any outlying data point on the bias is far less than that suggested by the influence function.

A better way of assessing the robustness of the HD estimator is therefore to examine the estimating equation and consider how it might be affected

by changes to the distribution of the data or the bandwidth used to obtain the kernel density estimate. The $\sqrt{f_\theta}$ term in the estimating equation plays a key role in ensuring robustness because it has the effect of downweighting extreme values. Figure 3.6 (p.52) illustrates how the objective function (using the $f_\theta = N(\theta, 1)$ model for data from the $N(0, 1)$ distribution with 10% contamination at various points) changes as the contamination point gets larger. Although the shape of the distance function is determined by the position of the contamination point the global minimum moves very little irrespective. In common with other density based robust estimation methods, this method is more sensitive to contamination occurring at points which are likely under the model and therefore, as confirmed by Figure 3.6 (p.52), the method is less able to cope with contamination at 3 than at 5. The $\sqrt{f_\theta}$ term is greater at 3 than at 5 and so downweights this point to a lesser extent. The value of the integrand in the estimating equation (3.8) becomes extremely small (≤ 0.0001) for x greater than 7 and thus large outliers have very little influence on the estimating procedure.

Having considered how the position of the contamination point affects the estimating procedure the next step is to examine what happens as the percentage of contamination increases. The Hellinger Distance functions for N(0,1) data with various percentages of contamination at 10 are shown in Figure 3.7 (p.53). The global minimum jumps from the neighbourhood of

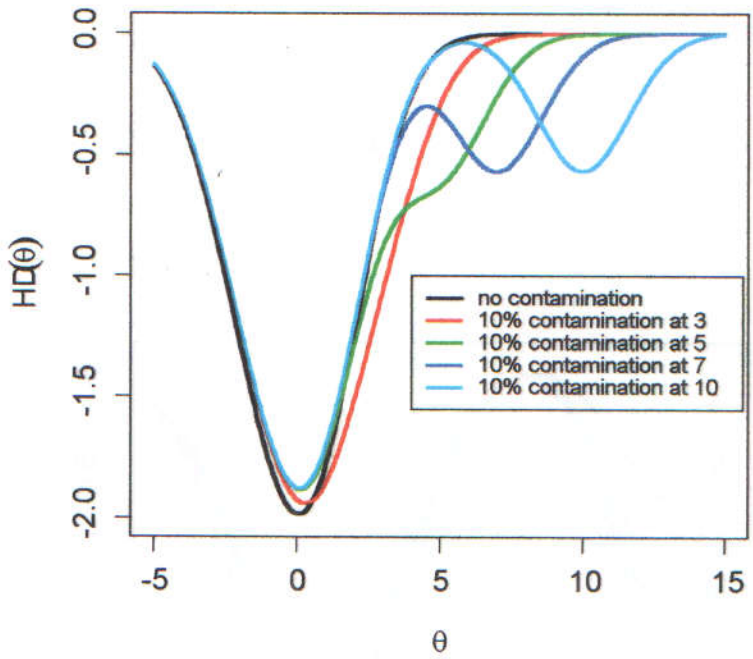

Figure 3.6: The Hellinger distance with bandwidth $= 0.5$ for $N(0,1)$ data with 10% contamination.

the target value to the contamination point when the contamination percentage is in excess of 50%. Thus the HD estimators are highly robust and can tolerate as much as 50% of the data being contaminated. When the contamination point is at 3, as illustrated in Figure 3.8 (p.54), the effect is less dramatic with the global minimum moving slowly in the direction of the contamination point rather than jumping.

The choice of bandwidth h for the kernel density estimate is another factor

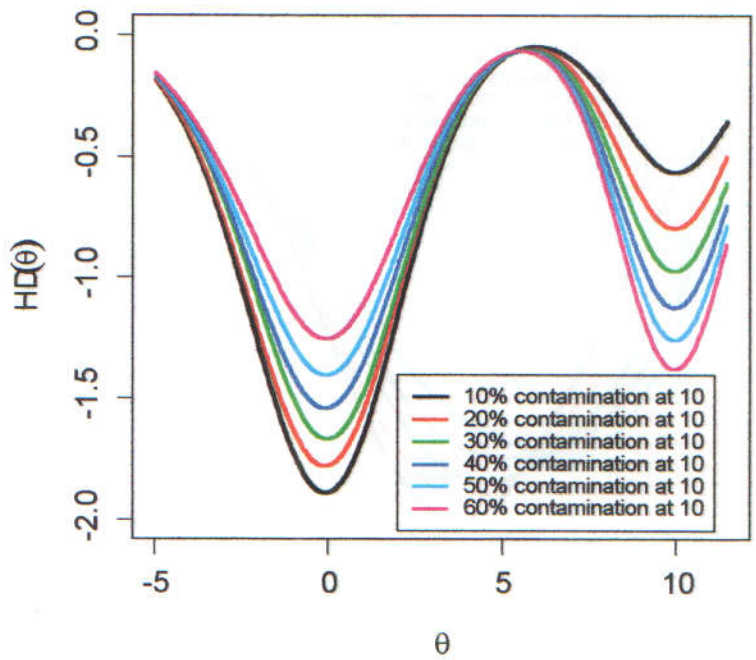

Figure 3.7: The Hellinger distance with bandwidth = 0.5 for N(0,1) data with differing percentages of contamination at 10.

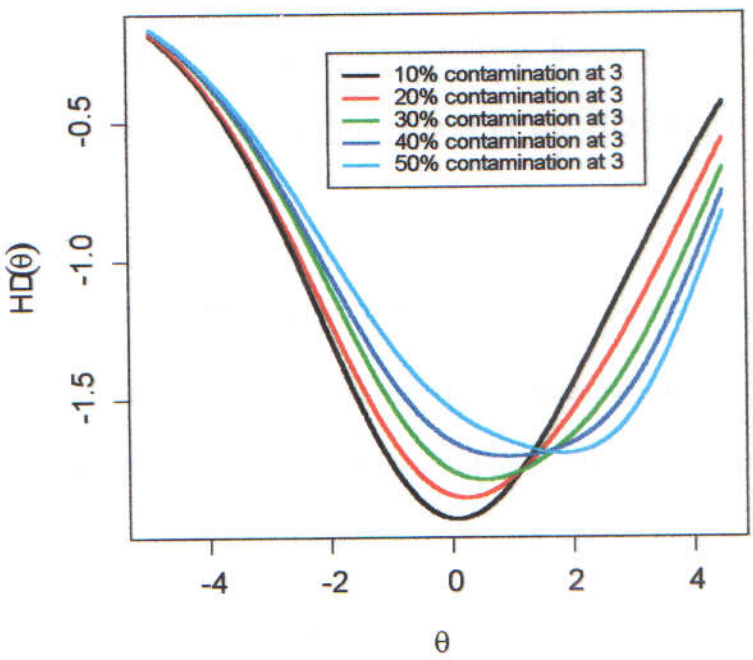

Figure 3.8: The Hellinger distance with bandwidth = 0.5 for N(0,1) data with differing percentages of contamination at 3.

which might be thought to affect robustness. As can be seen in Figure 3.9 (p.56), which shows how the kernel density estimate for a particular data set changes with the bandwidth, when h is small the kernel density estimate of g puts a very narrow spike of probability around the contamination point. As h increases the data density estimate becomes smoother and flatter making this spike less pronounced as a wider range of points are given a non-zero probability. The general shape of the density is, however, unchanged making it difficult to predict how the robustness of the HD estimators might be affected.

Figures 3.10 and 3.11 illustrate that the choice of bandwidth does not necessarily affect robustness greatly. When the contamination is at 10, as can be seen in Figure 3.10 (p.57), adjusting the bandwidth has little effect on the robustness of the method. When the contamination percentage is less than 50% the Hellinger distance estimate $\widehat{\theta}_h$ will be close to the target value (zero in this case) irrespective of the value chosen for h thus making $\sqrt{f_{\widehat{\theta}_h}(10)}$ also close to zero irrespective. However, when the contamination is at 3 for example, the robustness of the method is more readily affected by the choice of h. Figure 3.11 (p.58) shows how, when the degree of contamination at 3 is 20%, the minimum of the Hellinger Distance function shifts slowly away from the target value as the bandwidth increases. This suggests that in situations where the contamination is not fully downweighted by the model it

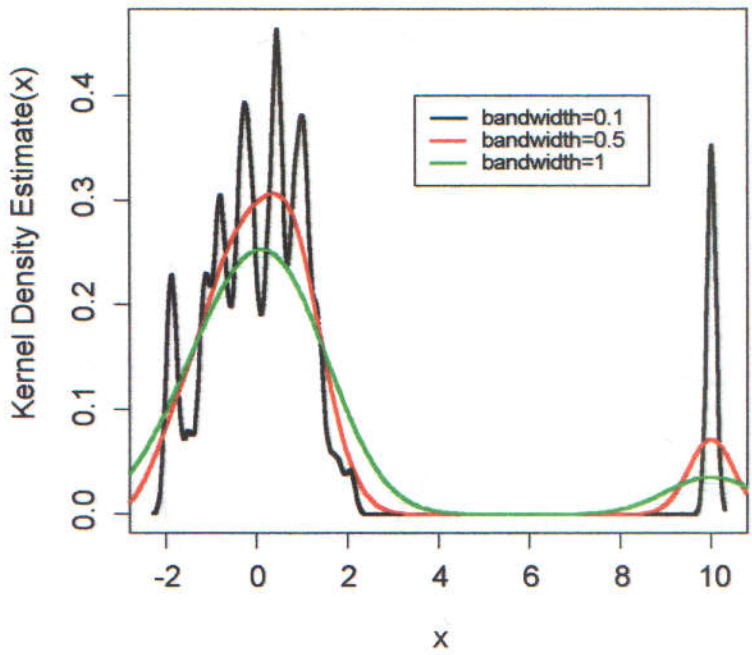

Figure 3.9: Kernel density estimates for N(0,1) data with 10% contamination at 10.

may be possible to improve robustness slightly by choosing h appropriately. Furthermore, it seems likely that the potential for improving robustness will be greater the higher the degree of contamination in the data.

3.2.3 Locating the estimates of θ_h

The HD estimates, $\widehat{\theta}_h$, can be obtained by solving the estimating equation (3.8) but, as illustrated in Figure 3.12 (p.60), this function has multiple roots for many types of data. Furthermore, these multiple roots occur even

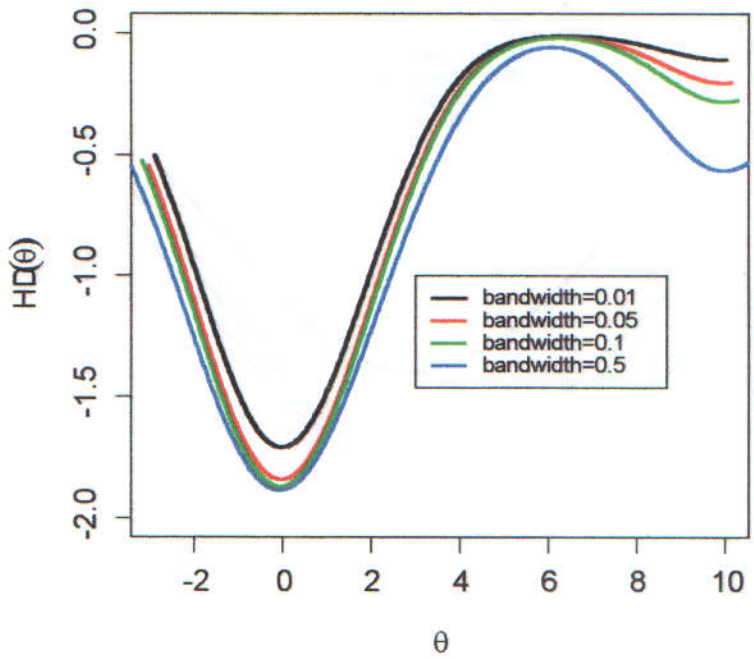

Figure 3.10: The Hellinger distance using different bandwidths for N(0,1) data with 10% contamination at 10.

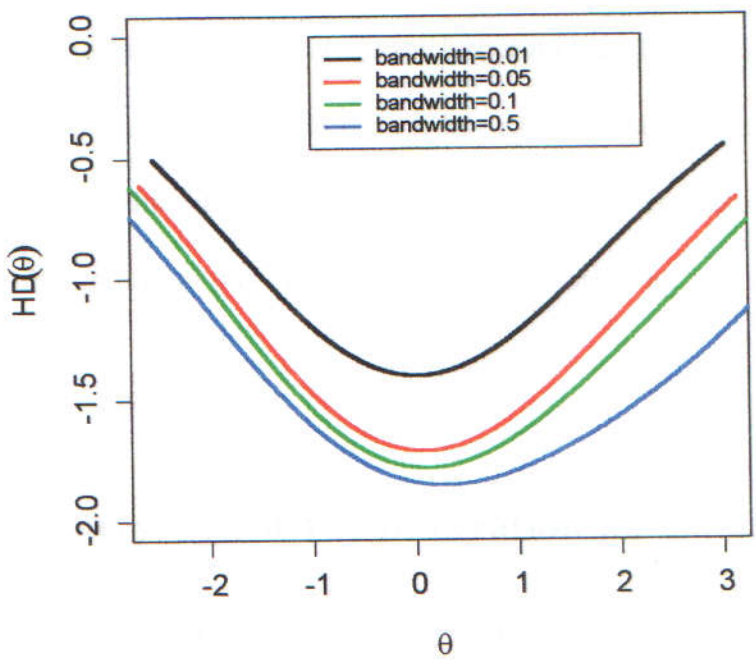

Figure 3.11: The Hellinger distance using different bandwidths for N(0,1) data with 20% contamination at 3.

when the percentage of contamination points in the data is small (10%) or heavy tailed (from the t_2 distribution). It is possible to solve this equation using numerical methods so long as the search area is restricted to ensure that it is the root relating to the global minimum which is found and not that for a local minimum. In the simulation setting, however, it is generally easier and safer to use a numerical optimisation procedure on the distance measure (Equation 3.7) instead. As with root finding it is necessary to direct the search into an appropriate region, but as demonstrated in Figure 3.13, the problem is simplified because for these types of data there are at most two possible minima, one close to zero and the other at the contamination point.

3.2.4 Carrying out the integration numerically

The integral in the estimating equation can not be derived explicitly so numerical methods are required but care needs to be taken because the function to be integrated is not standard. The kernel density estimate $\widehat{g}_n(x)$ is not defined for all values of x (typically existing only within the range of the data or the range of the data $\pm 3\sigma$) and although it is continuous within that range its smoothness depends on the bandwidth. This may lead to problems in applying numerical methods because the integrand changes direction rapidly and many numerical methods for integration are unable to

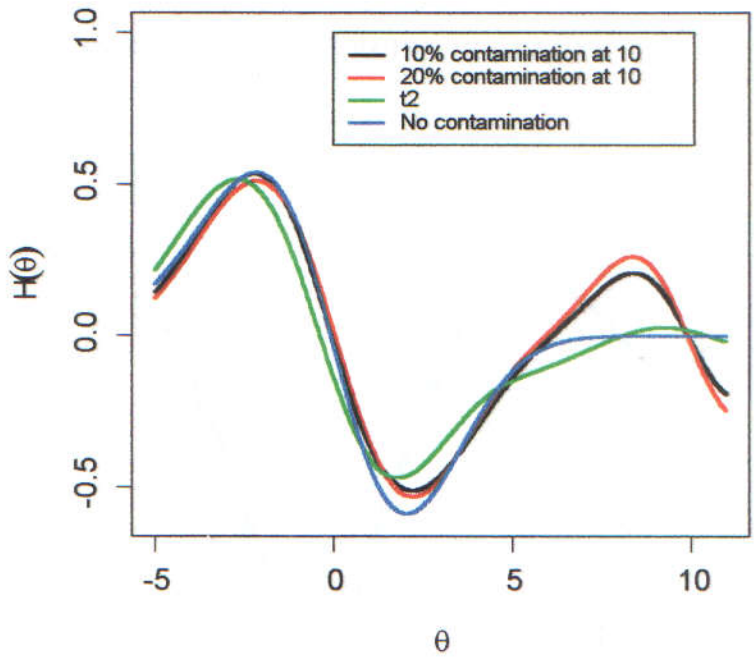

Figure 3.12: The estimating equation of the Hellinger distance estimator with bandwidth = 0.5 for data from different distributions.

cope with this. Furthermore, although in theory the kernel density estimate $(\widehat{g}_n)$ can be obtained for any value of x a considerable amount of computation can be avoided by opting for an integration method which requires the kernel density estimate to be known at predetermined points. In order to minimise computation the standard output from the kernel density estimation procedure was utilised and the integrals were therefore calculated using the trapezium rule with 512 strips.

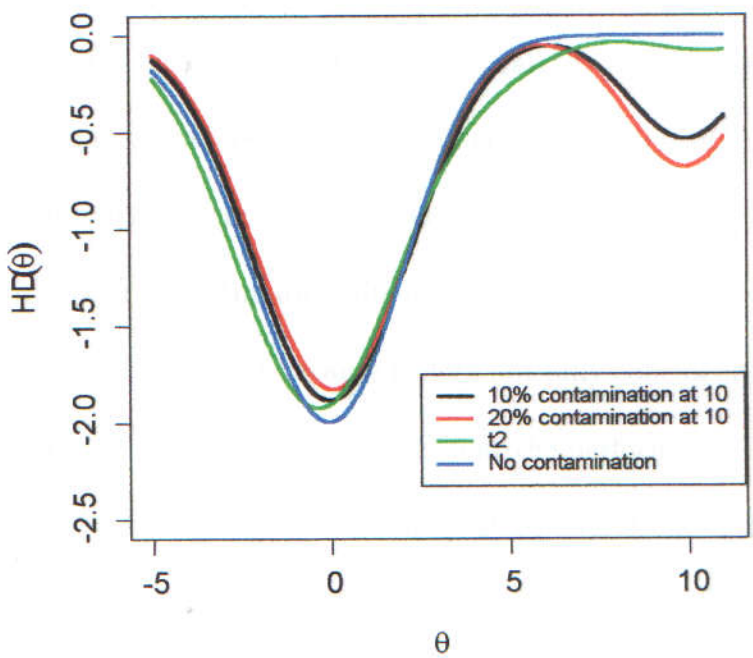

Figure 3.13: The Hellinger distance with bandwidth = 0.5 for data from different distributions.

3.2.5 Simulations

This method was applied to several simulated data sets to investigate what effect the choice of bandwidth might have on the resulting estimators. The results of these simulations, summarised in Table 3.2 (p.63), confirm that HD estimators are robust to both symmetric and asymmetric contamination whilst being highly efficient at the model. The choice of bandwidth does appear to affect the performance of these estimators but not greatly so. The potential for improving performance by choice of bandwidth is greatest when the contamination is at 3 where taking the bandwidth equal to 0.05 rather than 0.5 results in the MSE being halved. As a general rule it seems that choosing any small bandwidth will ensure good results but it not clear how, in practice, one would decide how small it should be. The role of the bandwidth h in HD estimation differs from that of the tuning parameter p in BHHJ (Section 3.1, p.34) because it does not fully control either robustness or efficiency. It is not clear, therefore, that the performance of the HD estimator could be optimised in the same way as BHHJ but such optimisation may not be needed since the method performs quite well for any value of h. The aim of any proposed bandwidth selection procedure should therefore be to provide sensible suggestions for bandwidth rather than optimal ones.

Table 3.2: Simulation results for the HD estimator when $f_\theta = N(\theta, 1)$

Distribution	h	**Average** $\widehat{\theta}_h$	**Mean Squared Error** $(\widehat{\theta}_h)$
$\phi(x)$	0.05	0.0025	0.0094
	0.1	0.0017	0.0090
	0.25	0.0014	0.0086
	0.5	0.0016	0.0087
	0.75	0.0020	0.0090
Maximum Likelihood	-	**0.0007**	**0.0084**
$0.9\phi(x) + 0.1\Delta_{10}(x)$	0.05	0.0083	0.0101
	0.1	0.0092	0.0097
	0.25	0.0081	0.0102
	0.5	0.0067	0.0108
	0.75	0.0058	0.0113
Maximum Likelihood	-	**0.9661**	**1.0123**
$0.8\phi(x) + 0.2\Delta_{10}(x)$	0.05	-0.0042	0.0154
	0.1	-0.0037	0.0150
	0.25	-0.0032	0.0145
	0.5	0.0062	0.0146
	0.75	0.0052	0.0147
Maximum Likelihood	-	**1.9798**	**4.0941**
$0.9\phi(x) + 0.1\Delta_3(x)$	0.05	0.0603	0.0158
	0.1	0.0862	0.0201
	0.25	0.1255	0.0286
	0.5	0.1464	0.0345
	0.75	0.1514	0.0362
Maximum Likelihood	-	**0.2868**	**0.0967**
t_2	0.05	0.0135	0.0250
	0.10	0.0144	0.0245
	0.25	0.0141	0.0241
	0.5	0.0138	0.0238
	0.75	0.0138	0.0237
	1	0.0136	0.0236
	1.25	0.0133	0.0236
	1.5	0.0127	0.0235
	1.75	0.0121	0.0235
Maximum Likelihood	-	**-0.0237**	**0.0930**

3.3 Öztürk and Hettmansperger's criterion function

This distribution based distance measure was proposed by Öztürk and Hettmansperger [23] and is defined as

$$d_F(\theta;p) = \int \left[G^p(x) - F_\theta^p(x)\right]^2 dx + \int \left[(1-G(x))^p - (1-F_\theta(x))^p\right]^2 dx \tag{3.10}$$

where $F_\theta(x)$ is the model distribution function, $G(x)$ is the true distribution function and $p > 0$. As in the section before last, the authors' initials will serve as an identifier so that from now on the method will be referred to simply as OH.

Estimating the true distribution function, G, by the empirical distribution function of the data, $F_n(x)$, gives the distance function

$$d_{F_n}(\theta;p) = \int \left[F_n^p(x) - F_\theta^p(x)\right]^2 dx + \int \left[(1-F_n(x))^p - (1-F_\theta(x))^p\right]^2 dx. \tag{3.11}$$

The OH estimate of θ, denoted $\widehat{\theta}_p$, is the minimiser of $d_{F_n}(\theta;p)$ which is

obtained by differentiating (3.11) with respect to θ and setting equal to zero as follows

$$\begin{aligned} 0 &= \int \left(F_n^p(x) - F_\theta^p(x)\right) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} dx \\ &\quad - \int \left[(1 - F_n(x))^p - (1 - F_\theta(x))^p\right] [1 - F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} dx. \end{aligned} \tag{3.12}$$

When $f_\theta = N(\theta, \sigma^2)$, by rewriting $F_n(x)$ as $\frac{1}{n}\sum I(X_i \leq x)$, revised estimating equations in which the integrals are replaced by sums of order statistics can be obtained and are as follows

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left\{c_i^* \left[1 - F_\theta\left(X_{(i)}\right)\right]^p - c_i F_\theta^p\left(X_{(i)}\right)\right\} \\ 0 &= -K(\infty, p-1) + \sum_{i=1}^{n} c_i K(X_{(i)}, p-1) + K(\infty, 2p-1) \\ &\quad + \sum_{i=1}^{n} c_i^* K^*\left(X_{(i)}, p-1\right) - K^*(\infty, 2p-1) \end{aligned}$$

where $c_i = (i/n)^p - ((i-1)/n)^p$, $c_i^* = ((n+1-i)/n)^p - ((n-i)/n)^p$, $K^*(t,m) = \int_{-\infty}^{t} [1 - \Phi_\sigma(x-\theta)]^{p-1} \phi_\sigma(x-\theta) \left(\frac{x-\theta}{\sigma}\right) dx$, $K(t,m) = \int_{-\infty}^{t} \Phi_\sigma^m(x-\theta) \phi_\sigma(x-\theta) \left(\frac{x-\theta}{\sigma}\right) dx$ and $X_{(i)}$ is the ith order statistic.

Thus for the Normal family of models, integrals involving F_n can be avoided and the amount of computation required significantly reduced. Unfortu-

nately this simplification does not hold for every choice of model because the integration by parts required is not always possible. (See Appendix C on page 246 for further details.)

The first term in the distance function (3.10) serves to reduce the effect of extreme positive values when $p < 1$ and extreme negative values when $p > 1$. For a very large positive observation the empirical distribution function will equal 1 and the model distribution function will be very close to, but slightly less than, 1. This makes the integrand approximately $[1 - F_\theta^p(x)]^2$ which is very small for $p < 1$. Similarly, for large negative observations the empirical distribution function will be very close to zero so this integrand will be approximately $F_\theta^{2p}(x)$ and the observation downweighted for $p > 1$. The second term has the opposite effect, downweighting extreme negative values when $p < 1$ and extreme positive values when $p > 1$. The combination of these two terms ensures that, in addition to being robust, the OH estimators are highly efficient at the model.

3.3.1 Asymptotic properties

The asymptotic properties of OH estimators are obtained by using a Taylor series approximation to the estimating equation. Full details are given in Appendix A.3 on page 227 which shows that (in common with the BHHJ

estimator in Section 3.1, p.34 and HD estimator in Section 3.2, p.47) the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}_p - \theta_p\right)$ is *Normal* with mean 0 and variance $J^{-1}KJ^{-1}$ but in this case J and K are as follows

$$
\begin{aligned}
J \;=\; & 2p(1-2p)\int F_\theta^{2p-2}(x)\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
& +2p(p-1)\int G^p(x)\,F_\theta^{p-2}(x)\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
& +2p(1-2p)\int (1-F_\theta(x))^{2p-2}\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
& -2p\int (1-G(x))^p(1-F_\theta(x))^{p-1}\frac{d^2F_\theta(x)}{d\theta^2}dx - 2p\int F_\theta^{2p-1}(x)\frac{d^2F_\theta(x)}{d\theta^2}dx \\
& +2p(p-1)\int (1-G(x))^p(1-F_\theta(x))^{p-2}\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
& +2p\int (1-F_\theta(x))^{2p-1}\frac{d^2F_\theta(x)}{d\theta^2}dx + 2p\int G^p(x)\,F_\theta^{p-1}(x)\frac{d^2F_\theta(x)}{d\theta^2}dx,
\end{aligned}
$$

$$
K = 2p^2\iint_{s<t}(r(s)+u(s))(r(t)+u(t))^T G(s)(1-G(t))\,ds\,dt
$$

where $r(x) = G^{p-1}(x)\frac{dF_\theta(x)}{d\theta}F_\theta^{p-1}(x)$ and $u(x) = (1-G(x))^{p-1}\frac{dF_\theta(x)}{d\theta}(1-F_\theta(x))^{p-1}$.

3.3.2 Robustness

The influence functions for the OH estimators of location and scale were derived by Öztürk and Hettmansperger in [23]. An outline of this derivation, for the location parameter only, is given in Appendix B.3 on page 241 and

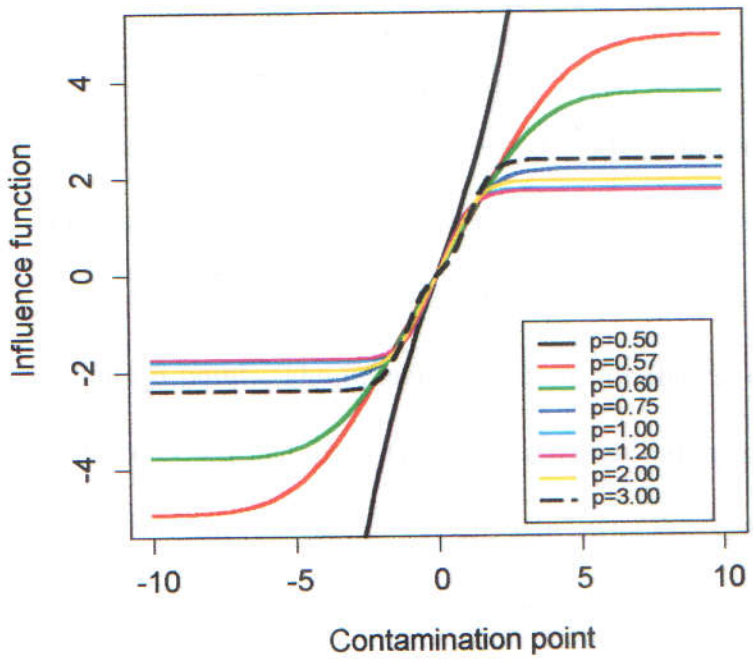

Figure 3.14: Influence function for the OH estimator when $f_\theta = N(\theta, 1)$.

shows that in this case the influence function is

$$IF(\xi) = \frac{\int \left[(1 - F_\theta(x))^{2p-2} + F_\theta^{2p-2}(x)\right] \left[F_\theta(x) - \Delta_\xi(x)\right] f_\theta(x)\, dx}{\int \left[(1 - F_\theta(x))^{2p-2} + F_\theta^{2p-2}(x)\right] f_\theta^2(x)\, dx} \tag{3.13}$$

Figure 3.14 shows the influence function of the location parameter with the $N(\theta, 1)$ model. Since this influence function is bounded only when $\int \left[(1 - F_\theta(x))^{2p-2} + F_\theta^{2p-2}(x)\right] dx$ is finite [23] this estimation procedure is not robust for any $p \leq 0.5$. For $p > 0.5$ the degree of robustness attained is determined by the magnitude of the lower and upper bounds of the influence

function. The absolute values of these bounds continue to decrease until p is a little larger than 1 and then slowly rises thereafter. Considering the influence function for p between 0.8 and 1.5 in steps of 0.1 indicates that maximum robustness is attained when $p = 1.2$. It is interesting to note that, in contrast to the minimum density power divergence, the optimal value of p for robustness does not appear to depend on the magnitude of the contamination point. The influence function for the criterion function does not redescend which means that an extremely large data point will have more influence on this estimation procedure than on either of the two density based methods. Although this suggests that this method may well be the least robust of the three methods considered in Chapter 3 its other redeeming features, such as single roots, make further study worthwhile nonetheless.

In Figure 3.15 (p.70) the data-based estimate of the criterion function, $d_{F_n}(\theta, p)$, is plotted for several values of p using a sample of $N(0,1)$ data with 10% contamination at 10. In each case the distance measure has a global minimum which is close to the target value 0 and furthermore, there are no local minima to confuse the numerical optimisation procedure which is needed to locate the minimiser. Although the shape of the distance measure differs for various values of p the location of the minimum does not

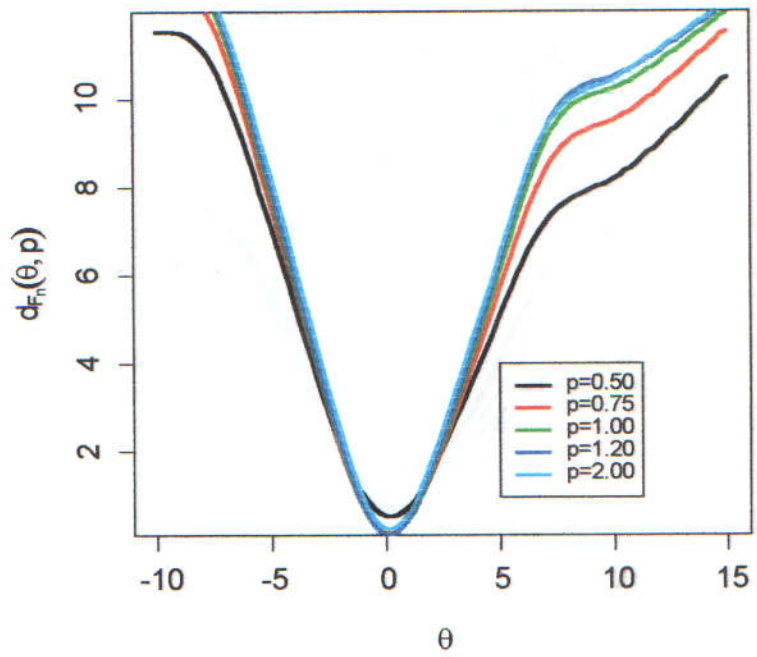

Figure 3.15: $d_{F_n}(\theta; p)$ for $N(0,1)$ data with 10% contamination at 10.

suggesting that, for these data at least, the choice of p will have little effect on the performance of this estimation procedure.

It is interesting, therefore, to investigate how the behaviour of the distance measure might change as the percentage of contaminated data points increases. In Figure 3.16 (p.71) $d_{F_n}(\theta, p)$ for $p = 1.2$ is plotted using $N(0,1)$ data with various percentages of contamination. As the degree of contamination increases a second minimum develops at the contamination point but the global minimum remains close to 0 until the percentage of con-

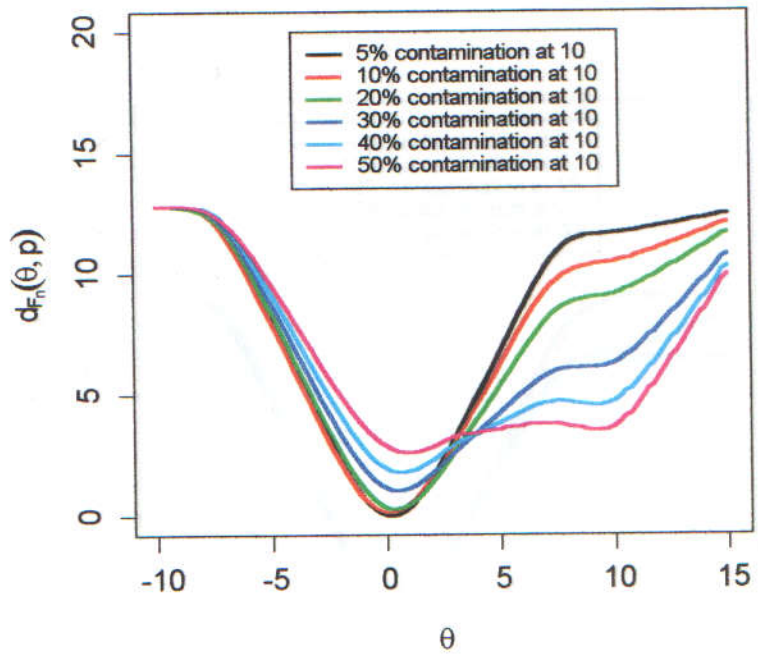

Figure 3.16: $d_{F_n}(\theta; p)$ with $p = 1.2$ for $N(0, 1)$ data with varying degrees of contamination at 10.

taminated data is approximately 50%. At this point the method breaks down, as Öztürk and Hettmansperger [23] suggest it should, and the global minimum shifts to the contamination point. Furthermore, Öztürk and Hettmansperger also noted that when the scale is known the breakdown point is independent of p. The behaviour of this distance measure as the magnitude of the contamination point increases is illustrated in Figure 3.17 (p.72). As indicated by the influence function (Figure 3.14, p.68), the global minimum remains in roughly the same place irrespective of whether the con-

tamination is at 3 or 30.

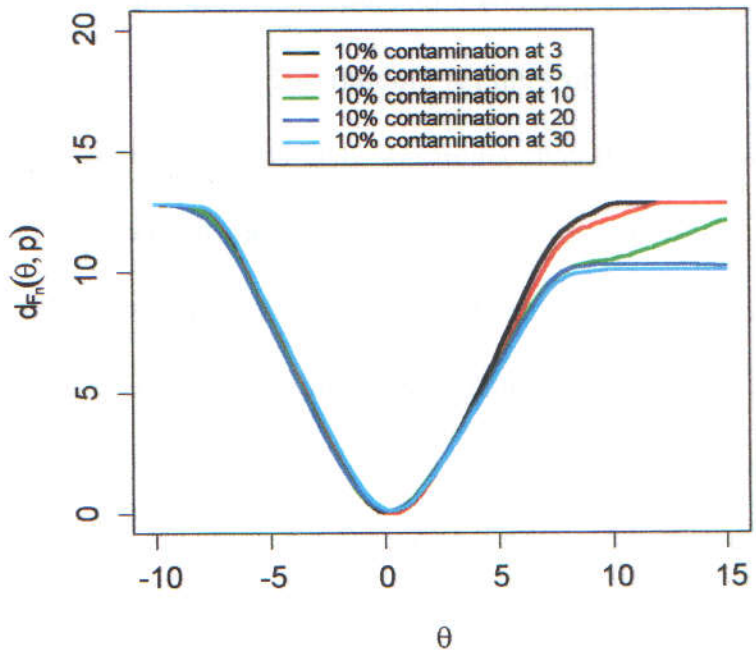

Figure 3.17: $d_{F_n}(\theta; p)$ with $p = 1.2$ for $N(0, 1)$ data with 10% contamination at various points.

3.3.3 Locating the estimates of θ_p

The estimating equation (3.12) cannot be solved explicitly and so, in common with the other two minimum distance estimators considered in Sections 3.1 (p.34) and 3.2 (p.47), a numerical procedure is needed to locate the estimates of θ_p. In this case, however, there are no multiple roots and so restrictions on the search area are not necessary.

3.3.4 Simulations

To further illustrate the behaviour of this estimation procedure, several simulated data sets were generated and used to obtain estimates of location. Each data set comprised 100 random samples of size 100 taken from one of the five distributions studied. Assuming that the dispersion parameter is known, the method was then applied to all the samples in each data set using a particular value of p to obtain 100 parameter estimates of location for the underlying distribution. The results obtained are summarised in Table 3.3 (p.75) which gives the average of the parameter estimates and their mean squared error for each distribution and a range of values of p. When there is no contamination in the data the performance of this method is very close to that of maximum likelihood for all values of p with the smallest mean squared error occurring when $p = 0.5$. For all the other distributions the mean squared error is minimised at $p \simeq 1$ as previously suggested by inspection of the influence function. This demonstrates that the criterion function can be used to provide estimators which are highly efficient when the data is not contaminated and robust when it is. By choosing $p = 0.5$ maximum efficiency will be attained but at the expense of robustness. Choosing $p > 0.5$ does the reverse by offering robust estimators with sub-optimal efficiency at the model. Without some reliable way of choosing p for a particular data set, however, these desirable theoretical

properties may not be attainable in practice.

Table 3.3: Simulation results for the OH estimator when $f_\theta = N(\theta, 1)$

Distribution	p	**Average** $\widehat{\theta}_p$	**Mean Squared Error** $(\widehat{\theta}_p)$
$\phi(x)$	0.25	-0.0027	0.0149
	0.50	0.0003	0.0083
	1.00	0.0019	0.0092
	2.00	0.0019	0.0089
	3.00	0.0019	0.0086
Maximum Likelihood	-	**0.0007**	**0.0084**
$0.9\phi(x) + 0.1\Delta_{10}(x)$	0.25	0.9343	0.9078
	0.50	0.3294	0.1254
	1.00	0.1953	0.0529
	2.00	0.2126	0.0600
	3.00	0.2577	0.0822
Maximum Likelihood	-	**0.9661**	**1.0123**
$0.8\phi(x) + 0.2\Delta_{10}(x)$	0.25	1.5003	2.3239
	0.50	0.6764	0.4968
	1.00	0.4480	0.2307
	2.00	0.4483	0.2648
	3.00	0.5799	0.3737
Maximum Likelihood	-	**1.9798**	**4.0941**
$0.9\phi(x) + 0.1\Delta_3(x)$	0.25	0.3541	0.1390
	0.50	0.2709	0.0871
	1.00	0.1871	0.0481
	2.00	0.2035	0.0549
	3.00	0.2459	0.0757
Maximum Likelihood	-	**0.2868**	**0.0967**
t_2	0.25	-0.0181	0.0625
	0.50	0.0001	0.0287
	1.00	0.0053	0.0246
	2.00	0.0049	0.0251
	3.00	0.0035	0.0272
Maximum Likelihood	-	**-0.0237**	**0.0930**

Chapter 4

A method for choosing α in the BHHJ estimator

4.1 Background

The family of density power divergences was introduced by Basu, Harris, Hjort and Jones [4] and is defined as

$$d_\alpha(g, f) = \int \left\{ f_\theta^{\alpha+1}(z) - \left(1 + \frac{1}{\alpha}\right) g(z) f_\theta^{\alpha}(z) + \frac{1}{\alpha} g^{\alpha+1}(z) \right\} dz \text{ for } \alpha > 0 \tag{4.1}$$

where f_θ is the model density, g is the true density.

The minimum density power divergence estimators, denoted θ_α, are the solutions to the following estimating equation

$$0 = \int u_{\theta_\alpha}(x) f_{\theta_\alpha}^{\alpha+1}(x)\, dx - \int u_{\theta_\alpha}(x) f_{\theta_\alpha}^{\alpha}(x)\, g(x)\, dx. \qquad (4.2)$$

The second term in the above is the expected value of the function $u_{\theta_\alpha}(x) f_{\theta_\alpha}^{\alpha}(x)$ and so can be replaced by the average of the function over the data to give a revised estimating equation as follows

$$0 = \int u_{\widehat{\theta}_\alpha}(x) f_{\widehat{\theta}_\alpha}^{\alpha+1}(x)\, dx - \frac{1}{n}\sum_{i=1}^{n} u_{\widehat{\theta}_\alpha}(X_i) f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) \qquad (4.3)$$

where X_i is the ith data point and $\widehat{\theta}_\alpha$ is the minimum density power divergence (BHHJ) estimate. (See Section 3.1, p.34, for further details.)

The tuning parameter, α, controls the robustness and efficiency properties of the resulting estimators. When $\alpha = 0$, $d_\alpha(g, f)$ reduces to the Kullback-Leibler divergence (which is equivalent to maximum likelihood) so the method produces estimators which are highly efficient but lack robustness. In contrast, when $\alpha = 1$, the $d_\alpha(g, f)$ gives the L_2 distance and the method therefore offers robustness at the expense of efficiency. Given that α equal to 0 or 1 leads to these two extremes, it seems plausible that by selecting α between 0 and 1 a trade-off between robustness and efficiency might be attained. Furthermore, given some appropriate measure of performance it may be possible to select α to optimise this trade-off for a particular set of data.

4.2 Estimation of the asymptotic mean squared error

As explained in Section 2.7 (p.18) the asymptotic mean squared error ($AMSE$) measures both robustness and efficiency and so by minimising an expression for the $AMSE$ of the BHHJ estimator, which is a function of α, it is hoped that the optimal value for α might be obtained. The $AMSE$ function of the BHHJ estimator is

$$As.E\left[\left(\hat{\theta}_\alpha - \theta_*\right)\left(\hat{\theta}_\alpha - \theta_*\right)^T\right] = \left(\theta_\alpha - \theta_*\right)\left(\theta_\alpha - \theta_*\right)^T + As.var\left(\hat{\theta}_\alpha\right).$$

where $\hat{\theta}_\alpha$ is the solution to equation (4.3), θ_α the solution to equation (4.2) and θ_* is the true parameter. This function is obtained by substituting θ_α for θ and $\hat{\theta}_\alpha$ for $\hat{\theta}$ in the expression for the multi-parameter $AMSE$ (equation D.1) of Appendix D (p.253).

The asymptotic variance of $\sqrt{n}\left(\widehat{\theta}_\alpha - \theta_\alpha\right)$ is $J^{-1}KJ^{-1}$ where

$$\begin{aligned} K = & \int f_{\theta_\alpha}^{2\alpha}(x) u_{\theta_\alpha}(x)\, u_{\theta_\alpha}^T(x)\, g(x) dx \\ & - \left[\int f_{\theta_\alpha}^{\alpha}(x) u_{\theta_\alpha}(x)\, g(x) dx\right]\left[\int f_{\theta_\alpha}(x) u_{\theta_\alpha}(x)\, g(x) dx\right]^T \end{aligned} \tag{4.4}$$

$$J = \int u_{\theta_\alpha}(x)\, u_{\theta_\alpha}^T(x)\, f_{\theta_\alpha}^{\alpha+1}(x)dx$$
$$+ \int \left(i_{\theta_\alpha}(x) - \alpha u_{\theta_\alpha}(x)\, u_{\theta_\alpha}^T(x)\right)\left(g(x) - f_{\theta_\alpha}(x)\right) f_{\theta_\alpha}^{\alpha}(x)dx \qquad (4.5)$$

where u_θ is the score function and $i_\theta(x) = \frac{-\partial\{u_\theta(x)\}}{\partial\theta}$ the information function of the model $f_\theta(x)$. (See Appendix A.1, p.209, for details of how this asymptotic variance is obtained.)

The $AMSE$ function of the BHHJ estimator is therefore

$$AMSE(\alpha) = (\theta_\alpha - \theta_*)(\theta_\alpha - \theta_*)^T + \frac{1}{n}J^{-1}KJ^{-1} \qquad (4.6)$$

where J and K are as defined in equations (4.4) and (4.5) respectively, θ_α is the solution to equation (4.2) and θ_* is the true parameter.

As explained in Appendix D (p.253), in the multi-parameter case the $AMSE$ is a matrix so the trace is used to provide a global measure of the $AMSE$ for minimisation. Thus when there are two unknown parameters to be estimated (θ and σ for example) the expression to be minimised is

$$AMSE(\hat{\theta}) \simeq As.var(\widehat{\theta}) + As.var(\widehat{\sigma}) + (\theta - \theta_*)^2 + (\sigma - \sigma_*)^2 \qquad (4.7)$$

The optimal choice of value for α is the minimiser of this function which is found using numerical methods.

It was my original intention to find a quadratic approximation to the $AMSE$ by expanding $f_\theta^\alpha(x)$ about $\alpha = 0$ using Taylor series. Then finding the optimal α would have been a matter of simple arithmetic. Unfortunately this approach was not possible because the terms of this Taylor series contain powers of $log\ f_\theta(x)$ so the series does not converge to $f_\theta^\alpha(x)$ across the whole of the region of interest $(0 \leq \alpha \leq 1)$.

Instead, by replacing g by the distribution of the data $\widehat{g}$, $\theta_\alpha$with $\widehat{\theta}_\alpha$ and θ_* with $\widehat{\theta}_*$ in equation 4.6, a data-based estimate of the $AMSE$ as a function of α can be obtained.

$$AMSE = (\widehat{\theta}_\alpha - \widehat{\theta}_*)(\widehat{\theta}_\alpha - \widehat{\theta}_*)^T + \frac{1}{n}\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1} \tag{4.8}$$

where

$$\widehat{K} = \frac{1}{n}\sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^{2\alpha}(X_i) u_{\widehat{\theta}_\alpha}(X_i)\, u_{\widehat{\theta}_\alpha}^T(X_i) - \frac{1}{n^2}\left[\sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) u_{\widehat{\theta}_\alpha}(X_i)\right]\left[\sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) u_{\widehat{\theta}_\alpha}(X_i)\right]^T$$

and $\widehat{J} = \int u_{\widehat{\theta}_\alpha}(x)\, u_{\widehat{\theta}_\alpha}^T(x)\, f_{\widehat{\theta}_\alpha}^{\alpha+1}(x)dx - \int \left(i_{\widehat{\theta}_\alpha}(x) - \alpha u_{\widehat{\theta}_\alpha}(x)\, u_{\widehat{\theta}_\alpha}^T(x)\right) f_{\widehat{\theta}_\alpha}^{\alpha+1}(x)dx$

$$+ \frac{1}{n}\sum_{i=1}^{n}\left[\left(i_{\widehat{\theta}_\alpha}(X_i) - \alpha u_{\widehat{\theta}_\alpha}(X_i)\, u_{\widehat{\theta}_\alpha}^T(X_i)\right) f_{\widehat{\theta}_\alpha}^{\alpha}(X_i)\right]$$

Since g is replaced by $\widehat{g}$, the integrals over g in K and J become summations over the data set in $\widehat{K}$ and $\widehat{J}$ respectively and the $\widehat{\theta}_\alpha$ are obtained by solving the estimating equation (4.3) with $g = \widehat{g}$ for a given α.

In the case where f_θ is the N(θ,1) distribution the $\widehat{K}$ and $\widehat{J}$ reduce to

$$\widehat{K} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\theta}_\alpha)^2 f_{\widehat{\theta}_\alpha}^{2\alpha}(X_i) - \frac{1}{n^2}\left[\sum_{i=1}^{n}(X_i - \widehat{\theta}_\alpha) f_{\widehat{\theta}_\alpha}^{\alpha}(X_i)\right]^2, \qquad (4.9)$$

$$\begin{aligned}\widehat{J} &= (\alpha+1)^{-\frac{3}{2}}(2\pi)^{-\frac{\alpha}{2}} + \frac{1}{n}\sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) - (\alpha+1)^{-\frac{1}{2}}(2\pi)^{-\frac{\alpha}{2}} \\ &\quad + \alpha(\alpha+1)^{-\frac{3}{2}}(2\pi)^{-\frac{\alpha}{2}} - \frac{\alpha}{n}\sum_{i=1}^{n}(X_i - \widehat{\theta}_\alpha)^2 f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} f_{\widehat{\theta}_\alpha}^{\alpha}(X_i) - \frac{\alpha}{n}\sum_{i=1}^{n}(X_i - \widehat{\theta}_\alpha)^2 f_{\widehat{\theta}_\alpha}^{\alpha}(X_i). \qquad (4.10)\end{aligned}$$

The problem of how to estimate θ_* in the bias part of the formula (Equation 4.8) is less easily solved because θ_* is the target parameter, the location of the true density g which is unknown. Furthermore, if a simple, reliable estimate of this target parameter was available there would be no need to utilise this alternative estimation method anyway. The argument therefore becomes circular; in order to obtain a good estimate of θ_* we must find the value of α which minimises the $AMSE$ but the $AMSE$ is itself a function of θ_*. Since θ_* can be viewed as a nuisance parameter in the α selection procedure, it could be replaced by any easily obtainable robust estimate. Therefore, putting $\theta_* = median$ seems to be an obvious first choice although several other robust estimates, such as the L_2 distance and $\widehat{\theta}_{0.2}$ (which is the

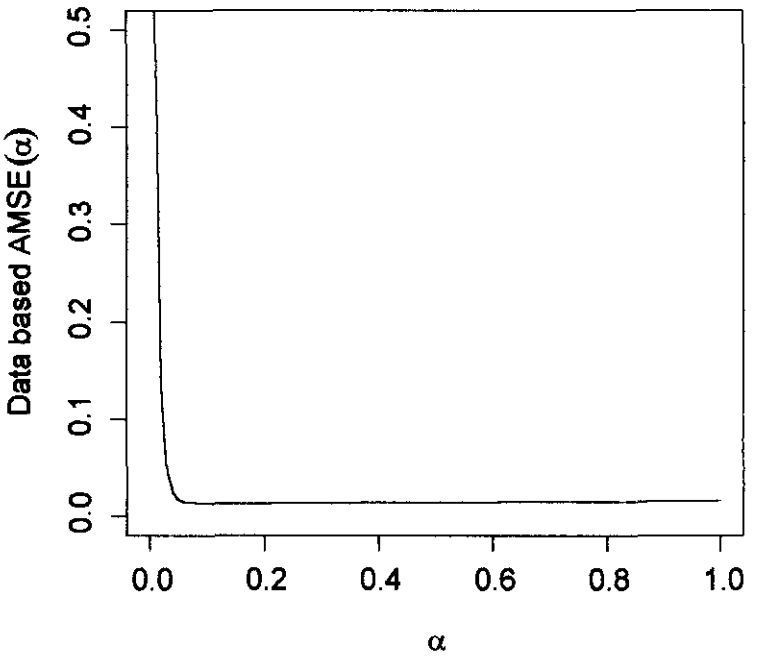

Figure 4.1: Data based estimate of the $AMSE$ of the BHHJ estimator using $f_\theta = N(\theta, 1)$ for $N(0, 1)$ data with 10% contamination at 10.

BHHJ estimate of θ when α=0.2), could also be considered. An example of this estimated $AMSE$ function is plotted for $N(0, 1)$ data with 10% contamination at 10 in Figure 4.1 (p.82). The curve falls very rapidly as α moves away from zero which suggests that the selection procedure will choose $\alpha > 0$ and gives hope that robustness will be attained.

4.3 Assessing the performance of this new method

For the new method to be considered a practical alternative to maximum likelihood estimation, in addition to offering robustness under departures from the model, the resulting estimators must also be highly efficient. These properties can be assessed by applying the method to simulated data for a particular family of models. Simulated data which contains some degree of contamination can be obtained by taking random samples from mixture densities, as described in Section 2.2 (p.8). Thus data which is predominantly $N(0,1)$ distributed but contains outliers is obtained by sampling $g_\varepsilon(x) = (1-\varepsilon)\phi(x) + \varepsilon\Delta_\xi(x)$ where there is $\varepsilon\%$ contamination at ξ and data which is very similar to $N(0,1)$ data but with heavy-tails is created by sampling the t distribution with k degrees of freedom where k= 2, 3 or 4. In both these cases the target density is the $N(0,1)$ distribution so when using the model $f_\theta = N(\theta, 1)$ one would hope that the resulting estimates of the location parameter will be close to the target value 0. Similarly, if dispersion is also unknown, using the model $f_\theta = N(\theta, \sigma^2)$ should lead to estimates of θ and σ which are close to their targets of 0 and 1 respectively. The efficiency of the new method can also be assessed in this way by applying the method to random samples from the $N(0,1)$ distribution

and comparing its performance to maximum likelihood estimation, which is known to be optimal when $f_\theta = g$.

Of course, these simulations are a rather simplistic representation of the estimation problem because in practice the distribution of the data is not the only unknown. One must also decide which family of models to use and, if this choice is inappropriate, the resulting estimators may be neither robust nor efficient. However, the problem of model suitability arises in many branches of statistics and has been widely studied elsewhere. Therefore, whilst acknowledging that in practice uncertainty regarding the model choice may well undermine confidence in the estimates obtained, I have not considered such possibilities here.

These simulations therefore focus on using the *Normal* family of models for imperfect, but predominantly $N(0,1)$ data, of the type commonly encountered in practice. To provide reassurance that the new method is suitable for use with other families of models, a reduced set of simulations were carried out using the *Gamma* $(4,\theta)$ family for uncontaminated and contaminated samples from the *Gamma* distribution. Further details of how these samples were obtained are given in Section 4.5.1 (p.92).

4.4 Theoretical asymptotic mean squared error

An alternative to applying this method to simulated data as a means of testing its performance, is to obtain theoretical results by substituting the probability density function from which the data was generated, g_ε, for g in equation 4.2. Solving this estimating equation for θ gives the theoretical θ_α, denoted θ_α^*, which are the values of θ_α which one would expect to get by applying the BHHJ method to a sample of data from g_ε. The variance part of the theoretical asymptotic mean squared error can be obtained by applying the same substitution to equations 4.4 and 4.5 and putting $\theta_\alpha = \theta_\alpha^*$. The bias term is calculated as the difference between the theoretical θ_α and the true parameter value θ_*.

The theoretical $AMSE$ function is therefore $B_* B_*^T + \frac{1}{n} J_*^{-1} K_* J_*^{-1}$ where $B_* = \theta_\alpha^* - \theta_*$,

$$K_* = \int f_\theta^{2\alpha}(x) u_\theta(x)\, u_\theta^T(x)\, g_\varepsilon(x) dx - \left[\int f_\theta^{\alpha}(x) u_\theta(x)\, g_\varepsilon(x) dx\right] \left[\int f_\theta^{\alpha}(x) u_\theta(x)\, g_\varepsilon(x) dx\right]^T$$

$$J_* = \int u_\theta(x)\, u_\theta^T(x)\, f_\theta^{\alpha+1}(x)dx$$
$$+ \int \left(i_\theta(x) - \alpha u_\theta(x)\, u_\theta^T(x)\right)\left(g_\varepsilon(x) - f_\theta(x)\right) f_\theta^\alpha(x)dx$$

and $\theta = \theta_\alpha^*$, f_θ is the model, u_θ is the score function and g_ε is the distribution of the data.

By making appropriate substitutions for f_θ and g_ε in B_*, K_* and J_* the optimal value of α for the simulated data sets can be obtained. These theoretical results can then be compared to the data-based ones and allow another aspect of the effectiveness of the new method to be assessed.

For contaminated standard normal data with $f_\theta = N(\theta, 1)$, the bias term has $\theta_* = 0$ and the expressions for J^* and K^* are

$$K_* = (1-\varepsilon)\int f_{\theta_\alpha^*}^{2\alpha}(x)(x-\theta_\alpha^*)^2\phi(x)dx + \varepsilon(\xi-\theta_\alpha^*)^2 f_{\theta_\alpha^*}^{2\alpha}(\xi)$$
$$- \left[(1-\varepsilon)\int f_{\theta_\alpha^*}^{\alpha}(x)(\theta_\alpha^*-x)\phi(x)dx + \varepsilon(\xi-\theta_\alpha^*) f_{\theta_\alpha^*}^{\alpha}(\xi)\right]^2$$

$$\text{and } J_* = (1-\varepsilon)\int f_{\theta_\alpha^*}^{\alpha}(x)\phi(x)dx + \varepsilon f_{\theta_\alpha^*}^{\alpha}(\xi)$$
$$-\alpha(1-\varepsilon)\int f_{\theta_\alpha^*}^{\alpha}(x)(x-\theta_\alpha^*)^2\phi(x)dx - \alpha\varepsilon f_{\theta_\alpha^*}^{\alpha}(\xi)(\xi-\theta_\alpha^*)^2$$

where ε is the degree of contamination and ξ is the contamination point.

When $f_\theta = N(\theta, 1)$ with $g_\varepsilon = t_k$ where k is the degrees of freedom, the bias term is as above but J_* and K_* are as follows

$$K_* = \int f_{\theta_\alpha^*}^{2\alpha}(x)(x-\theta_\alpha^*)^2 t_k(x)dx - \left[\int f_{\theta_\alpha^*}^{\alpha}(x)(x-\theta_\alpha^*)t_k(x)dx\right]^2$$

$$J_* = \int (x-\theta_\alpha^*)^2 f_{\theta_\alpha^*}^{\alpha+1}(x)dx + \int \left(1-\alpha(x-\theta_\alpha^*)^2\right)\left(t_k - f_{\theta_\alpha^*}(x)\right) f_{\theta_\alpha^*}^{\alpha}(x)dx$$

Similarly, using the model $f_\theta = Gamma(4,\theta)$ for data from any g_ε J^* and K^* as

$$K_* = \int f_{\theta_\alpha^*}^{2\alpha}(x) u_{\theta_\alpha^*}^2(x)\, g_\varepsilon(x)dx - \left[\int f_{\theta_\alpha^*}^{\alpha}(x) u_{\theta_\alpha^*}(x)\, g_\varepsilon(x)dx\right]^2$$

$$\text{and } J_* = \int f_{\theta_\alpha^*}^{\alpha+1}(x) u_{\theta_\alpha^*}^2(x)\, dx + \int \left(i_{\theta_\alpha^*}(x) - \alpha u_{\theta_\alpha^*}^2(x)\right)\left(g_\varepsilon - f_{\theta_\alpha^*}(x)\right) f_{\theta_\alpha^*}^{\alpha}(x)dx$$

where $u_\theta(x) = \frac{x}{\theta^2} - \frac{4}{\theta}$ and $i_\theta(x) = \frac{2x}{\theta^3} - \frac{4}{\theta^2}$. For contaminated $Gamma(4,1)$ data, the true parameter value in the bias term is $\theta_* = 1$.

The theoretical $AMSE$ functions (in the one and two parameter cases) for $N(0,1)$ data with 10% contamination at 10 are plotted in Figure 4.2 (p.88).

Both these curves falls very rapidly until reaching their respective minima and level out thereafter confirming that, for this data, choosing any $\alpha > 0.1$ would lead to better performance than maximum likelihood irrespective of whether the variance of the data is known.

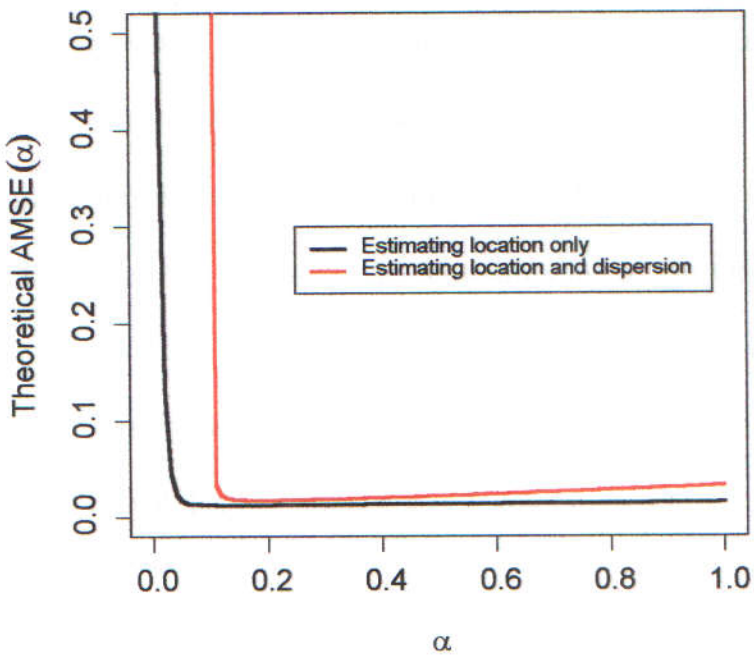

Figure 4.2: Theoretical $AMSE$ functions of the BHHJ estimator for $N(0,1)$ data with 10% contamination at 10.

The theoretical optimal α 's and the theoretical minimum $AMSE$ possible when using the model $f_\theta = N(\theta, 1)$ for data from a variety of distributions are shown in Table 4.1 (p.90) and were found numerically by evaluating

the $AMSE$ function over a grid of 100 evenly spaced values for α between 0 and 1. It it perhaps a little surprising that the theoretical optimum α for data with contamination at 5 should be greater than that for data contaminated at 10 but this can be explained by considering the influence function, Figure 3.1 (Chapter 3, p.39). For all $\alpha > 0$, the influence function redescends towards zero but the rate of that descent is determined by the value of α. The closer α is to 1 the more rapid the descent and the smaller a contamination point need be to be downweighted by the model. Thus for the procedure to be robust to contamination at 5 α needs to be larger than it does to ensure robustness to contamination at 10. Of course, when $\alpha = 0.5$, for example, the procedure is robust to contamination at both 5 and 10, but if the contamination in the sample is at 10 only, making α larger than it need be will lead to unnecessary losses in efficiency.

The theoretical optimal values of α for the same data when both location and dispersion are unknown are given in Table 4.2 (p.91). As one might expect, the optimal values of α are higher than in the one parameter case but generally follow the same pattern. However the optimal values of α for data from the t distribution increase quite significantly from around 0.3 to 1 which reflects the increased difficulty involved in obtaining a robust estimate of dispersion from data which has heavy-tails. Similarly, the min-

Table 4.1: Theoretical optimal values of α for the BHHJ estimator when $f_\theta = N(\theta, 1)$.

Distribution of data	**Theoretical Optimal α**	**Theoretical Minimum AMSE**
$\phi(x)$	0	0.010
$0.95\phi(x) + 0.05\Delta_{10}(x)$	0.11	0.011
$0.9\phi(x) + 0.1\Delta_{10}(x)$	0.13	0.011
$0.8\phi(x) + 0.2\Delta_{10}(x)$	0.14	0.013
$0.95\phi(x) + 0.05\Delta_5(x)$	0.29	0.012
$0.9\phi(x) + 0.1\Delta_5(x)$	0.36	0.013
$0.8\phi(x) + 0.2\Delta_5(x)$	0.44	0.015
$0.9\phi(x) + 0.1\Delta_{-10}(x)$	0.13	0.011
$0.9\phi(x) + 0.1\Delta_3(x)$	0.79	0.018
$0.8\phi(x) + 0.2\Delta_3(x)$	1.00	0.023
$0.9\phi(x) + 0.1\Delta_4(x)$	0.51	0.014
$0.8\phi(x) + 0.2\Delta_4(x)$	0.63	0.017
t_2	0.35	0.017
t_3	0.27	0.015
t_4	0.23	0.014

imum $AMSE$'s attainable are generally much larger and, in the case of 20% contamination at 3, increase considerably from 0.023 when dispersion is known to 0.416 when it is estimated from the data. This suggests that, even when the optimal value of α is used, the BHHJ method will not cope too well with this type of contamination.

Similar theoretically optimal results were obtained for $f_\theta = Gamma\,(4, \theta)$ with clean and heavy-tailed *Gamma* data and are shown in Table 4.3 (p.91). As with the previous examples, the magnitude of the optimal value for α is determined by the magnitude of the contamination in relation to the model. Thus contamination from the $Gamma(8, 1)$ distribution requires α

Table 4.2: Theoretical optimal values of α for the BHHJ estimator when $f_\theta = N(\theta, \sigma^2)$.

Distribution of data	**Theoretical Optimal α**	**Theoretical Minimum AMSE**
$\phi(x)$	0	0.015
$0.95\phi(x) + 0.05\Delta_{10}(x)$	0.17	0.017
$0.9\phi(x) + 0.1\Delta_{10}(x)$	0.18	0.018
$0.8\phi(x) + 0.2\Delta_{10}(x)$	0.20	0.022
$0.95\phi(x) + 0.05\Delta_5(x)$	0.45	0.021
$0.9\phi(x) + 0.1\Delta_5(x)$	0.54	0.026
$0.8\phi(x) + 0.2\Delta_5(x)$	0.64	0.048
$0.9\phi(x) + 0.1\Delta_{-10}(x)$	0.18	0.018
$0.9\phi(x) + 0.1\Delta_3(x)$	1.00	0.050
$0.8\phi(x) + 0.2\Delta_3(x)$	1.00	0.416
$0.9\phi(x) + 0.1\Delta_4(x)$	0.67	0.036
$0.8\phi(x) + 0.2\Delta_4(x)$	1.00	0.079
t_2	1.00	0.075
t_3	1.00	0.049
t_4	0.57	0.026

to be larger than when contamination is from $Gamma(16, 1)$.

Table 4.3: Theoretical optimal values of α when $f_\theta = Gamma(4, \theta)$.

Distribution of data	**Theoretical Optimal α**	**Theoretical Minimum AMSE**
$Gamma(4, 1)$	0	0.002
$0.9Gamma(4, 1) + 0.1Gamma(8, 1)$	0.76	0.009
$0.9Gamma(4, 1) + 0.1Gamma(16, 1)$	0.68	0.005

Comparison between the theoretical optimal α's and those obtained from simulated data will be discussed in Section 4.6.3 (p.105).

4.5 Simulations

As discussed in Section 4.3 (p.83) the performance of this new method was investigated by its application to several sets of simulated data. The data sets chosen are intended to represent a variety of practical situations in which robust methods are often recommended, namely where the data is heavy tailed or contains outliers. In addition to this, several simpler robust methods were also applied to the simulated data to assess whether any improvements in performance offered by this method are sufficient to justify the additional computation involved.

4.5.1 Generation of the simulated data sets

Each data set consists of 100 random samples of size 100 from the distribution g_ε. Where g_ε is a mixture density this selection process is carried out in two stages. First a random sample of size n is taken from the Binomial$(1, \varepsilon)$ distribution to give n indicator variables with approximately $100\varepsilon\%$ equal to 0 and $100(1-\varepsilon)\%$ equal to 1. The value 0 indicates a contamination point and 1 indicates a point from the true density. Thus the second stage involves randomly selecting a point from the true density for each indicator variable equal to 1. In considering $\varepsilon\%$ to be a random variable in this way the number of contamination points will vary from sample to sample. If

g_ε is a single density, t_2 for example, then the first part of this selection procedure is not necessary and random samples are obtained in one stage from the true density.

4.5.2 Estimation of θ_α

The BHHJ estimates, $\widehat{\theta}_\alpha$ were used to estimate θ_α and were obtained using the numerical optimisation procedure described in detail in Section 3.1.3 (p.45).

4.5.3 Estimation of θ_*

Several potential estimators for θ_* were tested in these simulations. These were the sample estimates of the median, $\widehat{\theta}_{0.2}$, $\widehat{\theta}_{0.3}$ and $\widehat{\theta}_1$ which is the L_2 estimate. In the two parameter case, the median absolute deviation (mad), $\widehat{\sigma}_{0.2}$, $\widehat{\sigma}_{0.3}$ and $\widehat{\sigma}_1$ were used to estimate scale. The scale parameter in the model $f_\theta = Gamma(4, \theta)$ is estimated by mad, $\widehat{\sigma}_{0.2}$, $\widehat{\sigma}_{0.3}$ and $\widehat{\sigma}_1$.

4.5.4 Minimising the $AMSE$ function

The α which minimises this cannot be found analytically therefore numerical methods must be used. However, care is needed in applying such methods

because the $AMSE$ function may change rapidly around the global minimum and there is the risk that the optimisation procedure will miss the region in which the global minimum lies altogether and find some local minimum instead. Therefore to avoid this problem, the $AMSE$ was evaluated across a grid of 100 points between 0 and 1 to find the minimiser directly. This is computationally intensive but ensures that the global minimum is found each time.

4.5.5 Other Robust Methods

The simplest robust alternative to this new method is to use the sample median and mean absolute deviation for location and dispersion respectively. The notion of using a fixed value of α for all data sets, rather than this data dependent approach, was also investigated with $\alpha = 0.2$, $\alpha = 0.3$ and $\alpha = 1$ (which is the L_2 distance estimator) being tried on the data sets.

4.6 Results and Discussion

To allow the performance of the various methods to be compared easily the mean squared errors (MSE) of the estimates obtained are summarised in Tables 4.4 to 4.8 on pages 96 to 100. Table 4.4 gives the MSE's when $f_\theta = N(\theta, 1)$ for contaminated and uncontaminated $N(0, 1)$ data and Ta-

bles 4.5 and 4.6 give the results for the location and dispersion parameters respectively when both are estimated (but a single value of α, optimised for joint performance, is used). Table 4.7 gives the overall MSE's for the two parameter case and are simply the sum of the individual MSE's from the previous two tables. The results obtained when the method, with $f_\theta = Gamma(4, \theta)$, was applied to contaminated *Gamma* data are given in Table 4.8.

Table 4.4: Simulation results: Mean squared errors of the BHHJ estimates of θ when $f_\theta = N(\theta, 1)$.

	Other Estimators					Minimising $AMSE$			
Distribution	*mean*	*median*	$\widehat{\theta}_{0.2}$	$\widehat{\theta}_{0.3}$	L_2	$\theta_* = \widehat{\theta}_{0.2}$	$\theta_* = \widehat{\theta}_{0.3}$	$\theta_* = L_2$	$\theta_* = median$
$\phi(x)$	0.008	0.013	0.009	0.009	0.014	0.009	0.009	0.009	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.331	0.028	0.013	0.014	0.019	0.013	0.013	0.014	0.014
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.012	0.012	0.019	0.011	0.012	0.012	0.015
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	0.015	0.016	0.023	0.015	0.015	0.015	0.036
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.084	0.023	0.014	0.012	0.017	0.013	0.013	0.013	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.022	0.015	0.019	0.015	0.015	0.015	0.018
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.038	0.017	0.019	0.015	0.014	0.014	0.028
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.111	0.043	0.012	0.012	0.017	0.011	0.011	0.012	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.045	0.030	0.016	0.060	0.043	0.043	0.022
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.226	0.155	0.028	0.316	0.226	0.034	0.316
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.201	0.052	0.041	0.025	0.022	0.021	0.020	0.020	0.021
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.672	0.123	0.128	0.045	0.021	0.036	0.018	0.018	0.036
t_2	0.093	0.027	0.024	0.024	0.031	0.025	0.024	0.025	0.024
t_3	0.023	0.021	0.014	0.015	0.023	0.015	0.015	0.015	0.015
t_4	0.016	0.018	0.012	0.013	0.017	0.013	0.013	0.014	0.014

Table 4.5: Simulation results: Mean squared errors of the BHHJ estimates of θ when $f_\theta = N(\theta, \sigma^2)$.

	Other Estimators					Minimising $AMSE$			
Distribution	*mean*	*median*	$\widehat{\theta}_{0.2}$	$\widehat{\theta}_{0.3}$	L_2	$\theta_* = \widehat{\theta}_{0.2}$	$\theta_* = \widehat{\theta}_{0.3}$	$\theta_* = L_2$	$\theta_* = median$
$\phi(x)$	0.008	0.013	0.009	0.010	0.014	0.010	0.010	0.010	0.013
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.331	0.028	0.013	0.014	0.019	0.012	0.012	0.012	0.017
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.011	0.012	0.018	0.012	0.012	0.012	0.016
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	1.473	0.065	0.020	0.018	0.018	0.018	0.019
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.084	0.023	0.019	0.013	0.016	0.011	0.011	0.012	0.015
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.099	0.039	0.017	0.014	0.014	0.015	0.023
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.614	0.436	0.016	0.016	0.016	0.016	0.037
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.111	0.043	0.011	0.012	0.016	0.014	0.015	0.017	0.015
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.046	0.065	0.051	0.060	0.050	0.015	0.023
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.313	0.283	0.094	0.308	0.281	0.094	0.121
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.201	0.052	0.098	0.062	0.021	0.098	0.062	0.021	0.029
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.672	0.123	0.486	0.396	0.029	0.486	0.396	0.029	0.069
t_2	0.093	0.027	0.024	0.024	0.027	0.021	0.021	0.021	0.025
t_3	0.023	0.021	0.014	0.014	0.020	0.015	0.015	0.015	0.018
t_4	0.016	0.018	0.012	0.012	0.016	0.015	0.015	0.015	0.015

Table 4.6: Simulation results: Mean squared errors of the BHHJ estimates of σ when $f_\theta = N(\theta, \sigma^2)$.

	Other Estimators					Minimising $AMSE$			
Distribution	s	mad	$\widehat{\sigma}_{0.2}$	$\widehat{\sigma}_{0.3}$	L_2	$\sigma_* = \widehat{\sigma}_{0.2}$	$\sigma_* = \widehat{\sigma}_{0.3}$	$\sigma_* = L_2$	$\sigma_* = mad$
$\phi(x)$	0.005	0.015	0.005	0.006	0.010	0.008	0.008	0.008	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	2.078	0.026	0.007	0.008	0.014	0.018	0.018	0.018	0.019
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	4.426	0.034	0.005	0.006	0.012	0.012	0.012	0.012	0.012
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	9.541	0.196	4.755	0.166	0.043	0.042	0.042	0.042	0.042
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.224	0.016	0.039	0.011	0.009	0.013	0.013	0.013	0.015
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.616	0.051	0.295	0.112	0.020	0.044	0.045	0.043	0.062
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	1.388	0.168	1.319	1.137	0.037	0.075	0.075	0.073	0.172
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	4.698	0.039	0.006	0.007	0.016	0.016	0.016	0.017	0.018
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.107	0.046	0.103	0.097	0.037	0.101	0.096	0.037	0.056
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.243	0.168	0.285	0.302	0.214	0.288	0.303	0.214	0.239
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.279	0.037	0.184	0.124	0.017	0.183	0.124	0.017	0.050
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.721	0.184	0.761	0.746	0.073	0.761	0.745	0.073	0.275
t_2	3.910	0.084	0.233	0.153	0.065	0.095	0.095	0.095	0.089
t_3	0.546	0.038	0.095	0.065	0.029	0.030	0.030	0.036	0.038
t_4	0.178	0.031	0.063	0.047	0.025	0.025	0.025	0.029	0.031

Table 4.7: Simulation results: Combined mean squared errors of the BHHJ estimates of θ and σ when $f_\theta = N(\theta, \sigma^2)$.

	Other Estimators					Minimising $AMSE$			
Distribution	*mean* s	*med* *mad*	$\widehat{\theta}_{0.2}$ $\widehat{\sigma}_{0.2}$	$\widehat{\theta}_{0.3}$ $\widehat{\sigma}_{0.3}$	L_2 L_2	$\theta_* = \widehat{\theta}_{0.2}$ $\sigma_* = \widehat{\sigma}_{0.2}$	$\theta_* = \widehat{\theta}_{0.3}$ $\sigma_* = \widehat{\sigma}_{0.3}$	$\theta_* = L_2$ $\sigma_* = L_2$	$\theta_* = med$ $\sigma_* = mad$
$\phi(x)$	0.013	0.028	0.014	0.016	0.024	0.018	0.018	0.018	0.022
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	2.409	0.054	0.020	0.022	0.033	0.030	0.030	0.030	0.036
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	5.438	0.077	0.016	0.018	0.030	0.024	0.024	0.024	0.028
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	13.64	0.322	6.228	0.201	0.063	0.060	0.060	0.060	0.061
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.308	0.039	0.059	0.024	0.025	0.024	0.024	0.025	0.030
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.921	0.105	0.394	0.151	0.037	0.058	0.059	0.058	0.086
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	2.368	0.275	1.933	1.573	0.053	0.091	0.091	0.089	0.209
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	5.809	0.082	0.017	0.019	0.032	0.030	0.031	0.034	0.033
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.204	0.079	0.149	0.162	0.088	0.161	0.146	0.052	0.079
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.615	0.297	0.598	0.585	0.308	0.611	0.584	0.308	0.360
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.480	0.089	0.282	0.186	0.038	0.281	0.186	0.038	0.079
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	1.393	0.307	1.247	1.142	0.102	1.247	1.142	0.102	0.344
t_2	4.003	0.111	0.257	0.177	0.092	0.116	0.116	0.116	0.114
t_3	0.569	0.059	0.109	0.079	0.049	0.045	0.045	0.051	0.056
t_4	0.194	0.049	0.075	0.059	0.041	0.040	0.040	0.044	0.046

Table 4.8: Simulation results: Mean squared errors of the BHHJ estimates of θ when $f_\theta = Gamma(4, \theta)$.

	Other Estimators					Minimising $AMSE$			
Distribution	mle	mad	$\widehat{\theta}_{0.2}$	$\widehat{\theta}_{0.3}$	L_2	$\theta_* = \widehat{\theta}_{0.2}$	$\theta_* = \widehat{\theta}_{0.3}$	$\theta_* = L_2$	$\theta_* = mad$
$Gamma(4,1)$	0.003	0.018	0.054	0.051	0.034	0.003	0.003	0.003	0.003
$0.9Gamma(4,1) + 0.1Gamma(8,1)$	0.016	0.023	0.013	0.012	0.011	0.014	0.014	0.013	0.014
$0.9Gamma(4,1) + 0.1Gamma(16,1)$	0.099	0.024	0.018	0.010	0.006	0.006	0.006	0.006	0.007

4.6.1 One parameter case

The lack of robustness of the maximum likelihood estimator is clearly demonstrated in column 2 of Table 4.4 (p.96). Whilst it produces estimators with the lowest MSE for $N(0,1)$ data, its performance is less good for data with symmetric contamination and very poor for data with asymmetric contamination. The median offers much smaller MSEs for both symmetric and asymmetric contamination but, although it is generally much more robust than the maximum likelihood estimator, its performance deteriorates quite rapidly as the percentage of contamination increases. There is little to choose between the other three estimators ($\widehat{\theta}_{0.2}, \widehat{\theta}_{0.3}$ and L_2). The performance of the L_2 estimator when $g = f_\theta = N(0,1)$ seems surprisingly good given that it is known to be very inefficient at the model. However, despite the mean squared errors being small in magnitude in this case, the ratio of the variance of the *mle* to that of the L_2 estimator is 0.6, confirming that the L_2 estimator does indeed perform badly when considered in terms of efficiency alone.

The results from the methods which minimise the $AMSE$ are generally better than those from the other robust methods. None of the methods proposed initially for estimating θ_* outperforms the others for every data set but on balance it seems that the L_2 distance might be preferred. This view is supported by Andrews *et al* [1] who suggest that when estimating a

nuisance parameter the goal of robustness is to be preferred over efficiency. It therefore seems reasonable that the L_2 should be used to initially estimate θ_*.

The results for contamination at 3 and 4 demonstrate that the BHHJ method, used either directly or to estimate θ_* in the $AMSE$, copes less well with contamination at points which are likely under the model than with outliers. The MSE when $\widehat{\theta}_{0.2}$ is used for data with 10% contamination at 10 is 0.012 whereas for 10% contamination at 3 it is 0.045. This counter-intuitive behaviour occurs because the BHHJ method weights data points by their probability under the model and thus contamination at 3 is more influential than contamination at 10. Nevertheless, in terms of MSE's, the performance of the BHHJ for such data is always at least as good as maximum likelihood and sometimes very much better.

The results obtained when applying the method to *Gamma* data are given in Table 4.8 (p.100). For $Gamma(4,1)$ data the methods which minimise the $AMSE$ perform much better than any of the simpler robust alternatives and equal that of maximum likelihood. It does not appear to matter how the true parameter θ_* is estimated because the MSE's are the same in each case. For the same data with 10% contamination from the $Gamma(8,1)$ distribution the results obtained by minimising the $AMSE$ are better than

those for maximum likelihood but the simpler BHHJ methods with α set equal to a predetermined value do offer the best performance for data of this type. When the contamination is from the $Gamma(16, 1)$ instead all the BHHJ based methods again perform better than either maximum likelihood or *mad* but in this case the methods which minimise the data-based estimate of the $AMSE$ lead to generally smaller MSE's than straightforward BHHJ. Since contamination from the $Gamma(8, 1)$ distribution is closer to the bulk of the uncontaminated data than $Gamma(16, 1)$ contamination it is a greater challenge to the BHHJ method and one would expect the MSE's for data of this type to be greater than in the latter case. The gains in efficiency offered by the $AMSE$ based methods greatly outweigh their shortcomings with respect to contamination which is not very unlikely under the model and offer the best overall performance. There is little to choose between the four $AMSE$ based methods but on balance it seems that the $\theta_* = L_2$ method should be preferred.

4.6.2 Two parameter case

The results for the location parameter, when estimating location and dispersion simultaneously, follow a similar pattern to those for the one parameter case. As one would expect, the MSE's are larger for the dispersion parameter than for location but the combined MSE's for all the robust methods

are considerably smaller than those for the mle's except, as one would expect, when the data is neither contaminated or heavy tailed. Once again the BHHJ method which minimises the $AMSE$ using the L_2 distance as an estimator for θ_* and σ_* performs well in most circumstances, as does the straightforward L_2 distance. There is little to choose between these two methods with the straightforward L_2 distance being less intense computationally but offering slightly less efficiency at the model than its $AMSE$ counterpart.

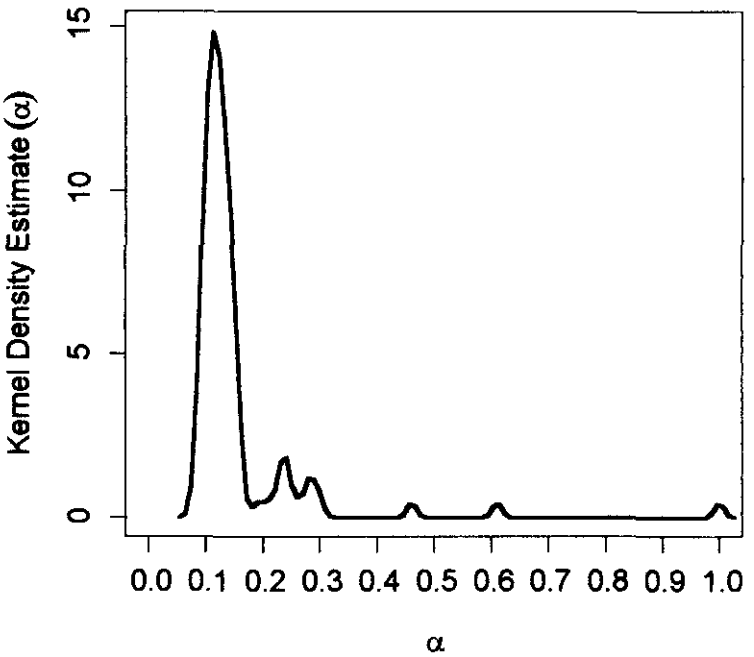

Figure 4.3: Kernel density estimate of the distribution of $\hat{\alpha}$ when $f_\theta = N(\theta, 1)$ for $N(0, 1)$ data with 10% contamination at 10.

4.6.3 Comparison to theoretical results

Tables 4.9 to 4.11 (p.113 - 115) give details of the data-based (with $\widehat{\theta}_* = L_2$) and theoretical optimal values of α and their associated MSE's. In the one parameter case, with $f_\theta = N(\theta, 1)$, the data-based estimate of α is very close to the theoretical optimal value so the method appears to work very well indeed. The data-based MSE's are just a little larger than the theoretical ones but they follow the same general pattern which clearly demonstrates the effectiveness of this new method. A kernel density estimate of the distribution of the data-based optimal α's in one particular case is plotted in Figure 4.3 (p.104) and shows how the data-based estimates of α cluster around 0.13 which is the theoretical optimal α for data of this type.

The method appears to work equally well for *Gamma* data with $f_\theta = Gamma(4, \theta)$ as indicated in Table 4.10. The $MSE's$ obtained from the data-based approach are very close to the theoretical optimal values and the estimates of α are generally quite good, except when there is contamination from the $Gamma(8, 1)$.

In the two parameter case (Table 4.11, p.115) the data-based estimates of α do not closely follow the theoretical optimal values and in many cases greatly over-estimate them. This is most apparent for the uncontaminated $N(0, 1)$ data where the data-based estimate is 0.69 as compared to the theoretical

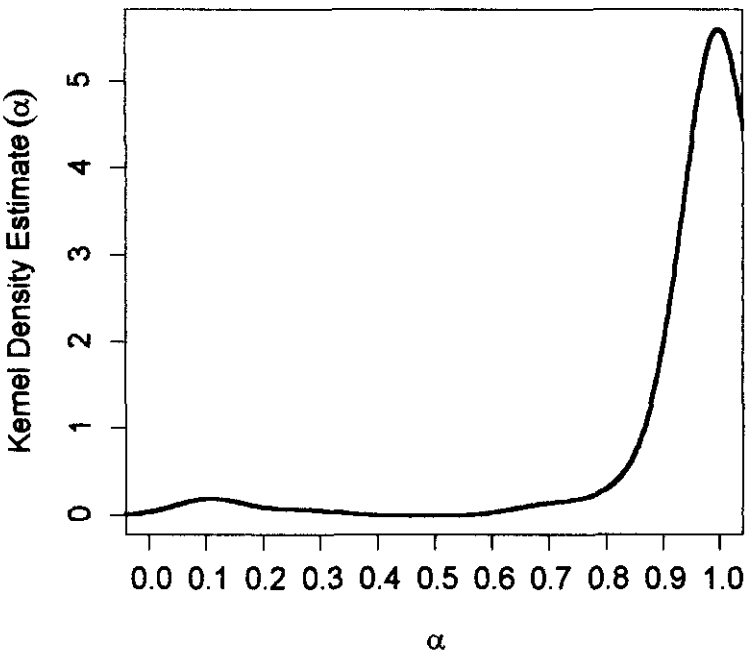

Figure 4.4: Kernel density estimate of the distribution of $\hat{\alpha}$ when $f_\alpha = N(\theta, \sigma^2)$ for $N(0,1)$ data with 10% contamination at 10.

optimal of 0, but despite this, the data-based MSE is very small (0.018) and only a little larger than the theoretical optimum (0.015). For contamination at 3 or 4 the new method performs very well indeed suggesting that α should be 1 (i.e. the L_2 estimator) which agrees with the theoretical optimum for 3 of the 4 such data sets. The method copes less well with heavy-tailed data, however, with the data-based estimates of α differing considerably from the optimal values but the resulting MSE's are small nonetheless. The kernel density estimate of the distribution of the data-based optimal

α's for $N(0,1)$ data with 10 % contamination at 10, plotted in Figure 4.4 (p.106), suggests that the estimates of α might in fact be even worse than the figures in Table 4.11 (p.115) imply. The average estimate of α is 0.62 but the majority of estimates are higher than this (clustered around $\alpha = 0.95$) and so considerably larger than the theoretical optimum, which is $\alpha = 0.17$.

The fact that the data-based MSE's are generally close to the optimal values even though the estimates of the optimal value of α are in some case very poor can be explained by considering the theoretical $AMSE$ functions (Figure 4.2, p.88). Both curves fall very rapidly as α moves away from 0 and rise relatively little after reaching their minima at $\alpha \simeq 0.1$ and $\alpha \simeq 0.2$ respectively. This makes estimating the optimal value of α very difficult because a slight change in the slope of the estimated curve may lead to relatively large shifts in the global minimum. However, this flatness also ensures that the data-based MSE's will be numerically close to the theoretical optima so long as the data-based estimate of α, however poor, is not too close to 0. It is not surprising, therefore, that in the two-parameter case the optimal BHHJ method does not perform consistently better than the straightforward L_2 distance. The potential for reducing the $AMSE$ by optimising the choice of α is quite small for data from many distributions and may well turn out to be far less than the error in estimating the $AMSE$ curve itself.

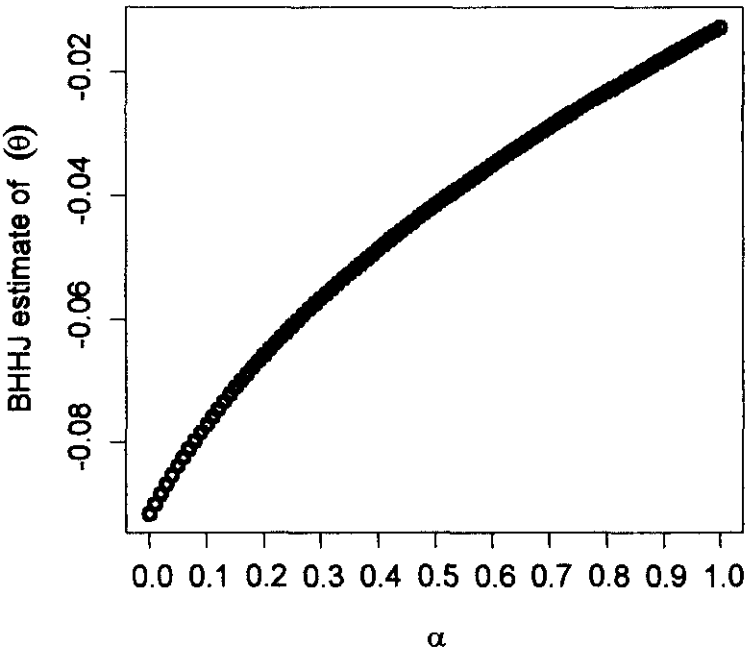

Figure 4.5: Sequence of BHHJ estimates for $N(0,1)$ data with $f_\theta = N(\theta, 1)$.

4.6.4 Potential for the development of diagnostic tools

An extremely useful aspect of the $BHHJ$ estimator is that by setting $\alpha = 0$, $0.01, 0.02, ..., 1$ in turn and repeatedly solving the estimating equation a sequence of estimates with increasing robustness can be obtained. Further research in this area is needed before the value of such plots can be fully assessed but it appears that, in some cases at least, the plot of this sequence of estimates against α may provide valuable information about the distribution of the data and might therefore increase one's confidence in the

estimates obtained. An example of the type of sequence of estimates one might obtain from a sample of $N(0,1)$ data is shown in Figure 4.5 (p.108). Scanning through similar plots, obtained from different realisations of data, indicates that this rough linearity (with the estimate of θ increasing with α) is typical of random samples from the $N(0,1)$ distribution and leads one to wonder whether these plots might have some diagnostic use. Another example which is typical of that obtained from samples which contain large outliers (i.e. contamination at 10 or t_2 data) is shown in Figure 4.6 (p.110). As α increases the estimates move rapidly away from the maximum likelihood estimator and converge towards a point elsewhere. This illustrates the effect of downweighting the contamination points and shows that once α is beyond 0.5 further increase in its magnitude has little effect.

The sequences of estimates can also be used to identify situations in which the method has broken down, as in Figure 4.7 (p.111) which shows the sequence obtained from a sample of $N(0,1)$ data with 45% contamination at 10. Initially, as α increases the estimates fall rapidly, as in the previous example, but when $\alpha \simeq 0.7$ breakdown occurs and the sequence of estimates jumps from around 0 to the contamination point.

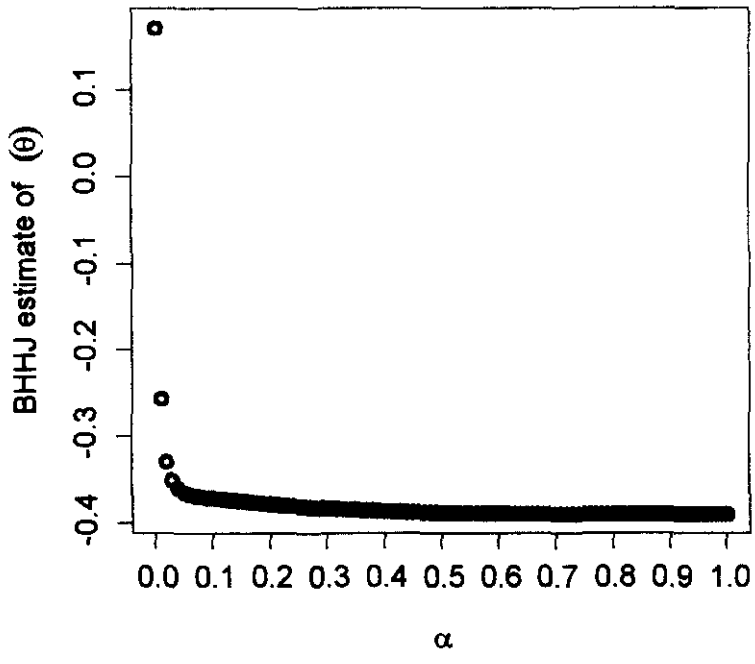

Figure 4.6: Sequence of BHHJ estimates for $N(0,1)$ data with 10% contamination at 10 and $f_\theta = N(\theta, 1)$.

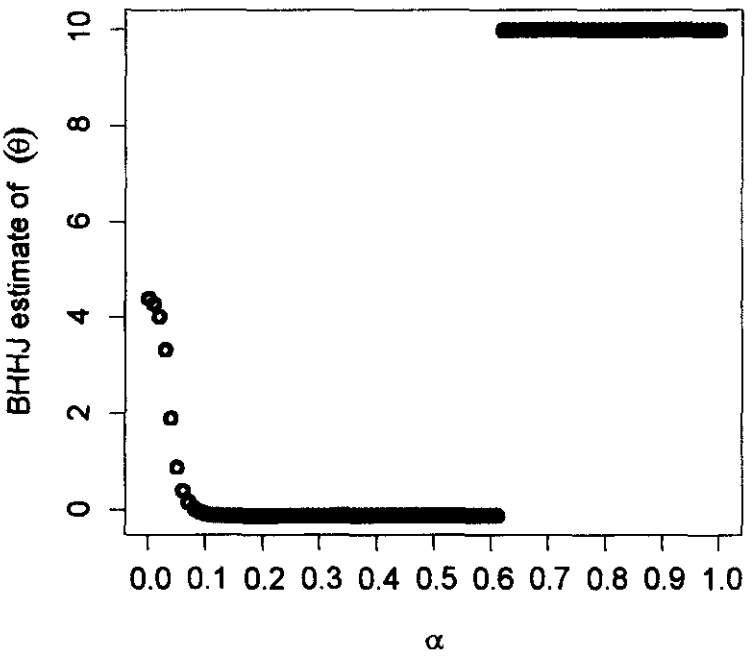

Figure 4.7: Sequence of BHHJ estimates for $N(0,1)$ data with 45% contamination at 10 and $f_\theta = N(\theta, 1)$.

Table 4.9: Theoretical and data-based mean squared errors of the BHHJ estimator when $f_\theta = N(\theta, 1)$.

Distribution	**Average Optimal α**		**Mean Squared Error**	
	Theoretical Results	Data-Based Results	Theoretical Results	Data-based Results
$\phi(x)$	0	0.06	0.010	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.11	0.16	0.011	0.014
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	0.13	0.15	0.011	0.012
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	0.14	0.19	0.013	0.015
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.29	0.27	0.012	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.36	0.32	0.013	0.015
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.44	0.45	0.015	0.014
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	0.13	0.17	0.011	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.79	0.51	0.018	0.043
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	1.00	0.86	0.023	0.034
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.51	0.44	0.014	0.020
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.63	0.64	0.017	0.018
t_2	0.35	0.44	0.017	0.025
t_3	0.27	0.40	0.015	0.015
t_4	0.23	0.30	0.014	0.014

Table 4.10: Theoretical and data-based mean squared errors of the BHHJ estimator when $f_\theta = Gamma(4, \theta)$.

Distribution	**Average Optimal α**		**Mean Squared Error**	
	Theoretical Results	Data-Based Results	Theoretical Results	Data-based Results
$Gamma(4,1)$	0	0.07	0.002	0.003
$0.9Gamma(4,1) + 0.1Gamma(8,1)$	0.76	0.11	0.009	0.013
$0.9Gamma(4,1) + 0.1Gamma(16,1)$	0.68	0.63	0.005	0.006

Table 4.11: Theoretical and data-based mean squared errors of the BHHJ estimator when $f_\theta = N(\theta, \sigma^2)$.

Distribution	**Average Optimal α**		**Mean Squared Error**	
	Theoretical Results	Data-based Results	Theoretical Results	Data-based Results
$\phi(x)$	0	0.69	0.015	0.018
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.17	0.62	0.017	0.030
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	0.18	0.69	0.018	0.024
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	0.20	0.91	0.022	0.060
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.45	0.58	0.021	0.025
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.54	0.52	0.026	0.058
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.64	0.64	0.048	0.089
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	0.18	0.72	0.018	0.034
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	1.00	1.00	0.050	0.052
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	1.00	1.00	0.416	0.308
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.67	1.00	0.036	0.038
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	1.00	1.00	0.079	0.102
t_2	1.00	0.73	0.075	0.116
t_3	1.00	0.64	0.049	0.051
t_4	0.57	0.75	0.026	0.044

Chapter 5

A method for choosing the bandwidth in Hellinger distance estimators

5.1 Background

The Hellinger distance (HD) is defined as follows

$$\begin{aligned} HD &= \int \left(f_\theta^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x) \right)^2 dx \\ &= 2 - 2 \int f_\theta^{\frac{1}{2}}(x) g^{\frac{1}{2}}(x) dx \end{aligned} \tag{5.1}$$

where f_θ is the model density and g is the true density.

The HD estimate of θ, denoted $\widehat{\theta}$, is the value of θ which minimises this distance and is obtained by differentiating equation (5.1) with respect to θ and setting it equal to 0. This leads to the estimating equation

$$H(\theta) = 0 = -\int g^{\frac{1}{2}}(x) f_\theta^{-\frac{1}{2}}(x) \frac{df_\theta(x)}{d\theta} dx \tag{5.2}$$

The true density distribution of the data, g, is not known therefore it must be estimated from the data. Kernel density estimation is the most widely used method for such problems and shall be used here.

5.1.1 Kernel density estimation

The kernel density estimate of the true density g is obtained via the following formula.

$$\widehat{g}_n(x) = \frac{1}{n} \sum_{i=1}^{n} N(X_i, h^2) \tag{5.3}$$

where n is the sample size, X_i is the ith data point, h is the bandwidth and $N(\theta, \sigma^2)$ denotes the normal density. Thus a series of n normal curves are placed around the data points, each centred at that point with the location parameter $\theta = X_i$ and variance $\sigma^2 = h^2$. The kernel density estimate at x is

taken to be the average value of the curves at that point. Thus the choice of bandwidth controls the extent to which these curves overlap and therefore, in a limited way, robustness. There is no necessity to use a normal kernel for this density estimation however and there are many alternatives, such as the rectangular and Epanechnikov kernels. In practice the choice of kernel has little impact on the estimation procedure and it is the bandwidth which should cause most concern [32],[33],[37].

Estimating the true density g with a kernel density estimate, denoted by $\widehat{g}_n$, therefore introduces a new problem; that of bandwidth selection. The HD estimates depend on the bandwidth so the HD estimator is denoted $\widehat{\theta}_h$ and the estimating equation becomes

$$\widehat{H}\left(\widehat{\theta}_h\right) = -\int \widehat{g}_n^{\frac{1}{2}}(x) f_{\widehat{\theta}_h}^{-\frac{1}{2}}(x) \left.\frac{df_\theta(x)}{d\widehat{\theta}}\right|_{\theta=\widehat{\theta}_h} dx = 0 \tag{5.4}$$

5.2 Estimation of the asymptotic mean squared error

Since $\widehat{\theta}_h$ is generally a biased estimator of the target parameter θ_* the asymptotic mean squared error ($AMSE$) again seems to be an appropriate choice for assessing the performance of this method. The aim is therefore to obtain

an expression for the $AMSE$ in terms of h which can then be optimised. The $AMSE$ function for the HD estimator is obtained by substituting $\hat{\theta}_h$ for $\hat{\theta}$ in equation (D.1) of Appendix D (p.253) and is

$$As.E\left[\left(\hat{\theta}_h-\theta_*\right)\left(\hat{\theta}_h-\theta_*\right)^T\right]=\left(\theta-\theta_*\right)\left(\theta-\theta_*\right)^T+As.var\left(\hat{\theta}_h\right)$$

where $\hat{\theta}_h$ is the solution to equation (5.4), θ the solution to equation (5.2) and θ_* is the true parameter.

The leading term in the asymptotic variance of $\sqrt{n}\left(\widehat{\theta}_h-\theta\right)$ is $J^{-1}KJ^{-1}$ where

$$J=\int\left[f_\theta^{-\frac{3}{2}}(x)\left[\frac{df_\theta(x)}{d\theta}\right]\left[\frac{df_\theta(x)}{d\theta}\right]^T-2f_\theta^{-\frac{1}{2}}(x)\frac{d^2f(x)}{d\theta^2}\right]g^{\frac{1}{2}}(x)\,dx \tag{5.5}$$

$$\text{and } K=\int\left[\frac{df_\theta(x)}{d\theta}\right]\left[\frac{df_{\theta(x)}}{d\theta}\right]^T f_\theta^{-1}(x)\,dx-\psi\psi^T \tag{5.6}$$

where $\psi=\int\left(\frac{df_\theta(x)}{d\theta}\right)f_\theta^{-\frac{1}{2}}(x)\,g^{\frac{1}{2}(x)}dx$. Details of how this expression is derived are given in Appendix A.2 (p.213).

The first-order $AMSE$ function for the HD estimator is therefore

$$AMSE=(\theta-\theta_*)(\theta-\theta_*)^T+\frac{1}{n}J^{-1}KJ^{-1} \tag{5.7}$$

where J and K are as defined above, θ is the solution to equation (5.2) and θ_* is the true parameter.

Since this expression does not depend on h it is, as it stands, of no use with regard to finding the optimal bandwidth. To resolve this problem a higher order approximation to the asymptotic variance was obtained instead by including further terms from the Taylor series expansion of the estimating equation (5.4), as detailed in Appendix A.2 (p.213). Using the same notation as above but taking the next term in the approximation leads to the following revised formula for K (J stays the same) as follows

$$\begin{aligned} K &= \int s(x)\, s^T(x)\, g(x)\, dx - \int\int s(x)\, s^T(y)\, g(x)\, g(y)\, dy\, dx \\ &+ \frac{h^2 K_2}{n}\left[\int g(x)\, s(x)\, [s''(x)]^T - \int\int s(x)\, g(x)\, s^T(y)\, g''(y)\, dy\, dx\right] \\ &- \frac{h^2 K_2}{2n}\left[\int s(x)\, s^T(x)\, g''(x)\, dx - \int\int s(x)\, s^T(y)\, g(x)\, g''(y)\, dy\, dx\right] \\ &- \frac{K_2}{2n^2 h}\left[\int s(x)\, s^T(x)\, dx - \int\int s(x)\, s^T(y)\, g(x)\, dy\, dx\right] \end{aligned}$$

where $s = \frac{df_\theta}{d\theta} f_\theta^{-\frac{1}{2}} g^{-\frac{1}{2}}$, $K_2 = \int K(u)\, u^2 du$, $s'' = \frac{d^2 s}{dx^2}$ and $g'' = \frac{d^2 g}{dx^2}$.

There were two problems with this formula which led to this approach being abandoned. The first arises because for the normal model $\int ss^T - \int s^T g \int s = \infty$ which would make this estimate of the variance ∞ and the second because the presence of h in the denominator of the multiplier $\frac{K_2}{2n^2 h}$ makes it unclear which of these terms will dominate. Furthermore, it could be that the multiplier $\frac{1}{n^2 h}$ arises in the higher order terms which have not been considered here, in which case, this expression for K would be unlikely to be useful. The results of simulation studies (my

own and others) do not support the assertion that the HD estimates are infinitely variable so, assuming that the above formula is algebraically correct, the problem must lie in its convergence. The convergence of the first order approximation to the asymptotic variance (i.e. $\frac{1}{4}J^{-1}KJ^{-1}$ where $K = \int s(x)\, s^T(x)\, g(x)\, dx - \int\int s(x)\, s^T(y)\, g(x)\, g(y)\, dy\, dx$) was rigorously proven by Beran [6] so it is certainly not the case that the expansion does not converge at all. It seems likely therefore that this approximation simply does not converge rapidly enough. A similar problem is encountered when using a Taylor series approximation to obtain the influence function for HD estimators. The Hellinger distance method shares the same influence function as maximum likelihood (i.e. unbounded) and yet is highly robust. Lindsay [19] explained this seemingly impossible behaviour by demonstrating that the influence function, which is a first order and very satisfactory approximation to bias for many estimators, is very poor indeed in the case of the Hellinger distance. He concluded that because the Hellinger distance method is first order equivalent to maximum likelihood, it is the second (and possibly later) terms in the approximation to bias which explain its robustness. Therefore a possible solution to the first problem (that of $\int s^2 = \infty$) might be to consider further terms in the approximation in the hope that one of them might cancel with the $\int s^2$ but the differentiation required is unwieldy to say the least and since the inverse h term might remain further

work in this direction was not attempted.

Another possible solution is to put boundaries on the range of integration so that $\int s^2 < \infty$ but the estimate of the asymptotic variance turns out to be negative because this term still swamps the others.

Therefore, since only the first term in the expansion for the asymptotic variance appeared to be reliable there was no alternative but to use the first order approximation to the $AMSE$ (equation 5.7) as the basis for optimising performance. Recall that this expression is not a function of h and so, as it stands, cannot be used to determine the optimal bandwidth. However, it is important to note that this independence is asymptotic and that in the finite case the performance of this method most certainly does depend on the bandwidth. Consider a random sample from the $N(0,1)$ distribution with 10% contamination at 10. Using a small bandwidth (say $h = 0.1$) will lead to a kernel density estimate which is bimodal with one peak at approximately 0 and another at 10. The peak at 10 is downweighted by the model and so robust parameter estimates will be obtained. By comparison using $h = 4$ will give a unimodal density estimate with a very heavy right tail. This right tail will also be downweighted by the model but to a much lesser degree and so the method will be less robust than in the previous case. Although much of the robustness and efficiency of HD estimators

is inherent in the method these properties are affected by the choice of bandwidth h so this general approach to optimising the performance of this method is reasonable, the challenge is therefore to modify the expression for the $AMSE$ to more closely reflect the method's behaviour in the finite case.

Replacing the unknown parameters in the $AMSE$ function by suitable estimates from data has the desired effect because the most obvious estimator for θ is $\widehat{\theta}_h$. The data-based estimate of the $AMSE$ is now a function of h which can be minimised to find the optimal bandwidth. This is much less *ad hoc* than it seems because (5.2) is replaced by $H_h(\theta) = 0 = -\int E(\widehat{g}_n(x))^{\frac{1}{2}} f_\theta^{-\frac{1}{2}}(x) \frac{df_\theta(x)}{d\theta} dx$ which is solved by θ_h (contrasted with θ). We are therefore justified in thinking of the HD estimator as θ_h which is estimated by $\widehat{\theta}_h$. The formula for the $AMSE$ also contains θ_*, the unknown true location, which must also be estimated from the data. The fact that it's difficult to obtain a reliable estimate of θ_* from data is the very reason that robust methods were developed and so the method stalls because in order to get a robust estimate of θ_* you need to know what value to place on h and you can't decide what h to use unless you know θ_*. Progress in solving this circular problem can again be made by treating θ_* as a nuisance parameter in the $AMSE$ function and replacing it with the most robust estimate which can be obtained simply. In this case, the median of the data

seems an appropriate choice for location and the median absolute deviation (MAD) for dispersion but there are many alternatives and so $\widehat{\theta}_*$ will be used to denote any such estimate of θ_*. The data-based estimate of the $AMSE$ is thus

$$\widehat{AMSE}\,(h) = \left(\widehat{\theta}_h - \widehat{\theta}_*\right)\left(\widehat{\theta}_h - \widehat{\theta}_*\right)^T + \frac{1}{n}\widehat{J}_h^{-1}\widehat{K}_h\widehat{J}_h^{-1} \tag{5.8}$$

where

$$\widehat{J}_h = \int \left[f_\theta^{-\frac{3}{2}} \left[\frac{df_\theta}{d\theta}\right]\left[\frac{df_\theta}{d\theta}\right]^T - 2f_\theta^{-\frac{1}{2}}\frac{d^2 f}{d\theta^2}\right] \widehat{g}_n{}^{\frac{1}{2}}\, dx \Bigg|_{\theta=\widehat{\theta}_h} \tag{5.9}$$

and

$$\widehat{K}_h = \int \left[\frac{df_\theta}{d\theta}\right]\left[\frac{df_\theta}{d\theta}\right]^T f_\theta^{-1}\, dx - \widehat{\psi}\widehat{\psi}^T \tag{5.10}$$

where $\widehat{\psi} = \int \left[\frac{df_\theta}{d\theta}\right] f_\theta^{-\frac{1}{2}} \widehat{g}_n^{\frac{1}{2}}\, dx$.

A graph of the data-based estimate of the $AMSE$ for $N(0,1)$ data with 10% contamination at 10 is given in Figure 5.1 (p.125). As the bandwidth decreases from 1.2 towards 0.1 the $AMSE$ curve falls slowly towards its minimum at $h = 0.1$. The curve rises very sharply as the bandwidth gets very small ($h < 0.01$) and the variance term tends to infinity.

Once again, the trace of the $AMSE$ matrix will be used to provide a global measure for minimisation. Thus when there are two unknown parameters

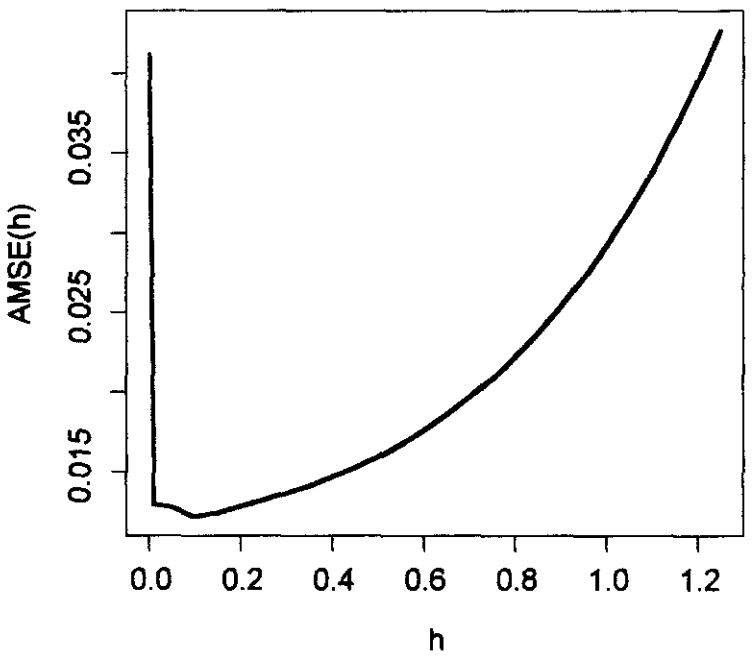

Figure 5.1: Data-based estimate of the $AMSE$ of the HD estimator using $f_\theta = N(\theta, 1)$ for $N(0, 1)$ data with 10% contamination at 10.

(θ and σ for example) the expression to be minimised is

$$AMSE(\hat{\theta}_p; \widehat{\sigma}_p) \simeq As.var(\widehat{\theta}_p) + As.var(\widehat{\sigma}_p) + (\theta_p - \theta_*)^2 + (\sigma_p - \sigma_*)^2$$

As in the previous chapter, the minimiser of this function is found using numerical methods.

5.3 Assessing the performance of the new method

The performance of the HD estimate in a variety of situations can be investigated by applying the method to simulated data. The data sets used are described in detail in Section 4.3 (p.83) of the previous chapter and consist of 100 random samples of size 100 from a perturbation of the true density g. Thus the simulated data has a variety of attributes, such as heavy-tails or outliers, and the performance of the method is fully assessed. As in the previous chapter, the correct family of models is used in each case; the *Normal* family for the predominantly $N(0, 1)$ data and the *Gamma* family for contaminated and uncontaminated *Gamma* data.

5.4 Theoretical Asymptotic Mean squared Error

The theoretical asymptotic mean squared error ($TAMSE$) for data of a particular type can be calculated by replacing g with g_ε in equations 5.2, 5.5 and 5.6. The theoretical HD estimate is thus the solution to this revised estimating equation and θ_* is the true parameter. Since the bandwidth only comes in at the estimation stage, the $TAMSE$ does not depend on h and is simply a point estimate. Obviously, this gives no information regarding the optimal value for h but is nevertheless a useful guide for assessing the performance of the method. Table 5.1 (p.128) gives the $TAMSE$'s for data from a variety of distributions when either location only or both location and dispersion are unknown. These figures confirm that this method will be generally robust and efficient in the one parameter case with the $AMSE$'s being very small for all types of data. In the two parameter case the $AMSE$'s are, as one would expect, larger than in the one parameter case and the method performs very well except when there is contamination at a point which is plausible under the model (at 3 or 4 for example) or when the data is heavy tailed.

The minimum $TAMSE$'s for the Gamma model are given in Table 5.2 (p.128) and show that the HD method should cope extremely well with all

Table 5.1: Theoretical minimum $AMSE$ of the HD estimator for the normal family of models.

Distribution of data	**Theoretical AMSE** $f_\theta = N(\theta, 1)$	**Theoretical AMSE** $f_\theta = N(\theta, \sigma^2)$
$\phi(x)$	0.010	0.015
$0.95\phi(x) + 0.05\Delta_{10}(x)$	0.011	0.016
$0.9\phi(x) + 0.1\Delta_{10}(x)$	0.011	0.017
$0.8\phi(x) + 0.2\Delta_{10}(x)$	0.013	0.019
$0.95\phi(x) + 0.05\Delta_5(x)$	0.011	0.017
$0.9\phi(x) + 0.1\Delta_5(x)$	0.011	0.018
$0.8\phi(x) + 0.2\Delta_5(x)$	0.013	0.022
$0.9\phi(x) + 0.1\Delta_{-10}(x)$	0.011	0.017
$0.9\phi(x) + 0.1\Delta_3(x)$	0.021	0.062
$0.8\phi(x) + 0.2\Delta_3(x)$	0.039	0.157
$0.9\phi(x) + 0.1\Delta_4(x)$	0.013	0.030
$0.8\phi(x) + 0.2\Delta_4(x)$	0.015	0.066
t_2	0.018	0.257
t_3	0.016	0.128
t_4	0.015	0.081

three data sets. Comparison between the theoretical minimum $AMSE$'s attainable and those obtained by applying the method to simulated data will be discussed in Section 5.6.3 (p.138).

Table 5.2: Theoretical minimum $AMSE$ of the HD estimator for the $Gamma(4, \theta)$ family of models.

Distribution of data	**Theoretical AMSE** $f_\theta = Gamma(4, \theta)$
$Gamma(4, 1)$	0.006
$0.9Gamma(4, 1) + 0.1Gamma(4, 8)$	0.009
$0.9Gamma(4, 1) + 0.1Gamma(4, 16)$	0.007

5.5 Simulations

As previously explained, the performance of this method was assessed by its application to several sets of simulated data. The data sets chosen were intended to represent practical situations in which robust methods are often recommended, namely when the data contains outliers or is heavy tailed. Simple random samples were also taken from the $N(0,1)$ distribution to enable the efficiency of the method to be evaluated. For a robust method to be worthy of further consideration it should not only perform better than maximum likelihood but better than its simpler robust rivals as well. Therefore the location and dispersion parameters were also estimated using the median and median absolute dispersion respectively so that comparisons could be made. Several bandwidth selection procedures (Sheather-Jones [31], Silverman [32], for example) are already available and widely used in other contexts so, in addition to the new method of minimising the $AMSE$ function, Hellinger distance estimates were also obtained using the Sheather-Jones bandwidth.

5.5.1 Generation of the simulated data sets

The new method was applied to the same simulated data as the BHHJ method in the previous chapter. Full details of how this data was generated

is therefore given in Chapter 4, Section 4.5.1 (p.92).

5.5.2 Estimation of θ_h

The HD estimates, $\widehat{\theta}_h$, for a range of bandwidths were obtained by minimising a data-based estimate of the distance measure $HD = 2 - 2\int f_\theta^{\frac{1}{2}}(x)\,\widehat{g}_n^{\frac{1}{2}}(x)\,dx$, using a quasi-newton procedure. Whilst it is intuitive that large bandwidths might reduce the robustness of this method it is less easy to imagine the effect of very small bandwidths. This function contains an integral of a kernel density estimate and so one might expect the optimal bandwidth to be quite small [17] and so twenty nine different bandwidths were considered ranging from $h = 0.00001$ to $h = 1.25$ in unequal steps. It is unlikely that the four very small bandwidths (0.00001, 0.0001, 0.001 and 0.01) chosen will be the optimal choice because one would expect the variance of the estimators to increase for h in this region. Their purpose is really to make certain that the optimal values suggested for h do not lie on the boundary of the search area. In the one parameter case $\widehat{HD}$ has two minima when based on asymmetrically contaminated data; one close to zero and the other at the contamination point. The method is extremely robust however and the global minimum remains close to zero for all the types of data considered therefore the issue of breakdown could be largely ignored. By restricting the search area for the minimisation procedure it was possible to ensure

that it was the global minimiser which was found each time.

5.5.3 Estimation of θ_*

The true density parameter θ_* was estimated in two ways. Firstly using the median of the data as an estimate of location and, in the two parameter case, the median absolute deviation to estimate dispersion. The alternative method was to use the Sheather-Jones bandwidth in the Hellinger distance estimating equation to obtain simpler, robust HD estimates, denoted $\widehat{\theta}_{sj}$ and $\widehat{\sigma}_{sj}$.

5.5.4 Minimising the $AMSE$ function

For each data set the data-based estimate of the $AMSE$ was calculated using each of the 29 bandwidths under consideration. The optimal value for the bandwidth was then chosen as the value of h which minimised this function.

5.6 Results and discussion

To allow the performance of these methods to be compared easily the mean squared errors (MSE) of the estimates obtained are summarised in Tables

5.3 to 5.7 on pages 132 to 136. Table 5.3 gives the $MSE's$ assuming that the variance of the data is known and Tables 5.4 and 5.5 give the results when both location and dispersion are unknown. Table 5.6 gives the overall $MSE's$ which are simply the sum of the individual $MSE's$ for each parameter. The results for the scale parameter in the $Gamma(4, \theta)$ model are given in Table 5.7.

Table 5.3: Simulation results: Mean squared errors of the HD estimates of θ when $f_\theta = N(\theta, 1)$.

	Other estimators			**Minimising AMSE**	
Distribution	*mean*	*median*	*HD(sj)*	$\theta_* = median$	$\theta_* = \widehat{\theta}_{sj}$
$\phi(x)$	0.008	0.013	0.009	0.009	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.331	0.028	0.013	0.015	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.011	0.017	0.010
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	0.015	0.032	0.015
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.084	0.023	0.012	0.012	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.014	0.016	0.014
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.013	0.031	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.111	0.043	0.011	0.013	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.034	0.022	0.025
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.113	0.095	0.089
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.201	0.052	0.020	0.026	0.016
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.672	0.123	0.023	0.052	0.017
$t_2(x)$	0.093	0.027	0.024	0.024	0.024
$t_3(x)$	0.023	0.021	0.014	0.016	0.015
$t_4(x)$	0.016	0.018	0.013	0.014	0.013

5.6.1 One parameter case

When $f_\theta = N(\theta, 1)$, as one would expect, maximum likelihood estimation leads to the smallest $MSE's$ when the data is not contaminated (Table

Table 5.4: Simulation results: Mean squared errors of the HD estimates of θ when $f_\theta = N(\theta, \sigma^2)$.

	Other estimators			**Minimising AMSE**	
Distribution	*mean*	*median*	*HD(sj)*	$\theta_* = median$	$\theta_* = \widehat{\theta}_{sj}$
$\phi(x)$	0.008	0.013	0.009	0.009	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.331	0.028	0.012	0.013	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.010	0.011	0.010
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	0.014	0.014	0.014
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.084	0.023	0.013	0.012	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.018	0.019	0.017
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.094	0.046	0.119
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.111	0.043	0.011	0.011	0.011
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.072	0.036	0.070
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.248	0.175	0.245
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.201	0.052	0.056	0.031	0.057
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.672	0.123	0.283	0.141	0.313
$t_2(x)$	0.093	0.027	0.026	0.025	0.027
$t_3(x)$	0.023	0.021	0.015	0.015	0.015
$t_4(x)$	0.016	0.018	0.013	0.014	0.013

5.3, p.132). When there is either symmetric or asymmetric contamination its lack of robustness means that maximum likelihood generally offers the largest $MSE's$ of any of the methods considered. The median offers much smaller $MSE's$ than maximum likelihood when there is asymmetric contamination in the data but performs less well for uncontaminated $N(0,1)$ data or data which has slightly heavy tails (i.e. t_3 or t_4 data). Using the Sheather-Jones bandwidth to obtain Hellinger distance estimates leads to further improvements in performance for all except the $N(0,1)$ data. Of the two methods which minimise the $AMSE$ function to determine the optimal value for h it seems that estimating the true location θ_* with $\widehat{\theta}_{sj}$ gives the best results on balance. The $MSE's$ for this method (column 6) are very

Table 5.5: Simulation results: Mean squared errors of the HD estimates of σ when $f_\theta = N(\theta, \sigma^2)$.

	Other estimators			**Minimising AMSE**	
Distribution	s	mad	$HD(sj)$	$\sigma_* = mad$	$\sigma_* = \widehat{\sigma}_{sj}$
$\phi(x)$	0.005	0.015	0.009	0.012	0.007
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	2.078	0.026	0.012	0.022	0.010
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	4.426	0.034	0.008	0.029	0.007
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	9.541	0.196	0.012	0.015	0.010
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.224	0.016	0.019	0.013	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.616	0.051	0.035	0.040	0.024
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	1.388	0.168	0.186	0.083	0.174
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	4.698	0.039	0.008	0.034	0.007
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.107	0.046	0.103	0.035	0.097
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.243	0.168	0.188	0.125	0.184
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.279	0.037	0.087	0.032	0.083
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.721	0.184	0.384	0.200	0.419
$t_2(x)$	3.910	0.084	0.314	0.078	0.286
$t_3(x)$	0.546	0.038	0.152	0.035	0.136
$t_4(x)$	0.178	0.031	0.113	0.027	0.100

similar to those obtained when putting the Sheather-Jones bandwidth directly into the Hellinger distance estimating equation (column 4) and thus appears to be a slight refinement which has the advantage of coping better with contamination at 3 and 4.

The MSE's when using maximum likelihood to estimate θ in the $Gamma(4, \theta)$ model (Table 5.7, p.136) increase considerably with the magnitude of the contamination. Using *mad* offers much smaller MSE's than maximum likelihood when the contamination is from $Gamma(4, 16)$ but for all other data performs less well than maximum likelihood. The HD method with Sheather-Jones bandwidth matches maximum likelihood for efficiency at

Table 5.6: Simulation results: Combined mean squared errors of the HD estimates of θ and σ when $f_\theta = N(\theta, \sigma^2)$.

	Other estimators			Minimising $AMSE$	
Distribution	*mean* / s	*median* / *mad*	*HD(sj)* / *HD(sj)*	$\theta_* = median$ / $\sigma_* = mad$	$\theta_* = \widehat{\theta}_{sj}$ / $\sigma_* = \widehat{\sigma}_{sj}$
$\phi(x)$	0.013	0.028	0.017	0.021	0.016
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	2.409	0.054	0.024	0.035	0.022
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	5.439	0.077	0.018	0.040	0.017
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	13.635	0.322	0.026	0.029	0.024
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.308	0.039	0.032	0.025	0.024
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.921	0.105	0.053	0.059	0.041
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	2.368	0.275	0.294	0.099	0.293
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	5.809	0.082	0.018	0.045	0.018
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.204	0.079	0.175	0.071	0.167
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.615	0.296	0.437	0.300	0.429
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.480	0.089	0.143	0.063	0.140
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	1.393	0.307	0.667	0.341	0.732
$t_2(x)$	4.003	0.110	0.340	0.103	0.312
$t_3(x)$	0.569	0.059	0.167	0.050	0.151
$t_4(x)$	0.194	0.049	0.125	0.041	0.113

the model and leads to greatly reduced MSE's for contamination from $Gamma(4, 16)$ but surprisingly performs slightly less well than maximum likelihood for contamination from $Gamma(4, 8)$. Minimising a data-based estimate of the $AMSE$ function with $\theta_* = mad$ leads to the smallest MSE's for contamination from $Gamma(4, 8)$ but performs no better than the alternative $\theta_* = \widehat{\theta}_{sj}$ for data from the other two distributions.

5.6.2 Two parameter case

When both θ and σ are unknown in the model $f_\theta = N(\theta, \sigma^2)$ the results for location are similar to those obtained when σ is known (Table 5.4, p.133).

Table 5.7: Simulation results: Mean squared errors of the HD estimates of θ when $f_\theta = Gamma(4, \theta)$.

	Other estimators			**Minimising AMSE**	
Distribution	*mle*	*mad*	$\widehat{\theta}_{sj}$	$\theta_* = mad$	$\theta_* = \widehat{\theta}_{sj}$
$Gamma(4,1)$	0.003	0.018	0.003	0.004	0.004
$0.9Gamma(4,1) + 0.9Gamma(4,8)$	0.016	0.023	0.017	0.015	0.016
$0.9Gamma(4,1) + 0.9Gamma(4,16)$	0.099	0.024	0.017	0.014	0.014

The HD based methods are considerably more efficient at the model than the median and cope very well with contamination at 10 but, as in the one parameter case, they are generally less able to handle contamination at 3 or 4. Although the $MSE's$ for this data are considerably smaller than those for maximum likelihood they can be twice the size of those obtained by just using the *median* and *mad* for θ and σ respectively. Another surprising result is that for all three of the HD based methods the $MSE's$ for $N(0,1)$ data with contamination at 10 are smaller when both parameters are estimated than in the one parameter case when σ is known. This counter-intuitive effect is repeated when the contamination is at -10 or from the t_2 distribution and leads to the uneasy recommendation that for data with large outliers estimating both parameters is to be preferred irrespective of whether σ is known. For all other types of data the results are, as one would expect, either the same as in the one parameter case or worse.

The results for the dispersion parameter (Table 5.5, p.134) again demonstrate that the HD based methods are much more efficient than *mad* but

less robust to contamination at 3, 4 and 5. The Hellinger distance method performs fairly well using the Sheather-Jones bandwidth but the results are improved when the choice of bandwidth is the minimiser of the $AMSE$ function. Neither of the two ways of estimating the true parameters θ_* and σ_* does consistently better than the other although on balance it seems that the *median/mad* should be the preferred choice.

The combined simulation results (Table 5.6, p.135) confirm that there is much to recommend the use of Hellinger distance estimators for many types of data. The robust methods perform better than maximum likelihood for all but the $N(0,1)$ data but none is clearly better than its rivals. The main disadvantage of using the *median* for location and *mad* for dispersion is that these estimators are very inefficient at the model and become rapidly less robust as the percentage of contaminants increases to 20%. The HD based methods do not cope well with contamination at 3 or 4, particularly when the contamination percentage is 20%. The performance of the two optimal HD methods can be very sensitive to the choice of estimator for the true location and dispersion but neither method consistently does better than the other. Using the *median* and *mad* respectively appears to be the best option for data with contamination at 3, 4 or 5 or where the data is heavy tailed, however, $\widehat{\theta}_{sj}$ and $\widehat{\sigma}_{sj}$ lead to smaller $MSE's$ when there is contamination at 10 and greater efficiency at the model.

Perhaps the most surprising result is that in these simulations none of the HD based methods were fully efficient despite being asymptotically equivalent to maximum likelihood for $N(0,1)$ data. Asymptotic equivalence does not imply equality in the finite case so one should not expect the performance of the HD estimators to rival that of maximum likelihood for every realisation of data, however, in a simulation setting where the method is being repeatedly applied to random samples one might reasonably expect the performance of these two methods to be more closely allied. The asymptotic behaviour of these estimators depends on the distance measure itself, the choice of bandwidth, the sample size and number of samples taken, therefore it would be interesting to repeat these simulations on larger samples of data and/or a larger number of samples to establish whether these results have simply occurred by chance or represent a real difference in performance.

5.6.3 Comparison to theoretical results

Since the first order $TAMSE$ function for HD estimators does not involve h it could not be used to determine the theoretical optimal bandwidth for any particular data set. Instead an estimate of the optimal bandwidth was obtained by inspection using a range of bandwidths ($h = 0.05, 0.1, 0.2, 0.3, 0.35, 0.4$) on the simulated data. The bandwidth which led to the smallest MSE for each dataset was used to estimate the theoreti-

cal optimal bandwidth and is shown in the column headed "By inspection" alongside the theoretical minimum $AMSE$ attainable and data-based results (with $\theta_* = median$ and $\sigma_* = mad$) in Tables 5.8 to 5.10.

For the model $f_\theta = N(\theta, 1)$ (Table 5.8, p.145) the estimates for the optimal value of h obtained by inspection appear to be a little inconsistent. The optimal bandwidth obtained in this way for $N(0, 1)$ data with contamination at either 5 or 10 is generally 0.3 but, strangely, came out at 0.1 when the percentage of contamination is 10%. For these two data sets the difference between the magnitude of the $AMSE$ when $h = 0.1$ and $h = 0.3$ is just 0.0003 and so the fact that these values jump around a little should not cause undue concern. However, with contamination at 3 or 4 the optimal value of h is also very small, at 0.05 or 0.1, but in this case the reduction in the MSE is much larger and suggests that these values are the true minimisers. The estimates of the optimal value of h obtained by minimising a data-based estimate of the $AMSE$ (column 4) are generally much smaller than those obtained by inspection (column 3). The MSE function is very flat over this range of bandwidths and so, despite being poor at estimating the optimal h, the new method performs very well with $MSE's$ being generally very close to the minimum values obtained by inspection. The exception to this occurs when the contamination is at 3, in which case the $MSE's$ obtained by minimising the $AMSE$ function are much larger (in percentage terms) than

those obtained by inspection being 0.095 and 0.033 respectively when the contamination percentage is 20%. Overall, the $MSE's$ for the new method compare reasonably favourably with the theoretical minimum values and confirm that choosing the value of h in this way ensures that the desirable theoretical properties of HD estimators are attained in practice.

The average optimal h obtained by applying the new method to *Gamma* data, shown in Table 5.9 (p.146), are close to the theoretical values obtained by inspection only when there is no contamination in the data. For contaminated *Gamma* data the optimal value for h found by inspection (column 3) was much smaller than the values found by minimising the $AMSE$ function (column 4). The corresponding optimal MSE's (column 5) were also considerably smaller than those obtained by minimising the data-based estimate of the $AMSE$ (column6) which leads one to suspect that, for this data, the first order approximation to the $AMSE$ used here is not sufficient. Furthermore, the smallest $MSE's$ were obtained by using very small bandwidths ($h = 0.05$ or $h = 0.1$) which confirm the notion, discussed in Section 5.2 (p.118), that for contaminated data small bandwidths lead to increased robustness. Thus it seems that for this data the new method is not able to select the optimal value for the bandwidth. These results contrast quite strongly those for the one parameter *Normal* model but it should be noted that in this case it is the scale parameter which is being estimated and not

location.

For the model $f_\theta = N(\theta, \sigma^2)$ (Table 5.10, p.147) the optimal values of h obtained using the new method are generally much larger than those obtained by inspection. For data which is heavy-tailed or has contamination at 3 or 4, the values for h suggested by the new method can be more than twice the size of those found by inspection. Thus the pattern seen for the scale parameter in the *Gamma* model is repeated for the *Normal* family of models when both location and dispersion are unknown. That is to say, that when dispersion is being estimated from data with contamination which is plausible under the model the robustness of the HD method can be greatly improved by making the bandwidth very small. Furthermore, the new method does not appear to recognise this feature and consequently performs less well than one would hope. There are two possible explanations for the new method's inability to find the optimal bandwidth for such data. Firstly, it could be that the first-order approximation to the $AMSE$ does accurately reflect the role of the bandwidth on the performance of the HD method and that it is the data-based estimate of this function which is the root of the problem. Alternatively, it could be that further terms in the approximation to the $AMSE$ are needed to fully explain the role of the bandwidth. Furthermore, the minimum $MSE's$ obtained by inspection are, in many cases, much lower than the theoretical minimum $AMSE's$ which

leads to further suspicion that this first-order approximation is not good enough.

As previously explained in Sections 5.1.1 and 5.2 it is not surprising that the optimal bandwidth for contaminated data should be very small. It is interesting, however, that this optimality should be more apparent for the scale parameter than location, occur only for data from particular distributions and that the magnitude of the potential improvement in performance is quite considerable. These issues can be explained by considering the interaction between the distance measure and kernel density estimate in more detail as follows. Contamination which is unlikely under the model, at 10 in $N(0,1)$ data for example, is fully downweighted by the factor $f_{\widehat{\theta}_h}^{-\frac{1}{2}}(x)\left.\frac{df(x)}{d\widehat{\theta}}\right|_{\theta=\widehat{\theta}_h}$ in the integrand of the estimating equation. Thus there is little to be gained by making the bandwidth very small. However when the contamination is not dealt with adequately by the HD method, at 3 or 4 for $f_\theta = N(\theta,1)$ for example, making the bandwidth very small can increase robustness because the normal curves which are placed around each data point are then very narrow. The effect of any one data point on the overall kernel density estimate is therefore very much reduced because the contribution each point makes will be effectively zero everywhere except within the narrow range of $\pm 3h$ around itself. Furthermore, small bandwidths mean that the tails of the kernel density estimate fall very rapidly to zero

outside the range of the data which means that the resulting HD estimates for dispersion are not unduly inflated.

5.7 Conclusions

The simulation results when dispersion is known confirm that the Hellinger distance method performs well for many bandwidths. It is only bandwidths which are either very small or very large in relation to the spread of the data (in this case $h < 0.01$ and $h > 0.5$) that should be avoided. The situation when the dispersion of the data is unknown is less clear and there is evidence to suggest that σ is far more sensitive to the choice of h than θ. Furthermore, it seems that by choosing a small bandwidth the robustness of the HD estimation procedure can be increased beyond that which one should expect from its asymptotic properties. Unfortunately, when the dispersion parameter is estimated the new method never suggests very small values for h and its performance, compared to the optimal values found by inspection, is disappointing. It may be that this could be improved by carrying out further work to obtain a more rapidly converging approximation to the asymptotic variance. However, it should be emphasised that although the new method is working less well than one would like, it nevertheless leads to considerably smaller MSE's than either maximum likelihood or

the *median/mad* combination when there is contamination in the data and is highly efficient at the model.

Table 5.8: Theoretical and data-based mean squared errors of the HD estimator when $f_\theta = N(\theta, 1)$.

Distribution	Theoretical Minimum AMSE	Data-based Results			
		Optimal h		Mean Squared Error	
		Inspection	Minimising $AMSE$ $\theta_* = median$	Inspection	Minimising $AMSE$ $\theta_* = median$
$\phi(x)$	0.010	0.30	0.07	0.009	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.011	0.30	0.09	0.012	0.015
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	0.011	0.10	0.08	0.010	0.017
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	0.013	0.30	0.12	0.015	0.032
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.011	0.30	0.09	0.012	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.011	0.10	0.18	0.010	0.016
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.013	0.30	0.36	0.015	0.031
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	0.011	0.30	0.11	0.011	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.021	0.05	0.15	0.016	0.022
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.039	0.05	0.30	0.033	0.095
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.013	0.05	0.22	0.013	0.026
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.015	0.10	0.44	0.014	0.052
t_2	0.018	0.40	0.06	0.024	0.024
t_3	0.016	0.40	0.07	0.014	0.016
t_4	0.015	0.40	0.07	0.013	0.014

Table 5.9: Theoretical and data-based mean squared errors of the HD estimator when $f_\theta = Gamma(4, \theta)$.

Distribution	Theoretical Minimum $AMSE$	Data-based Results			
		Optimal h		Mean Squared Error	
		Inspection	Minimising $AMSE$ $\theta_* = mad$	Inspection	Minimising $AMSE$ $\theta_* = mad$
$Gamma(4,1)$	0.006	0.75	0.76	0.003	0.004
$0.9Gamma(4,1) + 0.1Gamma(4,8)$	0.009	0.05	0.64	0.003	0.015
$0.9Gamma(4,1) + 0.1Gamma(4,16)$	0.007	0.10	0.61	0.004	0.014

Table 5.10: Theoretical and data-based mean squared errors of the HD estimator when $f_\theta = N(\theta, \sigma^2)$.

Distribution		**Data-based Results**			
	Theoretical Minimum *AMSE*	**Optimal** h		**Mean Squared Error**	
		Inspection	**Minimising** *AMSE* $\theta_* = median$, $\sigma_* = mad$	**Inspection**	**Minimising** *AMSE* $\theta_* = median$, $\sigma_* = mad$
$\phi(x)$	0.015	0.30	0.23	0.013	0.021
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.016	0.30	0.38	0.018	0.035
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	0.017	0.30	0.52	0.014	0.040
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	0.019	0.30	0.89	0.020	0.029
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.017	0.20	0.28	0.015	0.025
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.018	0.20	0.39	0.020	0.059
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.022	0.20	0.42	0.019	0.099
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	0.017	0.30	0.52	0.015	0.045
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.062	0.05	0.15	0.023	0.071
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.157	0.05	0.25	0.040	0.300
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.030	0.10	0.25	0.029	0.063
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.066	0.05	0.25	0.029	0.341
t_2	0.257	0.05	0.11	0.065	0.103
t_3	0.128	0.05	0.12	0.031	0.050
t_4	0.081	0.05	0.13	0.026	0.041

Chapter 6

A method for choosing the value of p in the OH estimator

6.1 Background

The family of criterion functions was introduced by Öztürk and Hettmansperger in 1996 [23] and is defined as

$$d_F(\theta; p) = \int [G^p(x) - F_\theta^p(x)]^2 \, dx + \int [(1 - G(x))^p - (1 - F_\theta(x))^p]^2 \, dx \tag{6.1}$$

where $F_\theta(x)$ is the model distribution function, $G(x)$ is the true distribution function and $p > 0$.

This distance measure is minimised by θ_p, the solution to the following equation, which is obtained by differentiating $d_F(\theta; p)$ with respect to θ and setting equal to zero.

$$\begin{aligned} 0 &= \int \left(G^p(x) - F_\theta^p(x)\right) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} dx \\ &\quad - \int \left[(1 - G(x))^p - (1 - F_\theta(x))^p\right] [1 - F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} dx. \end{aligned} \tag{6.2}$$

The true distribution function $G(x)$ is rarely known and is therefore estimated by the empirical distribution function $F_n(x)$. The Öztürk and Hettmansperger (OH) estimator of θ, denoted $\widehat{\theta}_p$, is therefore the minimiser of $d_{F_n}(\theta; p)$ and solution to the estimating equation

$$\begin{aligned} 0 &= \int \left(F_n^p(x) - F_\theta^p(x)\right) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} dx \\ &\quad - \int \left[(1 - F_n(x))^p - (1 - F_\theta(x))^p\right] [1 - F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} dx. \end{aligned} \tag{6.3}$$

When $f_\theta = N(\theta, 1)$ the two integrals not involving F_n can be integrated by parts and shown to sum to zero. In this case, therefore, the estimating equation simplifies to

$$0 = \int F_n^p(x) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} dx - \int [1 - F_n(x)]^p [1 - F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} dx. \tag{6.4}$$

The robustness and efficiency properties of these estimators are determined by the choice of p. The method is robust when $p > 0.5$ and highly efficient at estimating location for all values of p. It is less efficient at estimating dispersion, however, with efficiency in excess of 80% only when $p < 1$. When simultaneously estimating location and dispersion it is clear, therefore, that choosing $p > 1$ for data which is not contaminated would lead to inefficiency. Similarly, when the data is contaminated choosing $p < 1$ will result in sub-optimal robustness. Although this method has the potential to provide estimators which are both robust and efficient these desirable properties will not be attained in practice unless the value of p chosen is appropriate for the data to which it is applied.

6.2 Estimation of the asymptotic mean squared error

As in the previous two chapters, the asymptotic mean squared error ($AMSE$) will be used to jointly measure robustness and efficiency. It is hoped that by minimising an expression for this function, which is a function of p, the

optimal value for p can be determined. The $AMSE$ function for the OH estimator is

$$As.E\left[\left(\hat{\theta}_p-\theta_*\right)\left(\hat{\theta}_p-\theta_*\right)^T\right]=\left(\theta_p-\theta_*\right)\left(\theta_p-\theta_*\right)^T+As.var\left(\hat{\theta}_p\right)$$

where $\hat{\theta}_p$ is the solution to equation (6.3), θ_p the solution to equation (6.2) and θ_* is the true parameter. This function is obtained by substituting θ_p for θ and $\hat{\theta}_p$ for $\hat{\theta}$ in the expression for the multi-parameter $AMSE$ (equation D.1) of Appendix D (p.253).

The asymptotic variance of $\sqrt{n}\left(\widehat{\theta_p}-\theta_p\right)$ is $J^{-1}KJ^{-1}$ where

$$\begin{aligned}
J &= (1-2p)\int F_\theta^{2p-2}(x)\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&+\int G^p(x)\,F_\theta^{p-1}(x)\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&+(p-1)\int G^p(x)\,F_\theta^{p-2}(x)\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&-\int F_\theta^{2p-1}(x)\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&+(1-2p)\int (1-F_\theta(x))^{2p-2}\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&-\int (1-G(x))^p\,(1-F_\theta(x))^{p-1}\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&+(p-1)\int (1-G(x))^p\,(1-F_\theta(x))^{p-2}\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&+\int (1-F_\theta(x))^{2p-1}\,\frac{d^2F_\theta(x)}{d\theta^2}dx. \qquad (6.5)
\end{aligned}$$

$$K = 2p^2 \iint_{s<t} (r(s) + u(s))(r(t) + u(t))^T G(s)(1 - G(t))\, ds\, dt \quad (6.6)$$

and $r(x) = G^{p-1}(x) \frac{dF_\theta}{d\theta}(x) F_\theta^{p-1}(x)$, $u(x) = (1 - G(x))^{p-1} \frac{dF_\theta}{d\theta}(x)(1 - F_\theta(x))^{p-1}$ and $\theta = \theta_p$.

The derivation of the asymptotic properties of OH estimators is shown in detail in Appendix A.3 (p.227).

When $f_\theta = N(\theta, \sigma^2)$ the expression for J (equation 6.5) simplifies to give

$$J = p \int [r(s) + u(s)]\, g(s) ds \quad (6.7a)$$

$r(x) = G^{p-1}(x) \frac{dF_\theta(x)}{d\theta} F_\theta^{p-1}(x)$ and $u(x) = (1 - G(x))^{p-1} \frac{dF_\theta(x)}{d\theta}(1 - F_\theta(x))^{p-1}$.

The $AMSE$ function for the OH estimator is therefore

$$AMSE(p) = (\theta_p - \theta_*)(\theta_p - \theta_*)^T + \frac{1}{n} J^{-1} K J^{-1}$$

where J and K are as defined in equations (6.5) and (6.6) respectively, θ_p is the solution to equation (6.2) and θ_* is the true parameter.

Once again, the trace of the $AMSE$ matrix will be used to provide a global measure for minimisation. Thus when there are two unknown parameters (θ and σ for example) the expression to be minimised is

$$AMSE(\hat{\theta}_p; \widehat{\sigma}_p) \simeq As.var(\widehat{\theta}_p) + As.var(\widehat{\sigma}_p) + (\theta_p - \theta_*)^2 + (\sigma_p - \sigma_*)^2 \quad (6.8)$$

As in the previous chapters, the minimiser of this function is found using numerical methods.

Given the good results obtained with a data-based approximation to the $AMSE$ for the minimum density power divergence estimator (Chapter 4, p.76) and the problems encountered in trying to obtain a quadratic approximation to the $AMSE$ for both the density based estimators, the quadratic approximation was not attempted for this method. Instead, the unknown parameters in the $AMSE$ function were replaced by suitable estimates from data and this estimated function then minimised to find the optimal p. Therefore, with $\widehat{\theta}_p$ and $\widehat{\theta}_*$ replacing θ_p and θ_* respectively, the data-based estimate of the $AMSE$ is then

$$\widehat{AMSE}(p) = \left(\widehat{\theta}_p - \widehat{\theta}_*\right)\left(\widehat{\theta}_p - \widehat{\theta}_*\right)^T + \frac{1}{n}\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1} \tag{6.9}$$

where

$$\widehat{K} = 2p^2 \iint_{s<t} \left(\widehat{r}(s) + \widehat{u}(s)\right)\left(\widehat{r}(t) + \widehat{u}(t)\right)^T F_n(s)\left(1 - F_n(t)\right) dsdt \tag{6.10}$$

$$
\begin{aligned}
\widehat{J} &= (1-2p)\int F_\theta^{2p-2}(x)\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&+ \int F_n^p(x)\,F_\theta^{p-1}(x)\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&+ (p-1)\int F_n^p(x)\,F_\theta^{p-2}(x)\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&- \int F_\theta^{2p-1}(x)\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&+ (1-2p)\int (1-F_\theta(x))^{2p-2}\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&- \int (1-F_n(x))^p\,(1-F_\theta(x))^{p-1}\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&+ (p-1)\int (1-F_n(x))^p\,(1-F_\theta(x))^{p-2}\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&+ \int (1-F_\theta(x))^{2p-1}\,\frac{d^2F_\theta(x)}{d\theta^2}dx. \qquad (6.11)
\end{aligned}
$$

where $\widehat{r} = F_n^{p-1}\frac{dF_\theta}{d\theta}F_\theta^{p-1}$, $\widehat{u} = (1-F_n)^{p-1}\frac{dF_\theta}{d\theta}(1-F_\theta)^{p-1}$ and $\theta = \widehat{\theta}_p$.

Notice, however, that the expression for $\widehat{J}$ contains F_θ and F_n to the power $p-2$ so there will be singularities when $p < 2$.

When $f_\theta = N(\theta, \sigma^2)$ $\widehat{J}$ is

$$
\begin{aligned}
\widehat{J} &= p\int \left[[1-F_n(x)]^{p-1}[1-F_\theta(x)]^{p-1} - F_n^{p-1}(x)\,F_\theta^{p-1}(x)\right] f_\theta(x)\,f_n(x)dx \\
&= \frac{p}{n}\sum_{i=1}^{n}\left\{\left[[1-F_n(X_i)]^{p-1}[1-F_\theta(X_i)]^{p-1} - F_n^{p-1}(X_i)\,F_\theta^{p-1}(X_i)\right] f_\theta(X_i)\right\}
\end{aligned}
$$

which is evaluated at $\theta = \widehat{\theta}_p$. In this case, the distribution functions are to the power $p-1$ and so the estimation of the variance should be straightforward for $p > 1$.

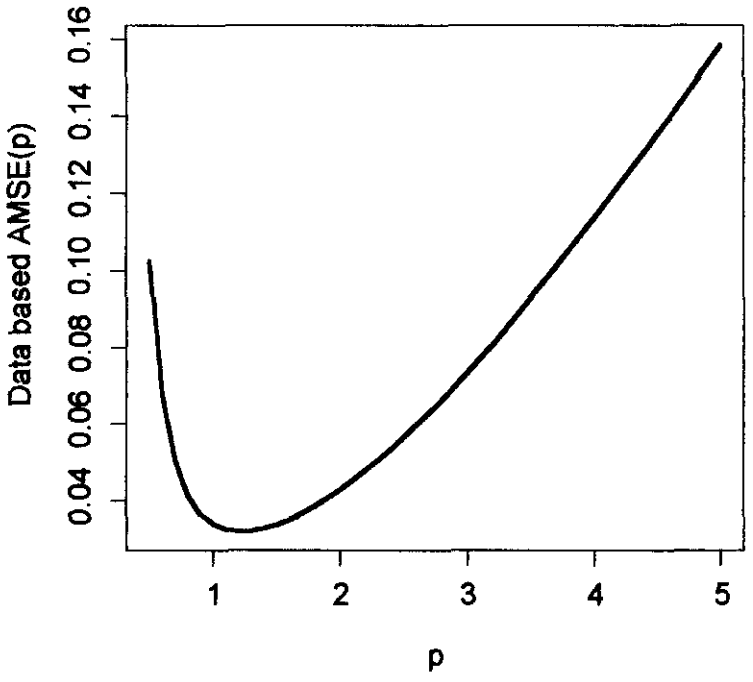

Figure 6.1: Data-based estimate of the $AMSE$ function of the OH estimator using $f_\theta = N(\theta, 1)$ for $N(0, 1)$ data with 10% contamination at 10.

The problem remains, however, of how to estimate θ_* in the bias term of the $AMSE$ function. Recall that this is the true parameter which is unknown and the fact that it is difficult to estimate from data is the very reason that robust methods were developed. Progress in solving this circular problem can be made by treating θ_* as a nuisance parameter in the $AMSE$ function and replacing it with the most robust estimate which can be obtained simply. In this case, the median of the data seems an appropriate choice for location and the median absolute deviation (*mad*) for dispersion but there are many

others which could be considered.

The data-based estimate of the $AMSE$ function for $N(0,1)$ data with 10% contamination at 10 (using the model $f_\theta = N(\theta, 1)$) is shown in Figure 6.1 (p.155). This function has a definite minimum which is close to $p = 1.2$ which suggests that the problems experienced with the BHHJ method (where the estimates of the optimal α varied considerably because the $AMSE$ function was very flat) are unlikely to recur here.

6.3 Assessing the performance of this new method

Continuing the approach to testing taken in the previous two chapters, the performance of this method was assessed by its application to simulated data. The same data sets were used here as with the BHHJ and HD methods so that the performance of these three methods could be compared easily. Full details of these data sets, along with the reasons for their choice, are given in Section 4.3 (p.83). They briefly comprise 100 random samples of size 100 from a perturbation of the true density g and thus provide simulated data with a variety of attributes such as heavy-tails or outliers.

As before, the correct family of models is used in each case; the *Normal*

family for the predominantly $N(0,1)$ data and the *Gamma* family for contaminated and uncontaminated *Gamma* data.

6.4 Theoretical asymptotic mean squared error

The behaviour of OH estimators can be explored by applying the method to simulated data drawn from the probability density function $g_\varepsilon(z)$, a perturbation of the true density $g(z)$. Then, taking a data-based approach, simulated data sets can be generated by taking random samples from $g_\varepsilon(z)$ and applying the method to each data set in turn. Alternatively, the theoretical asymptotic mean squared error ($TAMSE$) can be obtained by substituting $g_\varepsilon(z)$ for $g(z)$ and $G_\varepsilon(z)$ for $G(z)$ in equations (6.5) to (6.6). Applying the same substitution to the estimating equation (6.2) and solving equal to zero provides the theoretical θ_p's, denoted θ_p^*, which replace the $\widehat{\theta}_p$'s in the formulae for the $AMSE$ (6.9-6.11). The location parameter of the underlying distribution is now known and so θ_* replaces $\widehat{\theta}_*$ to give

$$TAMSE(p) = \left(\theta_p^* - \theta_*\right)\left(\theta_p^* - \theta_*\right)^T + \frac{1}{n}J^{-1}KJ^{-1} \qquad (6.12)$$

where

$$K = 2p^2 \iint_{s<t} (r(s) + u(s))(r(t) + u(t))^T G_\varepsilon(s)(1 - G_\varepsilon(t))\, ds\, dt \tag{6.13}$$

with $r = G_\varepsilon^{p-1} \frac{df_\theta}{d\theta} F_\theta^{p-1}$ and $u = (1 - G_\varepsilon)^{p-1} \frac{df_\theta}{d\theta} (1 - F_\theta)^{p-1}$

and

$$\begin{aligned} J &= p \int (G_\varepsilon^p(x) - F_\theta^p(x)) \left[F_\theta^{p-1}(x) \frac{d^2 F_\theta}{d\theta^2} + (p-1) F_\theta^{p-2}(x) \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T \right] dx \\ &\quad -p \int [(1 - G_\varepsilon(x))^p - (1 - F_\theta(x))^p] \left[(1 - F_\theta(x))^{p-1} \frac{d^2 F_\theta}{d\theta^2} \right] dx \\ &\quad +p \int [(1 - G_\varepsilon(x))^p - (1 - F_\theta(x))^p] \left[(p-1)(1 - F_\theta(x))^{p-2} \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T \right] dx \\ &\quad -p \int \left[F_\theta^{2p-2}(x) + (1 - F_\theta(x))^{2p-2}\right] \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx \end{aligned} \tag{6.14}$$

with both evaluated at $\theta = \theta_p$.

When $f_\theta = N(\theta, \sigma^2)$

$$J = p \int \left[[1 - G_\varepsilon(x)]^{p-1} [1 - F_\theta(x)]^{p-1} - G_\varepsilon^{p-1}(x) F_\theta^{p-1}(x) \right] f_\theta(x)\, g_\varepsilon(x)\, dx. \tag{6.15}$$

The p which minimises the $TAMSE$ can be found numerically to provide a theoretical optimal p which then provides a benchmark for the data-based results.

Because the formulae involve distribution functions, rather than density functions, to the power p the $TAMSE$ only simplifies in the case $f_\theta = g$. In

this case there is no bias and the variance term $\frac{1}{n}J^{-1}KJ^{-1}$can be obtained as follows

$$K = 2p^2 \iint_{s<t} \left(r\left(s\right)+u\left(s\right)\right)\left(r\left(t\right)+u\left(t\right)\right)^T G_\varepsilon\left(s\right)\left(1-G_\varepsilon\left(t\right)\right) ds\, dt$$

with $r\left(x\right) = \frac{df_\theta}{d\theta}\left(x\right) F_\theta^{2p-2}\left(x\right)$ and $u\left(x\right) = \frac{df_\theta}{d\theta}\left(x\right)\left(1-F_\theta\right)^{2p-2}\left(x\right)$ and

$$J = p\int \left[\left[1-F_\theta\left(x\right)\right]^{2p-2} + F_\theta^{2p-2}\left(x\right)\right]\left[\frac{dF_\theta}{d\theta}\right]\left[\frac{dF_\theta}{d\theta}\right]^T dx.$$

Note that in this case, the variance has singularities only when $p < \frac{1}{2}$.

When $g \neq f_\theta$ the $TAMSE$ can be obtained by making appropriate substitutions for f_θ, g_ε and G_ε into equations (6.12) to (6.14). For contaminated normal data with $\varepsilon\%$ contamination at ξ, $g_\varepsilon\left(z\right) = \left(1-\varepsilon\right)\phi\left(z\right) + \varepsilon\delta_\xi\left(z\right)$, where $\delta_\xi\left(z\right) = \infty$ if $\xi = z$ and 0 otherwise and $G_\varepsilon\left(z\right) = \left(1-\varepsilon\right)\Phi\left(z\right) + \varepsilon\Delta_\xi\left(z\right)$, where $\Delta_\xi\left(z\right) = 1$ if $\xi > z$ and 0 otherwise. So, for example, using the model $f_\theta = N\left(\theta, 1\right)$ for these data leads to

$$K = 2p^2 \iint_{s<t} \left(r\left(s\right)+u\left(s\right)\right)\left(r\left(t\right)+u\left(t\right)\right)^T G_\varepsilon\left(s\right)\left[1-G_\varepsilon\left(t\right)\right] dsdt$$

where $r\left(x\right) = \left[\left(1-\varepsilon\right)\Phi\left(x\right) + \varepsilon\Delta_\xi\left(x\right)\right]^{p-1}\phi_\sigma\left(x-\theta\right)\Phi_\sigma^{p-1}\left(x-\theta\right)$ and

$u\left(x\right) = \left[1-\left(1-\varepsilon\right)\Phi\left(x\right) - \varepsilon\Delta_\xi\left(x\right)\right]^{p-1}\phi_\sigma\left(x-\theta\right)\left(1-\Phi_\sigma\left(x-\theta\right)\right)^{p-1}$

and

$$\begin{aligned} J &= p\int \left[1-G_\varepsilon\left(x\right)\right]^{p-1}\left[1-\Phi_\sigma\left(x-\theta\right)\right]^{p-1}\phi_\sigma\left(x-\theta\right) g_\varepsilon\left(x\right) dx \\ &\quad -p\int G_\varepsilon^{p-1}\left(x\right)\Phi_\sigma^{p-1}\left(x-\theta\right)\phi_\sigma\left(x-\theta\right) g_\varepsilon\left(x\right) dx. \end{aligned}$$

Similarly, the bias is obtained as before as $\left(\theta_p^* - \theta_*\right)\left(\theta_p^* - \theta_*\right)^T$. The target parameter, θ_*, is known and therefore replaced by its true value and θ_p^* is the solution to the following estimating equation

$$\begin{aligned} 0 &= \int G_\varepsilon^p(x)\,\Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,dx \\ &\quad - \int [1-G_\varepsilon(x)]^p\,[1-\Phi_\sigma(x-\theta)]^{p-1}\,\phi_\sigma(x-\theta)\,dx \end{aligned}$$

Using the same model for data from a t distribution with k degrees of freedom, $g_\varepsilon = t_k$ and $G_\varepsilon = T_k$ and so K and J can be calculated as follows

$$K = 2p^2 \iint_{s<t} (r(s)+u(s))\,(r(t)+u(t))^T\,T_k(s)\,(1-T_k(t))\,dsdt$$

with $r(x) = T_k^{p-1}(x)\,\phi_\sigma(x-\theta)\,\Phi_\sigma^{p-1}(x-\theta)$ and

$u(x) = (1-T_k(x))^{p-1}\,\phi_\sigma(x-\theta)\,(1-\Phi_\sigma(x-\theta))^{p-1}$ and

$J = p^2 \int \left[[1-T_k(x)]^{p-1}\,[1-\Phi_\sigma(x-\theta)]^{p-1} - T_k^{p-1}(x)\,\Phi_\sigma^{p-1}(x-\theta)\right]\phi_\sigma(x-\theta)\,t_k(x)\,d$

As in the previous example, the target parameter in the bias is known and θ_p^* is obtained by solving

$$\begin{aligned} 0 &= \int T_k^p(x)\,\Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,dx \\ &\quad - \int [1-T_k(x)]^p\,[1-\Phi_\sigma(x-\theta)]^{p-1}\,\phi_\sigma(x-\theta)\,dx. \end{aligned}$$

The theoretical optimal values for p and their respective asymptotic mean squared errors for data from various distributions using $f_\theta = N(\theta,1)$ are

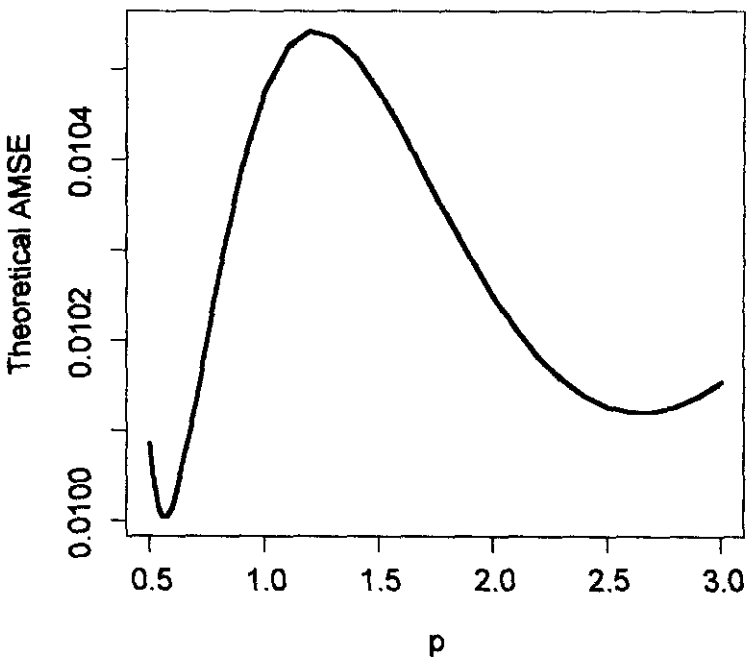

Figure 6.2: Theoretical $AMSE$ function for the OH estimator with $f_\theta = N(\theta, 1)$ for $N(0, 1)$ data.

Table 6.1: Theoretical optimal values of p for the OH estimator when $f_\theta = N(\theta, 1)$.

Distribution of data	**Theoretical Optimal p**	**Theoretical Minimum AMSE**
$\phi(x)$	0.57	0.010
$0.95\phi(x) + 0.05\Delta_{10}(x)$	1.20	0.021
$0.9\phi(x) + 0.1\Delta_{10}(x)$	1.20	0.053
$0.8\phi(x) + 0.2\Delta_{10}(x)$	1.20	0.222
$0.95\phi(x) + 0.05\Delta_5(x)$	1.20	0.021
$0.9\phi(x) + 0.1\Delta_5(x)$	1.20	0.053
$0.8\phi(x) + 0.2\Delta_5(x)$	1.20	0.222
$0.9\phi(x) + 0.1\Delta_{-10}(x)$	1.20	0.053
$0.9\phi(x) + 0.1\Delta_3(x)$	1.20	0.053
$0.8\phi(x) + 0.2\Delta_3(x)$	1.20	0.218
$0.9\phi(x) + 0.1\Delta_4(x)$	1.20	0.053
$0.8\phi(x) + 0.2\Delta_4(x)$	1.20	0.221
t_2	1.20	0.018
t_3	1.20	0.016
t_4	1.20	0.014

given in Table 6.1 (p.162). For $N(0, 1)$ data the minimum attainable $AMSE$ is 0.01 which is the same as that which could be achieved by using the optimal choice of α for these data with the BHHJ method (i.e. $\alpha = 0$). This means that when there is no contamination in the data, providing the optimal value for p can be identified, the performance of this method should rival that of BHHJ with optimally chosen α and also, therefore, maximum likelihood. For all the other types of data considered the optimal theoretical value for p is 1.2 with the theoretical minimum $AMSE$ attainable varying according to the degree and magnitude of the contamination.

The theoretical $AMSE$ function for $N(0, 1)$ data is plotted against p in

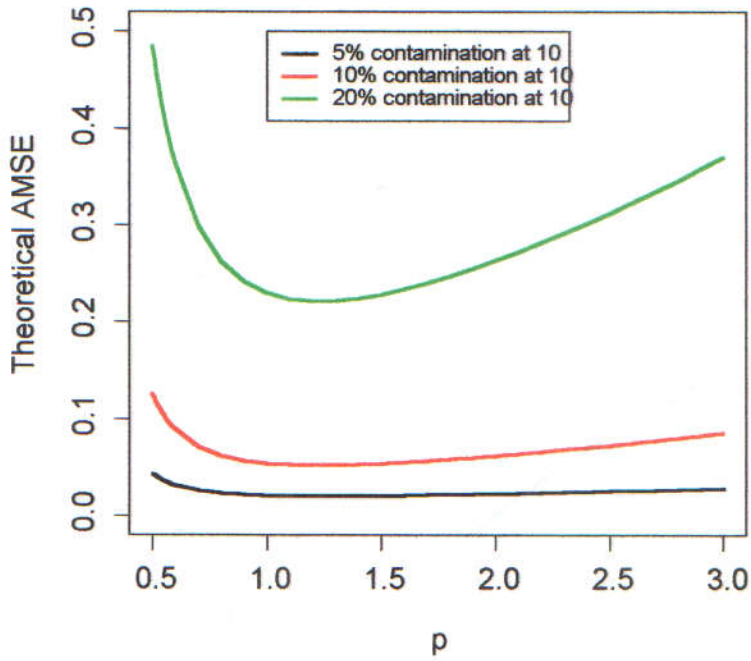

Figure 6.3: Theoretical $AMSE$ functions for the OH estimator with $f_\theta = N(\theta, 1)$ for data from various distributions.

Figure 6.2 (page 161) and has two minima: one at $p = 0.57$ and another at $p = 2.7$. The function is not necessarily bimodal however, as illustrated by the same plot for asymmetrically contaminated $N(0, 1)$ data in Figure 6.3 (page 163). In this case, the position of the global minimum remains fixed at $p = 1.2$ as the degree of contamination and value of the corresponding minimum increase.

The theoretical optimal values for p when both location and dispersion are unknown are given in Table 6.2. They show a similar pattern to those

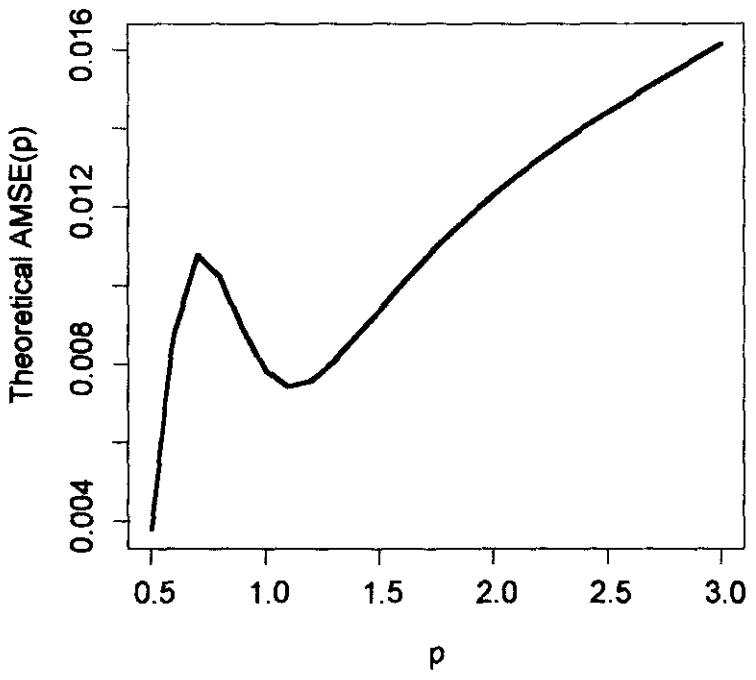

Figure 6.4: Theoretical $AMSE$ function for the OH estimator with $f_\theta = N(\theta, \sigma^2)$ for $N(0,1)$ data.

suggested when the dispersion is known with the optimal value of p turning out to be 1.2 in most cases. The only situation in which the optimal value for p is not 1.2 (or close to it) is when there is no contamination in the data in which case the theoretical optimal p is 0.5. The graph of the theoretical $AMSE$ function for these data is shown in Figure 6.4 (page 164) and, like the one parameter case, has a global minimum at $p = 0.5$ and a local one at $p = 1.2$. As one would expect, the theoretical minimum $AMSE$'s attainable are larger than in the one parameter case, particularly when the degree of contamination is 20%. With p chosen optimally the method should be highly efficient at the model (the theoretical minimum $AMSE$ for $N(0,1)$ data is just 0.004) and perform reasonably well when there is up to 10% contamination in the data. However, these results suggest that despite choosing p optimally the method will perform badly for data which has a high degree of contamination at one point.

The theoretical optimal value of p and corresponding $TAMSE$ for the $Gamma(4,\theta)$ model were obtained for the $f_\theta = g$ situation only and are 0.5 and 0.003 respectively. When $f_\theta \neq g$ singularities occur in J when $p < 2$ and in K when $p < 1$, so as an alternative, the influence function for this model was calculated and the optimal value for p determined by inspection. This graph, shown in Figure 6.5 (p.166), suggests that maximum robustness will be achieved when p is around 1.2 or 1.3. The curves for

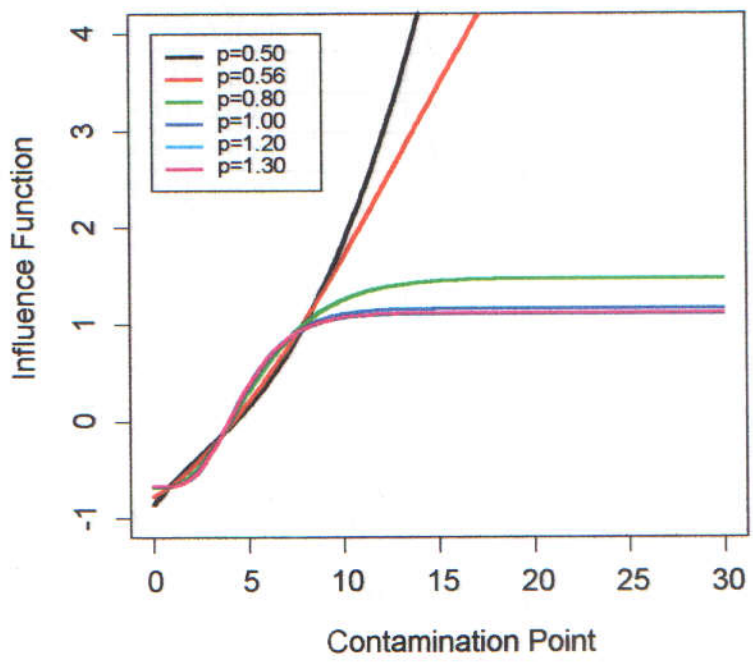

Figure 6.5: Influence function for the OH estimator when $f_\theta = Gamma(4, \theta)$.

Table 6.2: Theoretical optimal values of p for the OH estimator when $f_\theta = N(\theta, \sigma^2)$.

Distribution of data	**Theoretical Optimal p**	**Theoretical Minimum AMSE** σ
$\phi(x)$	0.5	0.004
$0.95\phi(x) + 0.05\Delta_{10}(x)$	1.2	0.028
$0.9\phi(x) + 0.1\Delta_{10}(x)$	1.2	0.136
$0.8\phi(x) + 0.2\Delta_{10}(x)$	1.2	2.398
$0.95\phi(x) + 0.05\Delta_5(x)$	1.2	0.028
$0.9\phi(x) + 0.1\Delta_5(x)$	1.2	0.136
$0.8\phi(x) + 0.2\Delta_5(x)$	1.2	1.466
$0.9\phi(x) + 0.1\Delta_{-10}(x)$	1.2	0.136
$0.9\phi(x) + 0.1\Delta_3(x)$	1.2	0.114
$0.8\phi(x) + 0.2\Delta_3(x)$	1.3	0.608
$0.9\phi(x) + 0.1\Delta_4(x)$	1.2	0.133
$0.8\phi(x) + 0.2\Delta_4(x)$	1.2	1.063
t_2	1.2	0.173
t_3	1.2	0.079
t_4	1.1	0.052

$p = 1.2$ and $p = 1.3$ are indistinguishable by eye but further investigation reveals $p = 1.2$ to be the optimum value. Unfortunately, it is not possible to estimate what the $TAMSE$ for the OH estimate of θ will be at that point.

Comparison between the theoretical optimal p's and those obtained by applying the method to simulated data will be discussed in Section 6.6.3 (p.176).

6.5 Simulations

As previously explained, the performance of this method can be assessed by applying the method to simulated data. Robust methods are often recommended in situations where the data is thought to be heavy-tailed or to contain outliers so data sets with these undesirable properties were created by taking random samples from several symmetric and asymmetric distributions. To enable the efficiency of the method to be assessed, uncontaminated random samples were also taken.

Since the new method of finding the optimal value of p could not be applied with the $Gamma(4, \theta)$ model parameter estimates were obtained using the OH method with p fixed at 1, 1.2 and 1.3 in turn. The value of p which leads to the smallest $MSE's$ could then be identified and (hopefully) confirm that the optimal value of p for robustness with this model is also 1.2.

6.5.1 Generation of the simulated data sets

Details of how the simulated data was generated is given in the BHHJ chapter, Section 4.5.1 (p.92).

6.5.2 Estimation of θ_p

The OH estimates, $\widehat{\theta}_p$, were obtained by solving the estimating equation (6.3) numerically. This estimating equation has only one root so the root-finding procedure can be applied less cautiously than with the density based methods because the global minimiser will be found irrespective of the choice of initial values and/or width of the search area. Further details given in Section 3.3.3 (p.72).

6.5.3 Estimation of θ_*

The sample median was used to estimate the unknown true location of the data, θ_*.

6.5.4 Minimising the $AMSE$ function

The asymptotic mean squared error function for criterion function estimators is not necessarily unimodal, as illustrated by Figure 6.2 (page 161), and cannot be minimised analytically. This means that numerical optimisation procedures must be used carefully to ensure that the global minimum is found rather than a local one. To avoid this problem the function was evaluated over a grid of 46 points between $p = 0.5$ and $p = 5$ in steps of 0.1. This, in effect, restricts the minimiser to one of 46 values when it should

really be continuous but has the benefit of greatly reducing the amount of computation involved whilst enabling the minimum to be found with sufficient accuracy.

6.5.5 Other robust methods

Once again several simpler methods were also applied to the simulated data. The *median/mad* combination was used to estimate location and dispersion respectively, the Cramér-von Mises method (which is OH with $p = 1$) and the OH method with $p = 1.2$.

6.6 Results and Discussion

To allow comparisons to be made easily between the performance of the various methods applied to the simulated data, the mean squared errors (MSE) of the estimates obtained for the contaminated and uncontaminated $N(0, 1)$ data are summarised in Tables 6.3-6.6 on pages 171-173. Table 6.3 gives the MSE's for the location parameter assuming that the variance of the data is known and Tables 6.4 and 6.5 give the results for the location and dispersion respectively when both parameters are unknown. Table 6.6 gives the overall MSE's for this latter case and are simply the sum of the individual MSE's for each parameter.

Table 6.3: Simulation results: Mean squared errors of the OH estimates of θ when $f_\theta = N(\theta, 1)$.

	Other estimators				Minimising AMSE
Distribution	mean	median	CVM	$p = 1.2$	$\theta_* = median$
$\phi(x)$	0.008	0.013	0.009	0.009	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.331	0.028	0.025	0.025	0.027
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.053	0.051	0.054
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	0.231	0.222	0.222
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.084	0.023	0.023	0.023	0.024
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.068	0.066	0.068
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.213	0.204	0.205
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.111	0.043	0.057	0.056	0.058
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.048	0.047	0.050
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.223	0.216	0.224
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.201	0.052	0.068	0.066	0.069
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.672	0.123	0.231	0.222	0.223
$t_2(x)$	0.093	0.027	0.025	0.025	0.025
$t_3(x)$	0.023	0.021	0.014	0.014	0.014
$t_4(x)$	0.016	0.018	0.012	0.012	0.012

The results obtained for the *Gamma* data with $f_\theta = Gamma(4, \theta)$ are given in Table 6.7 (p.173).

6.6.1 One parameter case

The second column of Table 6.3 contains the MSE's in using maximum likelihood to estimate location from samples of contaminated and uncontaminated $N(0, 1)$ data. The new method (column 6) is slightly less efficient than maximum likelihood, the known optimal method for the $f_\theta = g$ case, but highly efficient nonetheless. For all the other types of data the new method produces estimates with much smaller MSE's than maximum like-

Table 6.4: Simulation results: Mean squared errors of the OH estimates of θ when $f_\theta = N(\theta, \sigma^2)$.

	Other estimators				Minimising AMSE
Distribution	mean	median	CVM	$p = 1.2$	$\theta_* = median$
$\phi(x)$	0.008	0.013	0.009	0.009	0.009
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.071	0.066	0.067
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	0.830	0.747	0.832
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.094	0.088	0.090
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.507	0.472	0.475
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.059	0.057	0.059
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.291	0.287	0.294
$t_2(x)$	0.093	0.027	0.026	0.026	0.026
$t_3(x)$	0.023	0.021	0.014	0.014	0.014
$t_4(x)$	0.016	0.018	0.012	0.012	0.012

Table 6.5: Simulation results: Mean squared errors of the OH estimates of σ when $f_\theta = N(\theta, \sigma^2)$.

	Other estimators				Minimising AMSE
Distribution	s	mad	CVM	$p = 1.2$	$\sigma_* = mad$
$\phi(x)$	0.005	0.015	0.006	0.007	0.007
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	4.426	0.034	0.115	0.103	0.104
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	9.541	0.196	2.179	2.034	2.034
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.616	0.051	0.146	0.134	0.136
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	1.388	0.168	1.041	1.006	1.003
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.107	0.046	0.090	0.087	0.088
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.243	0.168	0.317	0.320	0.315
$t_2(x)$	3.910	0.084	0.201	0.189	0.190
$t_3(x)$	0.546	0.038	0.073	0.070	0.071
$t_4(x)$	0.178	0.031	0.049	0.047	0.048

Table 6.6: Simulation results: Combined mean squared errors of the OH estimates of θ and σ when $f_\theta = N(\theta, \sigma^2)$.

	Other estimators				Minimising AMSE
Distribution	mean s	median mad	CVM CVM	$p = 1.2$ $p = 1.2$	$\theta_* = median$ $\sigma_* = mad$
$\phi(x)$	0.013	0.028	0.015	0.016	0.016
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	5.439	0.077	0.186	0.169	0.170
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	13.635	0.322	3.009	2.781	2.866
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.921	0.105	0.240	0.222	0.206
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	2.368	0.275	1.548	1.478	1.478
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.204	0.079	0.149	0.144	0.147
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.615	0.296	0.611	0.610	0.609
$t_2(x)$	4.003	0.110	0.227	0.215	0.216
$t_3(x)$	0.569	0.059	0.087	0.084	0.085
$t_4(x)$	0.194	0.049	0.061	0.059	0.060

Table 6.7: Simulation results: Mean squared errors of the OH estimates of θ when $f_\theta = Gamma(4, \theta)$.

	Other estimators				
Distribution	*mle*	*mad*	*CVM*	$p = 1.2$	$p = 1.3$
$Gamma(4, 1)$	0.003	0.018	0.003	0.003	0.003
$0.9Gamma(4, 1) + 0.9Gamma(4, 8)$	0.016	0.023	0.016	0.016	0.014
$0.9Gamma(4, 1) + 0.9Gamma(4, 16)$	0.099	0.024	0.029	0.027	0.028

lihood, thus confirming its robustness. Comparison to the results obtained when using other robust estimation methods leads to the conclusion that this new method is not offering much improvement over its less computationally intense rivals. The new method is more efficient at the model than the median and offers slightly smaller $MSE's$ when the contamination is symmetric but these small benefits are outweighed by the methods' poor performance for asymmetrically contaminated data. The results for Cramer-von-Mises estimation (the $p = 1$ case of OH) and $p = 1.2$ are very similar to those obtained by the new method but there are consistent differences in performance which are worthy of mention. Setting $p = 1.2$, which is the theoretical optimal value for most of the contaminated data, leads to the smallest $MSE's$ for all the data sets despite the fact that when there is no contamination the optimal value of p is 0.57. The MSE's obtained by minimising a data-based estimate of the $AMSE$ function are generally just a little larger than, and in some cases equal to, those obtained when $p = 1.2$ which demonstrates that the new method either finds, or comes very close to finding, the optimum value of p in the majority of cases.

The results for the *Gamma* data and model (Table 6.7) also show that there is little to choose between the three OH based methods. As one might expect from the influence function (Figure 6.5, p.166) neither $p = 1.2$ nor $p = 1.3$ consistently leads to the smallest MSE's but both offer improvements over

mle, *mad* and CVM. Surprisingly, choosing the optimal value of p for robustness has little impact on efficiency.

6.6.2 Two parameter case

The simulation results for location and dispersion for the model $f_\theta = N(\theta, \sigma^2)$ are shown separately in Tables 6.4 and 6.5 respectively. In both cases the new method leads to much smaller mean squared errors than maximum likelihood for contaminated data and slightly less efficiency at the model. Using the median and median absolute deviation (mad) as estimators generally offers the most robustness but this is paid for in terms of lost efficiency at the model. When estimating the location parameter, there is little to choose between the three OH based methods in terms of either robustness or efficiency. However, setting $p = 1.2$ for all data does lead to consistently smaller MSE's than the new method or CVM. The results for the dispersion parameter are very similar to those for location. Minimising the $AMSE$ often leads to smaller MSE's than those obtained using CVM and is equally efficient. Once again, for contaminated data setting $p = 1.2$ leads to smaller MSE's than either of the OH based rivals. The combined mean squared errors for both parameters, obtained by adding the figures in Tables 6.4 and 6.5, are given in Table 6.6. These confirm that the generally good performance of the median and mad for contaminated data is over-

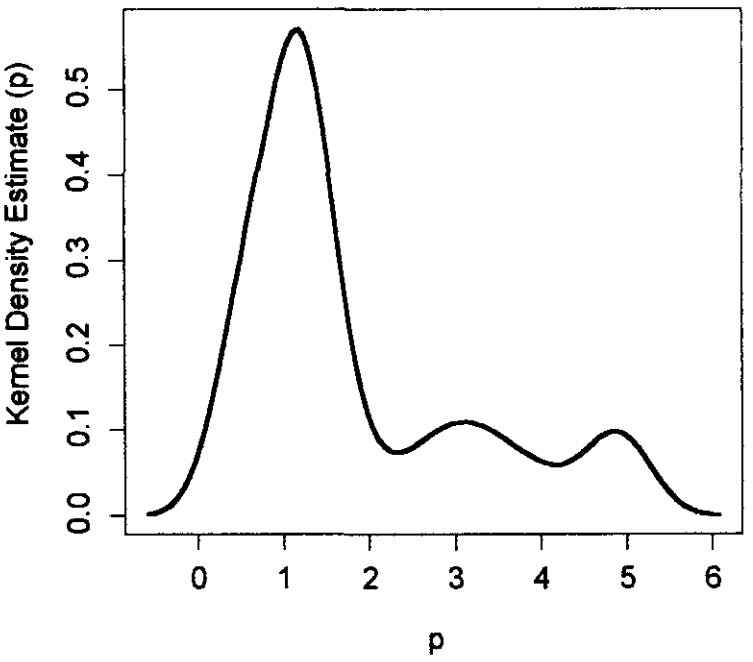

Figure 6.6: Kernel density estimate of the distribution of $\hat{p}$ when $f_\theta = N(\theta, 1)$ for $N(0, 1)$ data with no contamination.

shadowed by its huge losses in efficiency at the model. The optimal value for p is clearly 1.2 and it is very encouraging that the new method is able to locate this optimum is the majority of cases.

6.6.3 Comparison to theoretical results

Table 6.8 (p.182) gives details of the data-based and theoretical optimal values of p for the model $f_\theta = N(\theta, 1)$. For contaminated data the optimal

p's suggested by the data compare very well to those predicted in theory, but unfortunately the data-based method was not able to identify the global minimum for uncontaminated data at $p = 0.57$ nor the local minimum at $p = 2.7$. The kernel density estimate of the distribution of $\widehat{p}$ (Figure 6.6, p.176) shows that the estimated values of p are clustered around $p = 1.1$ and range over the entire range of possible values. One explanation for the data-based method's failure to find the minima in this case can be found by considering the theoretical $AMSE$ curve for $N(0,1)$ data (Figure 6.2). The magnitude of the theoretical $AMSE$ at its minimum is just 0.0005 less than its value at its highest point and so it is not surprising that this data-based estimate is not able to accurately reflect such fine detail.

In the two parameter case (Table 6.9, p.183) the data-based results are again close to the theoretical ones with the optimal value of p suggested by the data being 1.2 for most types of data. The theoretical $AMSE$ function, once again, has both local and global minima (Figure 6.4) for uncontaminated $N(0,1)$ data which are at $p = 1.2$ and $p = 0.5$ respectively. The data-based method does manage to locate the local minimum ($p = 1.2$) in many cases, as indicated by the kernel density estimate of $\widehat{p}$ which is shown in Figure 6.7 (p.178), with the average $\widehat{p}$ being recommended being 1.47. As in the one parameter case however, the theoretical $AMSE$ function is fairly flat with the global minimum being 0.004 at $p = 0.5$ and the local minimum

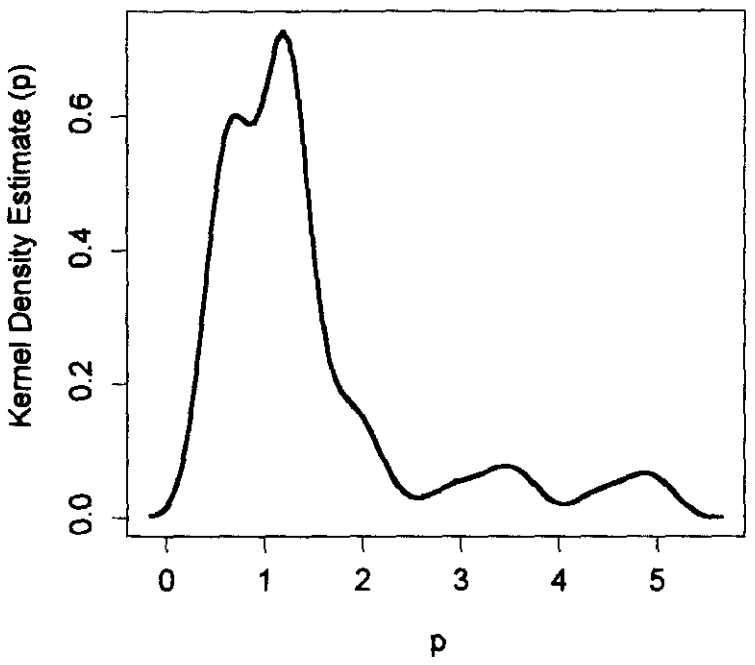

Figure 6.7: Kernel density estimate of the distribution of $\hat{p}$ when $f_\theta = N(\theta, \sigma^2)$ for $N(0,1)$ data with no contamination.

0.008 at 1.2 and the range of values suggested is fairly wide. Clearly, if the data-based method could be modified to enable the global minimum to be located, rather than the local one, efficiency could be greatly improved.

The fact that it has been possible to locate one of the minima but not the other may be due to the value of p at the respective minima and the effect that this has on our ability to estimate the $AMSE$. Notice, from equations (6.5) to (6.6), that the variance includes G^{p-1} terms so that when $p < 1$ we have G to a negative power. The true distribution function G is estimated by the empirical distribution function F_n which is generally a very satisfactory estimator of G, as is F_n^{p-1} of G^{p-1} for $p > 1$. When $p < 1$, however, the problem of how to estimate G^{p-1} is far less straightforward. Firstly, the empirical distribution function equals 0 at all points less than or equal to the minimum of the data set, which means that for $p < 1$ the left tail of F_n^{p-1} is infinity leading to computational difficulties. This problem can be solved, in part, by smoothing the function slightly so that it goes close to, but never equals, 0. This is done by taking the value of F_n to be the mid-point of the height of the step so that for a sample of size 100 it would have 0.005 as its lowest possible value and 0.995 as its largest. This modified empirical distribution function now behaves more like the true density distribution which also tends towards 0 and 1 but never equals them. Problems remain however, because these functions tend towards 0 and 1 at different rates and

the small numerical differences between them are magnified by the inverse powers. For example, taking $p = 0.5$, $G = \Phi$ and calculating F_n from a random sample of $N(0,1)$ data leads to the data-based estimate of G^{p-1} being 14.142 as compared to its true value of 1867.77. Similar problems are encountered when estimating $(1-G)^{p-1}$ with $(1-F_n)^{p-1}$.

The problems encountered in estimating the variance when $p < 1$ may be solved by further modifications to F_n or by taking a different approach entirely, such as bootstrap estimation. It might then be possible to improve efficiency by obtaining better suggestions for the optimal value of p given data which is not contaminated. There is little to be gained by such modifications when the data is contaminated, however, because the theoretical optimal value of p is greater than 1.

6.7 Conclusions

Although the parameter estimates obtained using the new method are indeed robust, and compare very favourably to those obtained via maximum likelihood, their performance is slightly disappointing nevertheless. The theoretical asymptotic mean squared error attainable for data with 10% contamination is typically about 0.13 and quickly rises thereafter. When the variance is known and there is 20% contamination at 10 in $N(0,1)$

data the minimum $AMSE$ obtainable in theory is 0.222 which compares very badly with the corresponding figure for the optimal $BHHJ$ method of 0.013. Furthermore, the OH estimate at $p = 1$ is the same as the Cramér-von Mises estimator which for the same data has a slightly larger theoretical $AMSE$ of 0.230. Even with p set at the optimum value of $p = 1.2$ the OH estimate is still only a little more robust than Cramér-von Mises and much less so than $BHHJ$. It should be emphasised, however, that this lack of robustness is inherent in the OH estimator and not the result of the new method selecting sub-optimal values for p.

The main goal of this research was, however, to determine how to choose a value for p when the true distribution of the data is unknown and in this respect the new method has been very successful. Using the $AMSE$ function as a joint measure of robustness and efficiency led to the surprising discovery that, for the data sets considered here at least, the optimal value for p is 1.2. As one would expect, the data-based selection procedure leads to slightly larger MSE's than one would obtain by simply setting $p = 1.2$ but it performs very well nonetheless and is a reliable alternative for optimising the performance of the OH estimator. The problems experienced in estimating the variance when $p < 1$ may have led tc losses in efficiency but improving the new method's performance in this respect would not be enough to compensate for its lack of robustness. It therefore seems sensible

Table 6.8: Theoretical and data-based mean squared errors of the OH estimator when $f_\theta = N(\theta, 1)$.

Distribution	Average Optimal p		Mean Squared Error	
	Theoretical Results	Data-based Results	Theoretical Results	Data-based Results
$\phi(x)$	0.57	1.81	0.010	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	1.20	1.24	0.021	0.027
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.20	1.19	0.053	0.054
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	1.20	1.20	0.222	0.222
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	1.20	1.26	0.021	0.024
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	1.20	1.26	0.053	0.068
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	1.20	1.21	0.222	0.205
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.20	1.20	0.053	0.058
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	1.20	1.17	0.053	0.050
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	1.20	1.14	0.218	0.224
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	1.20	1.29	0.053	0.069
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	1.20	1.20	0.221	0.226
t_2	1.20	1.26	0.018	0.025
t_3	1.20	1.22	0.016	0.014
t_4	1.20	1.34	0.014	0.012

to constrain p to be no less than 1 and accept that for uncontaminated $N(0, 1)$ data it may not be possible to locate the either minima. Although the resulting loss in efficiency is high in percentage terms the mean squared error of the estimates would still be very small.

Table 6.9: Theoretical and data-based mean squared errors of the OH estimator when $f_\theta = N(\theta, \sigma^2)$.

Distribution	Average Optimal p		Mean Squared Error	
	Theoretical Results	Data-based Results	Theoretical Results	Data-based Results
$\phi(x)$	0.5	1.47	0.004	0.016
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.2	1.23	0.136	0.170
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	1.2	1.25	2.398	2.866
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	1.2	1.22	0.136	0.206
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	1.2	1.27	1.466	1.478
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	1.2	1.39	0.114	0.147
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	1.3	1.44	0.608	0.609
t_2	1.2	1.19	0.173	0.216
t_3	1.2	1.18	0.079	0.085
t_4	1.1	1.23	0.052	0.060

Chapter 7

Comparison of the BHHJ, OH and Hellinger distance estimators

The results presented in the previous three chapters clearly demonstrate that all of these methods can produce estimates which are both robust and efficient. However, the methods were neither equally robust nor equally efficient so the purpose of this chapter is bring the results together and enable comparisons to be made. There are two key aspects to consider; firstly the features of the distance measures themselves and secondly the effectiveness of the new method in suggesting suitable values for α, p or h. For this reason two sets of results for each distance measure are presented in Tables

7.1 to 7.3 on pages 189 to 191. Table 7.1 shows the results obtained when estimating location with the model $f_\theta = N(\theta, 1)$. The results obtained with the $Gamma(4, \theta)$ model are summarised in Table 7.2 and those relating to the model $f_\theta = N(\theta, \sigma^2)$ in Table 7.3. The first set of results for each shows the mean squared error (MSE) obtained by minimising a data-based estimate of the asymptotic mean squared error function ($AMSE$) as described in detail in Chapters 4 to 6. The merits of using this new method are not clear because choosing values for α, p and h in some simpler way often leads to significant improvements in performance over maximum likelihood. The second set of results for each method therefore relate to the best simple alternative to minimising the $AMSE$ which, for the minimum density power divergence (BHHJ), Öztürk and Hettmansperger's criterion function (OH) and Hellinger distance (HD), turned out to be $\alpha = 1$, $p = 1.2$ and h =Sheather-Jones bandwidth respectively. The results using maximum likelihood and the *median/mad* approach are also shown.

7.1 Simulation results - estimating location only

With the model $f_\theta = N(\theta, 1)$ (Table 7.1, p.189), using OH or HD in either form leads to 89% efficiency. The performance of BHHJ with $\alpha = 1$ (i.e.

using the L_2-distance) is very poor (just 57% efficient) but choosing a value for α from the data via the $AMSE$ function does increase this to the more acceptable level of 89%. Comparing the robustness of the simplest versions of the methods first (i.e. $\alpha = 1$, $p = 1.2$ and h =Sheather-Jones bandwidth) identifies OH as being the least effective method, often leading to larger $MSE's$ than the median. The HD method is very robust to contamination at 10 but copes much less well than BHHJ when there is contamination at 3 or 4. Thus the MSE for the HD method when there is 20% contamination in the data at 3 is 0.113 as compared to 0.028 for BHHJ. Minimising the $AMSE$ function to suggest values for α, p and h has a mixed effect on the performance of these methods. For the OH method choosing p in this way has no effect on efficiency but leads to decreased robustness. In the case of the HD method for many types of data there is little justification for the additional computation involved in minimising the $AMSE$ function because using the Sheather-Jones bandwidth selection procedure to choose h appears to be equally effective. However, the new method does lead to smaller MSE's when there is contamination at 3 or 4 and so should not be dismissed out of hand. Applying the new data-based selection procedure to the BHHJ method led to significant improvements in its performance with small increases in robustness being supplemented by greatly increased efficiency.

7.2 Simulation results - estimating dispersion only

When estimating dispersion only with the $Gamma(4, \theta)$ model (Table 7.2, p.190), both OH and BHHJ could be fully efficient. The HD method performed much less well than expected with just 33% efficiency. The simpler alternatives to the new method generally led to increased robustness over the maximum likelihood estimator and *mad*. The BHHJ method with $\alpha = 1$ is very inefficient however but minimising the $AMSE$ to obtain an estimate of α greatly improves this. Unfortunately, this new method could not be applied for OH with the $Gamma(4, \theta)$ model and did not make the HD estimator perform any better than it does when simply using the Sheather-Jones bandwidth.

7.3 Simulation results - estimating location and dispersion

The results when both parameters are to be estimated (Table 7.3, p.191) echo those seen in the one parameter case. (Note that some entries relating to the OH method are blank because the full set of simulations were not carried out. It was the least promising and most computionally in-

tense of the three methods considered and so was not studied in as much detail as the others.) OH is the most efficient (81%) of the three simpler methods but is less robust than using the *median/mad* combination. BHHJ ($\alpha = 1$) is highly inefficient (54%) but performs generally better than *median/mad* except when there is contamination at 3. The HD method (h =Sheather-Jones) performs extremely well with contamination at 10 but quite badly when the data is contaminated at 3 or 4 or heavy tailed. The results obtained by minimising the $AMSE$ function show once again that OH performs worse when p is chosen from the data rather than fixed at 1.2. Selecting the bandwidth h for HD using this method generally leads to smaller $MSE's$ than using the Sheather-Jones bandwidth but the method is still unable to cope with the smallest degree of contamination at 3 or 4, 20% contamination at 5 or heavy tailed data and is just 76% efficient. Overall, BHHJ offers the lowest $MSE's$ for symmetrically or asymmetrically contaminated data but has the disadvantage of being fairly inefficient at the model (72%).

Table 7.1: Mean squared errors of the BHHJ, OH and HD estimates of θ when $f_\theta = N(\theta, 1)$.

			BHHJ		OH		HD	
Distribution	**mean**	**median**	**Minimising AMSE**	$\alpha = 1$	**Minimising AMSE**	**p = 1.2**	**Minimising AMSE**	h =S-J **bandwidth**
$\phi(x)$	0.008	0.013	0.009	0.014	0.009	0.009	0.009	0.009
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	0.331	0.028	0.014	0.019	0.027	0.025	0.013	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	1.012	0.043	0.012	0.019	0.054	0.051	0.010	0.011
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	4.094	0.126	0.015	0.023	0.222	0.222	0.015	0.015
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.084	0.023	0.013	0.017	0.024	0.023	0.012	0.012
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.305	0.054	0.015	0.019	0.068	0.066	0.014	0.014
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	0.980	0.107	0.014	0.019	0.205	0.204	0.013	0.013
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	1.111	0.043	0.012	0.017	0.058	0.056	0.012	0.011
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.097	0.033	0.043	0.016	0.050	0.047	0.025	0.034
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.372	0.129	0.034	0.028	0.224	0.216	0.089	0.113
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.201	0.052	0.028	0.022	0.069	0.066	0.016	0.020
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	0.672	0.123	0.018	0.021	0.223	0.222	0.017	0.023
$t_2(x)$	0.093	0.027	0.025	0.031	0.025	0.026	0.024	0.024
$t_3(x)$	0.023	0.021	0.015	0.023	0.014	0.014	0.015	0.014
$t_4(x)$	0.016	0.018	0.014	0.017	0.012	0.012	0.013	0.013

Table 7.2: Mean squared errors of the BHHJ, OH and HD estimates of θ when $f_\theta = Gamma(4, \theta)$.

			BHHJ		OH	HD	
Distribution	**mle**	**mad**	**Minimising AMSE**	$\alpha = 1$	**p = 1.2**	**Minimising AMSE**	**h =S-J bandwidth**
$Gamma(4,1)$	0.003	0.018	0.003	0.034	0.003	0.009	0.009
$0.9Gamma(4,1) + 0.1Gamma(8,1)$	0.016	0.023	0.013	0.011	0.016	0.012	0.012
$0.9Gamma(4,1) + 0.1Gamma(16,1)$	0.099	0.024	0.006	0.006	0.027	0.010	0.010

Table 7.3: Combined mean squared errors of BHHJ, OH and HD estimates of θ and σ when $f_\theta = N(\theta, \sigma^2)$.

			BHHJ		OH		HD	
Distribution	**mean s**	**median mad**	**Minimising AMSE**	$\alpha = 1$	**Minimising AMSE**	**p = 1.2**	**Minimising AMSE**	h =S-J bandwidth
$\phi(x)$	0.013	0.028	0.018	0.024	0.016	0.016	0.016	0.017
$0.95 \times \phi(x) + 0.05 \times \Delta_{10}(x)$	2.409	0.054	0.030	0.033	-	-	0.022	0.024
$0.9 \times \phi(x) + 0.1 \times \Delta_{10}(x)$	5.439	0.077	0.024	0.030	0.170	0.169	0.017	0.018
$0.8 \times \phi(x) + 0.2 \times \Delta_{10}(x)$	13.635	0.322	0.060	0.063	2.866	2.781	0.024	0.026
$0.95 \times \phi(x) + 0.05 \times \Delta_5(x)$	0.308	0.039	0.025	0.025	-	-	0.024	0.032
$0.9 \times \phi(x) + 0.1 \times \Delta_5(x)$	0.921	0.105	0.058	0.037	0.206	0.222	0.041	0.053
$0.8 \times \phi(x) + 0.2 \times \Delta_5(x)$	2.368	0.275	0.089	0.053	1.478	1.478	0.293	0.294
$0.9 \times \phi(x) + 0.1 \times \Delta_{-10}(x)$	5.809	0.082	0.034	0.032	-	-	0.018	0.018
$0.9 \times \phi(x) + 0.1 \times \Delta_3(x)$	0.204	0.079	0.052	0.088	0.147	0.144	0.167	0.175
$0.8 \times \phi(x) + 0.2 \times \Delta_3(x)$	0.615	0.296	0.308	0.308	0.609	0.610	0.429	0.437
$0.9 \times \phi(x) + 0.1 \times \Delta_4(x)$	0.480	0.089	0.038	0.038	-	-	0.140	0.143
$0.8 \times \phi(x) + 0.2 \times \Delta_4(x)$	1.393	0.307	0.102	0.102	-	-	0.732	0.667
$t_2(x)$	4.003	0.110	0.116	0.092	0.216	0.215	0.312	0.340
$t_3(x)$	0.569	0.059	0.051	0.049	0.085	0.084	0.151	0.167
$t_4(x)$	0.194	0.049	0.044	0.041	0.060	0.059	0.113	0.125

7.4 Key points

The key points arising from the above comparison are as follows and will be discussed in detail in the next few paragraphs. The main features of the OH method are that it's highly efficient but much less robust than using the *median* and *mad* to estimate location and dispersion respectively, particularly so when the percentage of contaminants is 20%. Furthermore, minimising a data based estimate of the $AMSE$ function leads to poorer performance than routinely setting $p = 1.2$. The HD method is highly efficient and very robust to large outliers but copes poorly with heavy tailed data and $N(0, 1)$ data with asymmetric contamination at points which are not unlikely under the model (such as at 3 or 4). In these cases the HD method performs better than maximum likelihood estimation but worse than using the *median/mad* combination. Once again, minimising the $AMSE$ function as a means of choosing what bandwidth to use does not appear to improve the performance of this method. The BHHJ method has the potential to provide estimators which are both robust and efficient providing α is chosen appropriately for the data. The optimal value of α does vary considerably for different types of data, however, so routinely using a specific value of α (such as 0.1 or 0.2) is not sufficient to guarantee the optimal performance of this method. Therefore the new approach of minimising the $AMSE$ function to find an appropriate value for α is very valuable indeed.

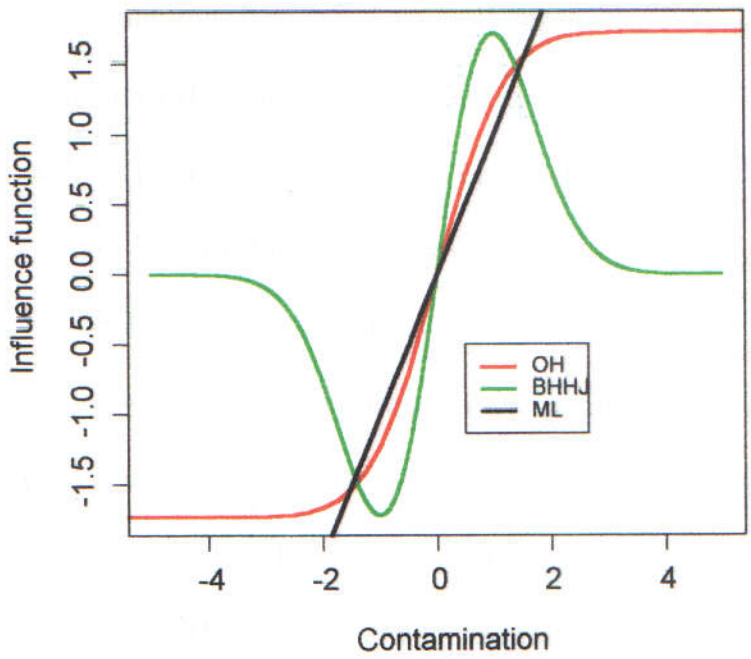

Figure 7.1: Influence functions for Öztürk and Hettmansperger's criterion function (OH), the minimum density power divergence (BHHJ) and maximum likelihood (ML)

7.5 Robustness

The behaviour of these three methods, in terms of their robustness, can be explained by considering how their estimators are affected by contamination in the data. The influence function is a useful first order approximation to asymptotic bias for OH and BHHJ and so, to facilitate easy comparisons, the influence functions for both these methods are shown together in Figure 7.1. The theoretical influence function for the Hellinger distance method is known to be a poor approximation to bias and is therefore not presented. Instead a data-based estimate of bias was obtained applying the Hellinger distance method to contaminated data. A random sample of 99 data points was taken from the $N(0,1)$ distribution and a single contamination point added over a range of 500 equally spaced values between + and - 15. Using a bandwidth of $h = 0.5$ the estimating equation was repeatedly solved to show how the HD estimate changes with the magnitude of the contamination point. This estimate of the bias function (shown in Figure 7.2) will vary with different realisations of data and may not have the same inverse symmetry that the theoretical version has when both g and f_θ are symmetric because it is based on the distribution of the sample, g_n, and not g. Nevertheless, each curve has the same general shape and suggests that a higher order approximation to asymptotic bias for Hellinger distance estimators, if one were to be obtained, would be similar to the influence function for BHHJ.

The influence function for OH tends to a specific, non-zero limit (approximately 1.6) as the magnitude of the contaminant increases. Thus a contamination point at 100 has no more influence than a contamination point at 10 which means the influence function is bounded and the method may be classed as robust. However, the magnitude of that influence is around 1.6 so increasing the percentage of contaminants at 10 from 5% to 10% and setting $p = 1.2$ leads to a similar doubling of the MSE from 0.025 to 0.051. It is clear therefore that this method will not perform well using a normal model for $N(0, 1)$ data with large outliers.

The influence function for BHHJ (with $p = 1$) redescends to zero for all contamination points greater than 3.5 and thus copes better with contamination at 4 than at 3. The position of the peak indicates that this method would be least robust to contamination with an absolute value close to 1.

The estimated bias curve for the HD method also redescends but falls much more slowly than the influence function for BHHJ and does not return to its baseline until the absolute value of the contamination point is greater than 5. This explains why the HD method is not robust to contamination at 3 or 4 and the peak at approximately 2.5 suggests that this method would cope better than BHHJ with contamination at 1.

Thus it seems reasonable to conclude that, as a result of their influence (or

bias) functions redescending to zero, the estimates obtained using BHHJ and HD are far more robust to large outliers than those obtained via OH. Furthermore, it is the rate at which the influence (or bias) functions of the two density based methods redescend which explains how well the two density based methods will cope with contamination at points which are likely under the model. Therefore it seems that in order for a method to be highly robust to large outliers its influence function must redescend and therefore lead to increased sensitivity to contamination at points which are plausible under the model.

7.6 Efficiency

A widely quoted advantage of the HD method is that it is asymptotically equivalent to maximum likelihood when $f_\theta = g$ and yet in these simulations the method was found to be slightly less efficient than its rival the OH method which does not claim full efficiency (76% vs. 81%). It may be that the bandwidth used was inappropriate and led to the sub-optimal performance of the method. It could also be the case that sample size or number of samples taken were too small for the asymptotic equivalence to be demonstrated. Since the poor small sample behaviour of the HD estimator has been reported elsewhere (Harris and Basu [12] developed the "Penalized

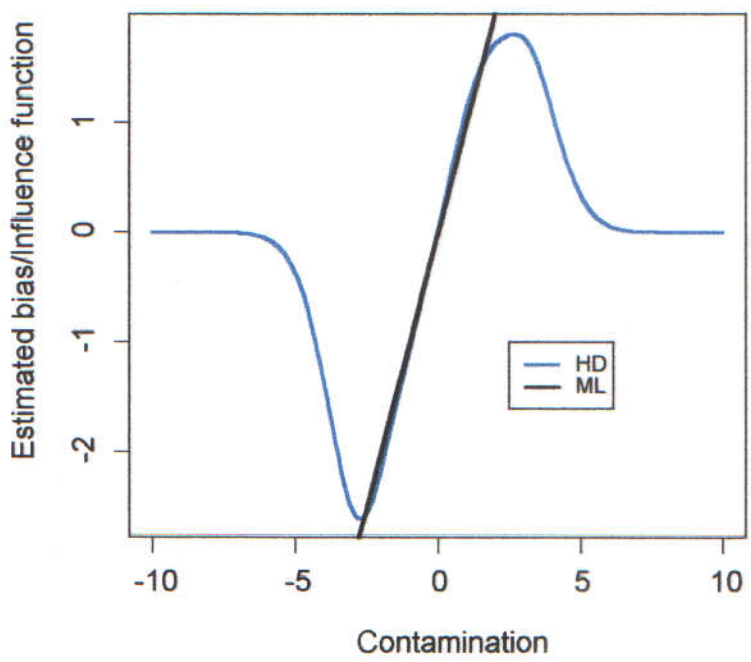

Figure 7.2: Estimated bias function for Hellinger distance estimators (HD) and influence function for maximum likelihood estimators (ML)

Hellinger Distance" estimator to address this problem) it seems likely that the samples used here were simply not large enough but nevertheless, two things are clear. Firstly, the HD method is not necessarily more efficient at estimating location or dispersion than many other robust methods and secondly for further progress to be made in understanding the interaction between bandwidth and efficiency for this method a second or higher order expression for asymptotic efficiency, as a function of h, must be obtained.

The BHHJ method differs from OH and HD in that it is not asymptotically equivalent to maximum likelihood when $f_\theta = g$ but equivalent to it when $\alpha = 0$. This means that when $f_\theta = g$ both OH and HD are generally quite efficient irrespective of the values placed on p and h whilst in contrast, the BHHJ is efficient for such data only when $\alpha = 0$. The choice of value for α is therefore vital to ensure the full efficiency of this method.

7.7 Optimising performance by minimising the $AMSE$ function

Although the same approach was used on all three of the methods studied (i.e. using the $AMSE$ as a combined measure of robustness and efficiency to optimise performance) the practical application of this was slightly different

in each case. For each distance measure the first task was to derive an expression for the $AMSE$ function for each of the three distance measures under consideration and this was used as the basis for deciding appropriate values for α, p and h. For BHHJ this function depended on α and the true density g (amongst other things) and so by replacing g by g_ε the relationship between the distribution of data and optimal value for α could be explored. This demonstrated that when estimating both location and dispersion the optimal value of α is 0 for uncontaminated $N(0,1)$ data and 1 for symmetric or asymmetric contamination. Replacing g_ε with the data $\widehat{g}$ gave a data based estimate of the $AMSE$ function which could then be minimised to suggest an optimal value for α. This seems to be a fairly effective way of improving the performance of BHHJ, as demonstrated by the increase in efficiency from 54% to a more acceptable 72%, but there is scope for further improvements in the performance of this new method, particularly at the model. One possible course of action would be to try to estimate the $AMSE$ more robustly in the hope that the value suggested for α would then be closer to the optimal.

In the case of OH, replacing g by g_ε and G by G_ε in the $AMSE$ function led to the surprising discovery that $p = 1.2$ would be optimal (or close to optimal) for all the types of data considered. For completeness, a data based estimate of the $AMSE$ function was obtained and minimised as with BHHJ

but, as one would expect, this data-based approach led to generally larger $MSE's$ than simply using the optimal $p = 1.2$. The fact that $p = 1.2$ was found to be optimal for the types of data studied here does not necessarily mean that this value would be optimal for every possible type of data but it does demonstrate the merits of assessing robustness and efficiency jointly in this way because the notion that $p = 1.2$ might be a sensible choice was not identified when these two properties were considered separately by Öztürk and Hettmansperger [23] in their paper.

The case with HD was slightly different again because the theoretical $AMSE$ function, obtained by replacing g by g_ε, did not depend on the bandwidth h at all. This meant that a data based estimate of the $AMSE$ function (which does depend on h) could be obtained and minimised as with the previous two methods but there is no means of knowing how far from optimal the suggested bandwidths are nor whether a single value or formula for the bandwidth might be sufficient to ensure the acceptable performance of this method. Furthermore, although applying the new method did lead to consistently smaller $MSE's$ it is not clear whether the improvement was so small because the Sheather-Jones bandwidth performs very well anyway, because the expression for the asymptotic variance term in the $AMSE$ was poor or because of the usual problems encountered in estimating a function from data. It is therefore imperative that further work is carried out

to obtain a reliable second or higher order approximation to bias so that the relationship between bandwidth, efficiency and robustness can be fully explored.

7.8 Computational issues

Each of the methods studied presented different computational challenges. The BHHJ method is the simplest to implement because there is no need to estimate the true density directly from data, however, multiple roots are possible when the degree of contamination is high so applying this method is not always straightforward nonetheless. The computer programs for the OH method took considerably longer to run than those for BHHJ because sums of order statistics are needed instead of averaging over the data. There are no multiple roots to contend with however and the empirical distribution function is simple to calculate. Applying the HD method requires significantly more computation than the other two methods because it involves kernel density estimates and numerical integration in addition to the numerical optimisation procedure which is required by all three methods. Furthermore, since the Hellinger distance estimating equation may also have multiple roots it is certainly the trickiest of the three methods to implement but it should be emphasised that none of these problems were insurmount-

able, for the two parameter problem at least, and the benefits of using this method greatly outweigh these disadvantages.

7.9 Some observations relating to density and distribution based estimators

A very useful benefit of using the $AMSE$ to jointly assess robustness and efficiency is that it makes it possible to compare the performance of several very different estimators within a common framework. Thus, having studied distance measures based on both density and distribution functions, it is now possible to make some general comments concerning the relative merits of these two classes of estimator. The following comments are based purely on intuition with no attempt being made to justify these suppositions with full mathematical rigour. However, they are interesting nonetheless and should, at the very least, provoke debate and will hopefully lead to further research.

An interesting feature of the OH method is that setting $p = 1.2$ for all types of data led to good all round performance. The fact that $p = 1.2$ is optimal for robustness can be easily determined by examination of the influence functions (Figure 3.14, p.68). These influence functions share the

same general shape for all values of p but are bounded at different points so the most robust method is clearly the one which gives large data points the least influence. The optimal value for p with regard to efficiency is $p = 0.57$ but despite being sub-optimal putting $p = 1.2$ lead to acceptable efficiency. This is because the influence functions for many values of p are superimposed within the region in which uncontaminated $N(0, 1)$ data is likely to fall (i.e. between + and - 3) so although a particular value of p may lead to greater efficiency there is little loss in choosing any $p > 0.5$. This hypothesis is supported by investigations regarding efficiency which were carried out by Öztürk and Hettmansperger when the distance measure was proposed and showed the method to be highly efficient at estimating location for many values of p.

Extending the notion that the shape of the influence function within the range + and -3 for $N(0, 1)$ data indicates efficiency to HD and BHHJ is also intriguing. If one considers the influence function for the maximum likelihood estimator to be the ideal for efficiency and compare it to the influence function for BHHJ (with $\alpha = 1$) and estimated bias function for HD the difference between the two methods in terms of efficiency can also be explained. Within the range of interest the estimated bias function for the Hellinger distance (Figure 7.2, page 197) is roughly the same as the influence function for the maximum likelihood estimator and so fully

efficient. In contrast, the influence function for BHHJ with $\alpha = 1$ (Figure 7.1, page 193) is very different to that of maximum likelihood in the region of interest, having peaks at ± 1, which may explain why this method is so inefficient.

In conclusion, my view is that if one were to construct the influence function of an ideal (i.e. simultaneously robust and efficient) estimation method it would have the following features. Firstly, to ensure robustness to large outliers it would redescend and be approximately zero for all values outside the likely range of the $N(0, 1)$ distribution. For maximum efficiency this influence function should closely resemble that of maximum likelihood within the range + and -3. The ideal estimation method would therefore be density based and very similar to the Hellinger distance.

It should be emphasised however that, as in the case of BHHJ, it is not strictly necessary to construct an estimator in this way to guarantee robustness and efficiency. The BHHJ method provides a bridge between maximum efficiency (with $\alpha = 0$) and extreme robustness ($\alpha = 1$) so providing there is a reliable way of choosing α appropriately its performance will rival (and possibly exceed) that of the Hellinger distance.

7.10 My preferred method

In comparing these methods the most important question one should ask is "which of these methods would you use in practice and why?". Despite there being so little to choose between the two density based methods in terms of performance $BHHJ$ would certainly be my first choice quite simply because it is the easiest to use and understand. Recall that the estimating equation is

$$0 = \int f_\theta^\alpha(x) \frac{df_\theta}{d\theta} dx - \frac{1}{n} \sum_{i=1}^{n} f_\theta^\alpha(X_i) u(X_i)$$

where X_i is the ith data point and u is the score statistic.

A useful feature of this method is that for given values of α, θ and σ the first term in this estimating equation is constant and the other is just the average of n functions which are centred at θ. Notice however, that when $f_\theta(x) = N(\theta, \sigma)$, x and θ are interchangeable so the second term could also be thought of as the average of n functions, each centred at the datapoint X_i. This means that for the normal family of models at least, it is quite easy to visualise the effect of each data point on the estimate of location which although of limited practical use, because the same does not apply to the estimate of dispersion nor indeed to location under many other models, does enable one to develop an intuitive feel for the method.

Another advantage of $BHHJ$ is that it is equivalent to maximum likelihood when $\alpha = 0$ irrespective of the distribution of the data. This means it is possible to calculate and plot a sequence of estimators with $\alpha = 0, 0.1, 0.2, ..., 1$ (as described in Section 4.6.4, p.108) which relates the most efficient estimate based on a data set to a highly robust one. The shape of this graph may be used to identify certain types of contamination in the data or situations in which the method has broken down, something which is not possible with either of the other two methods.

Chapter 8

Conclusions

- The asymptotic mean squared error is a useful joint measure of the efficiency and robustness of estimators.
- The asymptotic mean squared error function may be used to determine the optimum value for a tuning parameter in a distance measure.
- Asymptotic relative efficiency may be too harsh a measure of the performance of estimators at the model. Applying a fairly inefficient estimation method to uncontaminated data does not necessarily result in large mean squared errors.
- The optimum value for p in Öztürk and Hettmansperger's criterion function is 1.2.
- Minimising a data-based estimate of the asymptotic mean squared

error function leads to reasonable suggestions for α, p and h in the minimum density power divergence, Öztürk and Hettmansperger's criterion function and the Hellinger distance respectively.

- The derivation of some higher order approximation to bias (similar to the influence function) for Hellinger distance estimators would be extremely useful and informative.

- Of the three estimators considered here, my preference is for the minimum density power divergence (BHHJ).

Appendix A

Asymptotic properties of estimators

A.1 Derivation of the asymptotic mean and variance of the BHHJ estimator

The density power divergence is defined as

$$d_\alpha(g,f) = \int \left\{ f_\theta^{\alpha+1}(z) - (1+\frac{1}{\alpha})g(z)f_\theta^{\alpha}(z) + \frac{1}{\alpha}g^{\alpha+1}(z) \right\} dz$$

where $\theta = (\theta, \sigma, ..., \omega)^T$ is a vector of n parameters from the model density f_θ, g the true density and $0 \leq \alpha \leq 1$.

Differentiating with respect to θ gives the asymptotic estimating equation

$$\frac{dd_\alpha}{d\theta} = \int \left[(1+\alpha) f_\theta^\alpha(z) \frac{df_\theta}{d\theta} - (1+\frac{1}{\alpha}) g(z) \alpha f_\theta^{\alpha-1}(z) \frac{df_\theta}{d\theta}\right] dz$$

$$U(\theta) = \int f_\theta^{\alpha+1}(z) u_\theta(z) - \int f_\theta^\alpha(z) u_\theta(z)\, g(z)\, dz$$

which can be set equal to zero and solved to give the parameter θ_α.

The true density g is unknown and must therefore be estimated in some way. The second integral in the asymptotic estimating equation, $\int f_\theta^\alpha(z) u_\theta(z)\, g(z)\, dz$ is the expected value of $f_\theta^\alpha(z) u_\theta(z)$ over g and so can be estimated by taking the average of the function over the data. The estimating equation is therefore

$$U_n(\theta) = \int u_\theta(z)\, f_\theta^{\alpha+1}(z) dz - n^{-1} \sum_{i=1}^{n} f_\theta^\alpha(X_i) u_\theta(X_i)$$

The solution to $U_n(\theta) = 0$ is the BHHJ estimator, $\widehat{\theta}_\alpha$.

Subject to certain regularity conditions (see p.82 of Azzalini [2] for details), the asymptotic mean and variance of the BHHJ estimator $\left(\widehat{\theta}_\alpha\right)$ is obtained by expanding $U_n(\widehat{\theta}_\alpha)$ about $U_n(\theta_\alpha)$ using the Taylor series expansion

$$U_n(\hat{\theta}_\alpha) = U_n(\theta_\alpha) + (\widehat{\theta}_\alpha - \theta_\alpha)^T U_n'(\theta_\alpha) + \frac{1}{2}(\widehat{\theta}_\alpha - \theta_\alpha)^T U_n''(\theta_\alpha)(\widehat{\theta}_\alpha - \theta_\alpha) + \ldots$$

$$0 \simeq U_n(\theta_\alpha) + (\widehat{\theta}_\alpha - \theta_\alpha)^T U_n'(\theta_\alpha) + \ldots$$

$$(\widehat{\theta}_\alpha - \theta_\alpha) \simeq -\left[U_n'(\theta_\alpha)\right]^{-1} U_n(\theta_\alpha)$$

$$\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha) \simeq -\sqrt{n}\left[U_n'(\theta_\alpha)\right]^{-1} U_n(\theta_\alpha)$$

where $U_n'(\theta_\alpha)$ is the vector of first derivatives and $U_n''(\theta_\alpha)$ the matrix of second derivatives.

The denominator, $U_n'(\theta_\alpha)$, can be replaced with its expected value $E(U_n'(\theta_\alpha))$ (as a consequence of the "Law of Large Numbers") and the asymptotic mean and variance of $\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha)$ derived as follows

$$E(\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha)) \simeq -\sqrt{n}\left[E(U_n'(\theta_\alpha))\right]^{-1} E(U_n(\theta_\alpha))$$

$$var(\sqrt{n}(\widehat{\theta} - \theta)) \simeq n\left[E(U_n'(\theta_\alpha))\right]^{-1} var(U_n(\theta_\alpha))\left[E(U_n'(\theta_\alpha))\right]^{-1}$$

The expected value of $U_n(\theta)$ with respect to the true density is

$$\begin{aligned} E(U_n(\theta)) &= \int u_\theta(z)\, f_\theta^{\alpha+1}(z)dz - \frac{1}{n}\sum_{i=1}^{n} E\left\{f_\theta^{\alpha}(X_i)u_\theta(X_i)\right\} \\ &= \int u_\theta(z)\, f_\theta^{\alpha+1}(z)dz - \int f_\theta^{\alpha}(z)u_\theta(z)\, g(z)dz \end{aligned}$$

Thus, $E(U_n(\theta_\alpha)) \simeq U(\theta_\alpha) = 0$.

$$\begin{aligned} \frac{dU_n(\theta)}{d\theta} &= \frac{d}{d\theta}\left[\int u_\theta(z)\, f_\theta^{\alpha+1}(z)dz - n^{-1}\sum_{i=1}^{n} f_\theta^{\alpha}(X_i)u_\theta(X_i)\right] \\ &= \int (1+\alpha)u_\theta(z)\, u_\theta^T(z)\, f_\theta^{\alpha+1}(z)dz - \int f_\theta^{\alpha+1}(z)i_\theta(z)dz \\ &\quad - n^{-1}\sum_{i=1}^{n}\left[\alpha u_\theta(X_i)u_\theta^T(X_i)f_\theta^{\alpha}(X_i) - i_\theta(X_i)f_\theta^{\alpha}(X_i)\right] \end{aligned}$$

$$
\begin{aligned}
E(U_n'(\theta)) &= \int (1+\alpha) u_\theta(z)\, u_\theta^T(z)\, f_\theta^{\alpha+1}(z) dz - \int f_\theta^{\alpha+1}(z) i_\theta(z) dz \\
&\quad - n^{-1} \sum_{i=1}^{n} \alpha E\left[u_\theta(X_i)\, u_\theta^T(X_i) f_\theta^{\alpha}(X_i)\right] + n^{-1} \sum_{i=1}^{n} E\left[i_\theta(X_i) f_\theta^{\alpha}(X_i)\right] \\
&= \int (1+\alpha) u_\theta(z)\, u_\theta^T(z)\, f_\theta^{\alpha+1}(z) dz - \int f_\theta^{\alpha+1}(z) i_\theta(z) dz \\
&\quad - \alpha \int u_\theta(z)\, u_\theta^T(z)\, f_\theta^{\alpha}(z) g(z) dz + \int i_\theta(z) f_\theta^{\alpha}(z) g(z) dz \\
&= \int u_\theta(z)\, u_\theta^T(z)\, f_\theta^{\alpha+1}(z) dz \\
&\quad + \int \left(i_\theta(z) - \alpha u_\theta(z)\, u_\theta^T(z)\right) \left(g(z) - f_\theta(z)\right) f_\theta^{\alpha}(z) dz
\end{aligned}
$$

$$
\begin{aligned}
var(U_n(\theta)) &= \frac{1}{n^2} \sum_{i=1}^{n} var\left[f_\theta^{\alpha}(X_i) u_\theta(X_i)\right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} E\left[f_\theta^{2\alpha}(X_i) u_\theta(X_i)\, u_\theta^T(X_i)\right] \\
&\quad - \frac{1}{n^2} \sum_{i=1}^{n} \left[E\left[f_\theta^{\alpha}(X_i) u_\theta(X_i)\right]\right] \left[E\left[f_\theta^{\alpha}(X_i) u_\theta(X_i)\right]\right]^T \\
&= \frac{1}{n} \int f_\theta^{2\alpha}(z) u_\theta(z)\, u_\theta^T(z) g(z) dz \\
&\quad - \frac{1}{n} \left[\int f_\theta^{\alpha}(z) u_\theta(z)\, g(z)\right] \left[\int f_\theta^{\alpha}(z) u_\theta(z)\, g(z)\right]^T
\end{aligned}
$$

The central limit theorem applies to $U_n(\theta_\alpha)$ and so the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha)$ is Normal with

$$E(\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha)) \simeq 0$$

$$var(\sqrt{n}(\widehat{\theta}_\alpha - \theta_\alpha)) \simeq J^{-1} K J^{-1}$$

where

$$J = \int u_{\theta_\alpha}(z)\, u_{\theta_\alpha}^T(z)\, f_{\theta_\alpha}^{\alpha+1}(z)dz + \int \left(i_{\theta_\alpha}(z) - \alpha u_{\theta_\alpha}(z)\, u_{\theta_\alpha}^T(z)\right)\left(g(z) - f_{\theta_\alpha}(z)\right) f_{\theta_\alpha}^{\alpha}(z)dz$$

and

$$\begin{aligned} K &= \frac{1}{n}\int f_{\theta_\alpha}^{2\alpha}(z) u_{\theta_\alpha}(z)\, u_{\theta_\alpha}^T(z) g(z)dz \\ &\quad -\frac{1}{n}\left[\int f_{\theta_\alpha}^{\alpha}(z) u_{\theta_\alpha}(z)\, g(z)\right]\left[\int f_{\theta_\alpha}^{\alpha}(z) u_{\theta_\alpha}(z)\, g(z)\right]^T \end{aligned}$$

A.2 Derivation of the asymptotic mean and variance of the Hellinger distance estimator

The following derivation of the asymptotic mean and variance of Hellinger distance estimators is obtained by using Taylor series to approximate both the estimating equation and kernel density estimate $\widehat{g}_n^{\frac{1}{2}}$. The original aim was to obtain a second (or higher order) approximation to the asymptotic variance which would depend on the bandwidth h and therefore allow the asymptotic mean squared error ($AMSE$) function to be expressed as a function of h. This function could then be differentiated and solved to find the minimiser, the optimal bandwidth. Unfortunately, as explained in Section 5.2 (p.118), this asymptotic expansion does not appear to converge and so could not be used in this way. However, it should be emphasised

that the first term in this expression does agree to the formulae given by Beran [6] and was therefore used to obtain an expression for the $AMSE$ function. The full derivation (including the higher order terms which were later discarded) is presented here.

The Hellinger distance $H(\theta)$ is defined as

$$\begin{aligned} H(\theta) &= \int \left(f_\theta^{\frac{1}{2}}(x) - g^{\frac{1}{2}}(x) \right)^2 dx \\ &= 2 - 2 \int f_\theta^{\frac{1}{2}}(x)\, g^{\frac{1}{2}}(x)\, dx \end{aligned}$$

where $\theta = (\theta, \sigma, ..., \omega)^T$ is a vector of n parameters from the model density f_θ and g the true density of the data.

Differentiating this function with respect to θ gives the asymptotic estimating equation

$$\begin{aligned} \dot{H}(\theta) &= -\int g^{\frac{1}{2}}(x)\, f_\theta^{-\frac{1}{2}}(x)\, \dot{f}_\theta(x)\, dx \\ &= -\int s(x)\, g(x)\, dx \end{aligned}$$

where $\dot{f}_\theta = \frac{df_\theta}{d\theta}$ and $s = \frac{\dot{f}_\theta}{\sqrt{I_{\theta g}}}$.

Solving $\dot{H}(\theta) = 0$ yields the asymptotic Hellinger distance estimator, θ_1.

The true density of the data is not usually known and must therefore be estimated from the data. Replacing g with a kernel density estimate $\widehat{g}_n$

gives the estimating equation

$$\widehat{H}(\theta;h) = -\int \widehat{g}_n{}^{\frac{1}{2}}(x)\, f_\theta^{-\frac{1}{2}}(x)\, \dot{f}_\theta(x)\, dx$$

The HD estimates, denoted by $\widehat{\theta}_h$ to emphasise their dependence on the bandwidth used in the kernel density estimate, are obtained by solving $\widehat{H}(\theta;h) = 0$.

Subject to certain regularity conditions (see p.82 of Azzalini [2] for details), $\widehat{H}\left(\widehat{\theta}_h;h\right)$ is expanded about $\widehat{H}\left(\theta_1\right)$ using Taylor series to give the following approximation

$$\begin{aligned}
\widehat{H}(\hat{\theta}_h;h) &= \widehat{H}(\theta_1) + (\hat{\theta}_h - \theta_1)^T \widehat{H}'(\theta_1) + \frac{1}{2}(\hat{\theta}_h - \theta_1)^T \widehat{H}''(\theta_1)(\hat{\theta}_h - \theta_1) + ... \\
0 &= \widehat{H}(\theta_1) + (\hat{\theta}_h - \theta_1)^T \widehat{H}'(\theta_1) + ... \\
\sqrt{n}(\hat{\theta}_h - \theta_1) &\simeq -\sqrt{n}\left[\widehat{H}'(\theta_1)\right]^{-1} \widehat{H}(\theta_1)
\end{aligned}$$

where $\widehat{H}'(\theta_1)$ is the vector of first derivatives and $\widehat{H}''(\theta_1)$ the matrix of second derivatives.

The vector $\widehat{H}'(\theta_1)$ can be replaced with its expected value $E\left(\widehat{H}'(\theta_1)\right) = H'(\theta_1)$ ("Law of Large Numbers") and the asymptotic mean and variance of $\sqrt{n}\left(\widehat{\theta}_h - \theta_1\right)$ derived as follows

$$\begin{aligned}
E\left(\sqrt{n}(\hat{\theta}_h - \theta_1)\right) &\simeq -\sqrt{n}\left[H'(\theta_1)\right]^{-1} E\left(\widehat{H}(\theta_1)\right) \\
var\left(\sqrt{n}(\hat{\theta}_h - \theta_1)\right) &\simeq n\left[H'(\theta_1)\right]^{-1} var\left(\widehat{H}(\theta_1)\right)\left[H'(\theta_1)\right]^{-1}
\end{aligned}$$

Differentiating $H(\theta)$ with respect to θ gives

$$H'(\theta) = -\int \dot{s}(x)\, g(x)\, dx$$

where $\dot{f}_\theta = \frac{df_\theta}{d\theta}$, $s = \frac{\dot{f}_\theta}{\sqrt{f_\theta g}}$, $\ddot{f}_\theta = \frac{d^2 f_\theta}{d\theta^2}$ and $\dot{s} = \frac{ds}{d\theta}$.

Hence $H'(\theta_1) = H'(\theta)|_{\theta=\theta_1}$

Since $E(\widehat{H}(\theta;h)) = -\int E(\widehat{g}_n^{\frac{1}{2}}(x))\, f_\theta^{-\frac{1}{2}}(x)\, \dot{f}_\theta(x)\, dx$ we need an expression for $E(\widehat{g}_n^{\frac{1}{2}}(x))$. This is obtained by rewriting $\widehat{g}_n^{\frac{1}{2}}(x)$ as $\sqrt{g(x)}\left(1+\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^{\frac{1}{2}}$ and using the binomial expansion to express $\widehat{g}_n^{\frac{1}{2}}(x)$ as a power series. Thus, utilising the expression for the expected value of a kernel density estimate given in equation A.6 (p.226), gives

$$\begin{aligned}
E(\widehat{H}(\theta;h)) &= -\int \frac{\dot{f}_\theta(x)}{\sqrt{f_\theta}(x)} E\left(\sqrt{g(x)}\left[1+\frac{1}{2}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)\right.\right. \\
&\quad -\frac{1}{8}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^2+\frac{3}{48}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^3 \\
&\quad \left.\left.-\frac{15}{384}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^4+\ldots\right]\right) dx \\
&\simeq -\int \frac{\dot{f}_\theta(x)}{\sqrt{f_\theta(x)}}\sqrt{g(x)}\, dx - \frac{1}{2}\int \frac{\dot{f}_\theta(x)}{\sqrt{f_\theta(x)}}\frac{B(\widehat{g}_n(x))}{\sqrt{g(x)}} dx + \ldots \\
&\simeq -\int s(x)\, g(x)\, dx - \frac{h^2 K_2}{4}\int s(x)\, g''(x) dx + \ldots
\end{aligned}$$

where $K_2 = \int_{-\infty}^{\infty} K(u)u^2 du$, $\dot{f}_\theta = \frac{df_\theta}{d\theta}$, $s = \frac{\dot{f}_\theta}{\sqrt{f_\theta g}}$, $g''(x) = \frac{d^2 g(x)}{dx^2}$ and $B(\widehat{g}_n(x)) = E(\widehat{g}_n(x) - g(x))$.

The first term in this expansion is the asymptotic estimating equation $H(\theta)$ and so, assuming that the later terms are ignorable, $\widehat{\theta}_h$ converges to θ_1. In

general θ_1 is not the location of the true distribution g so $\widehat{\theta}_h$ is a biased estimate of the true location θ_*. However, when $f_\theta = g$ there is no bias because $E\left(\widehat{H}(\theta)\right) = 0$.

The variance of $\widehat{H}(\theta)$ is obtained from

$$E\left(\widehat{H}(\theta)\widehat{H}^T(\theta)\right) - \left[E\left(\widehat{H}(\theta)\right)\right]\left[E\left(\widehat{H}(\theta)\right)\right]^T$$

as follows

$$E\left(\widehat{H}(\theta)\widehat{H}^T(\theta)\right) = \int\int \left[\dot{f}_\theta(x) f_\theta^{-\frac{1}{2}}(x)\right]\left[\dot{f}_\theta(y) f_\theta^{-\frac{1}{2}}(x)\right]^T E\left(\sqrt{\widehat{g}_n(x)\widehat{g}_n(y)}\right) dy\, dx$$

Again making the substitution $\widehat{g}_n^{\frac{1}{2}}(x) = \sqrt{g(x)}\left(1 + \frac{\widehat{g}_n(x) - g(x)}{g(x)}\right)^{\frac{1}{2}}$ a power

series for $\sqrt{\widehat{g}_n(x)\widehat{g}_n(y)}$ can be obtained as follows

$$
\begin{aligned}
\sqrt{\widehat{g}_n(x)\widehat{g}_n(y)} &= \sqrt{g(x)g\,(y)}\left(1+\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)\right)^{\frac{1}{2}}\left(1+\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)\right)^{\frac{1}{2}} \\
&= \sqrt{g(x)g(y)}\left[1+\frac{1}{2}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)+\frac{1}{2}\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)\right. \\
&\quad -\frac{1}{8}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^2-\frac{1}{8}\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)^2 \\
&\quad +\frac{3}{48}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^3+\frac{3}{48}\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)^3 \\
&\quad +\frac{1}{4}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right) \\
&\quad -\frac{1}{16}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)^2 \\
&\quad -\frac{1}{16}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^2\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right) \\
&\quad -\frac{5}{128}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^4-\frac{5}{128}\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)^4 \\
&\quad +\frac{1}{32}\left(\frac{\widehat{g}(x)-g(x)}{g(x)}\right)\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)^3 \\
&\quad +\frac{1}{32}\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^3 \\
&\quad \left.+\frac{1}{64}\left(\frac{\widehat{g}_n(x)-g(x)}{g(x)}\right)^2\left(\frac{\widehat{g}_n(y)-g(y)}{g(y)}\right)^2+\ldots\right]
\end{aligned}
$$

Taking expectations over the true density g leads to

$$
\begin{aligned}
E\left(\sqrt{\widehat{g}_n(x)\widehat{g}_n(y)}\right) = \sqrt{g(x)}\sqrt{g(y)}\Bigg(1 &+ \frac{1}{2}\frac{B(\widehat{g}_n(x))}{g(x)} + \frac{1}{2}\frac{B(\widehat{g}_n(y))}{g(y)} \\
&- \frac{1}{8}\frac{1}{g(x)^2}\left(V(\widehat{g}_n) + B^2(\widehat{g}_n)\right)(x) - \frac{1}{8}\frac{1}{g(y)^2}\left(V(\widehat{g}_n) + B^2(\widehat{g}_n)\right)(y) \\
&+ \frac{1}{16}\frac{1}{g(x)^3}E\left(\widehat{g}_n(x) - g(x)\right)^3 + \frac{1}{16}\frac{1}{g(y)^3}E\left(\widehat{g}_n(y) - g(y)\right)^3 \\
&- \frac{5}{128}\frac{1}{g(x)^4}E\left(\widehat{g}_n(x) - g(x)\right)^4 - \frac{5}{128}\frac{1}{g(y)^4}E\left(\widehat{g}_n(y) - g(y)\right)^4 \\
&+ \frac{1}{4}\frac{Cov(\widehat{g}_n(x), \widehat{g}_n(y)) + B(\widehat{g}_n(x))B(\widehat{g}_n(y))}{g(x)g(y)} \\
&- \frac{1}{16}g(y)^{-1}g(x)^{-2}E\left[\left(\widehat{g}_n(x) - g(x)\right)^2\left(\widehat{g}_n(y) - g(y)\right)\right] \\
&- \frac{1}{16}g(x)^{-1}g(y)^{-2}E\left[\left(\widehat{g}_n(y) - g(y)\right)^2\left(\widehat{g}_n(x) - g(x)\right)\right] \\
&+ \frac{1}{32}g(x)^{-1}g(y)^{-3}E\left[\left(\widehat{g}_n(x) - g(x)\right)\left(\widehat{g}_n(y) - g(y)\right)^3\right] \\
&+ \frac{1}{32}g(y)^{-1}g(x)^{-3}E\left[\left(\widehat{g}_n(x) - g(x)\right)^3\left(\widehat{g}_n(y) - g(y)\right)\right] \\
&+ \frac{1}{64}g(x)^{-2}g(y)^{-2}E\left[\left(\widehat{g}_n(x) - g(x)\right)^2\left(\widehat{g}_n(y) - g(y)\right)^2\right] + \ldots\Bigg)
\end{aligned}
$$

where $V(\widehat{g}_n) = E\left[\left(\widehat{g}_n - E(\widehat{g}_n)\right)^2\right]$ and $B(\widehat{g}_n) = E(\widehat{g}_n - g)$.

Thus,

$$
\begin{aligned}
E\left(\widehat{H}(\theta)\widehat{H}^T(\theta)\right) = &\int\int s(x)\, s^T(y)\, g(x)g(y)dy\,dx \\
&+ \frac{1}{2}\int\int s(x)\, s^T(y)\, B(\widehat{g}_n(x))g(y)dy\,dx \\
&+ \frac{1}{2}\int\int s(x)\, s^T(y)\, g(x)B(\widehat{g}_n(y))dy\,dx \\
&- \frac{1}{8}\int\int r(x)\, s^T(y)\, g(y)\left(V(\widehat{g}_n(x)) + B^2(\widehat{g}_n(x))\right)dy\,dx \\
&- \frac{1}{8}\int\int s(x)\, r^T(y)\, g(x)\left(V(\widehat{g}_n(y)) - B^2(\widehat{g}_n(y))\right)dy\,dx
\end{aligned}
$$

$$+\frac{1}{16}\int\int t(x)\,s^T(y)\,g(y)\,E(\widehat{g}_n(x)-g(x))^3dy\,dx$$
$$+\frac{1}{16}\int\int s(x)\,t^T(y)\,g(x)\,E(\widehat{g}_n(y)-g(y))^3dy\,dx$$
$$-\frac{5}{128}\int\int u(x)\,s^T(y)\,g(y)\,E(\widehat{g}_n(x)-g(x))^4dy\,dx$$
$$-\frac{5}{128}\int\int s(x)\,u^T(y)\,g(x)\,E(\widehat{g}_n(y)-g(y))^4dy\,dx$$
$$+\frac{1}{4}\int\int s(x)\,s^T(y)\,cov(\widehat{g}_n(x),\widehat{g}_n(y))dy\,dx$$
$$+\frac{1}{4}\int\int s(x)\,s^T(y)\,B(\widehat{g}_n(x))B(\widehat{g}_n(y))dy\,dx$$
$$-\frac{1}{16}\int\int r(x)\,s^T(y)\,E\left[(\widehat{g}_n(x)-g(x))^2(\widehat{g}_n(y)-g(y))\right]dy\,dx$$
$$-\frac{1}{16}\int\int s(x)\,r^T(y)\,E\left[(\widehat{g}_n(x)-g(x))(\widehat{g}_n(y)-g(y))^2\right]dy\,dx$$
$$+\frac{1}{32}\int\int s(x)\,t^T(y)\,E\left[(\widehat{g}_n(x)-g(x))(\widehat{g}_n(y)-g(y))^3\right]dy\,dx$$
$$+\frac{1}{32}\int\int t(x)\,s^T(y)\,E\left[(\widehat{g}_n(x)-g(x))^3(\widehat{g}_n(y)-g(y))\right]dy\,dx$$
$$+\frac{1}{64}\int\int r(x)\,r^T(y)\,E\left[(\widehat{g}_n(x)-g(x))^2(\widehat{g}_n(y)-g(y))^2\right]dy\,dx...$$

where $B(\widehat{g}_n)=E(\widehat{g}_n-g)$, $V(\widehat{g}_n)=E(\widehat{g}_n-g)^2$, $\dot{f}_\theta=\frac{df_\theta}{d\theta}$, $s=\frac{\dot{f}_\theta}{\sqrt{f_\theta g}}$, $r=\frac{\dot{f}_\theta}{\sqrt{f_\theta}g^{\frac{3}{2}}}$, $t=\frac{\dot{f}_\theta}{\sqrt{f_\theta}g^{\frac{5}{2}}}$ and $u=\frac{\dot{f}_\theta}{\sqrt{f_\theta}g^{\frac{7}{2}}}$.

$$\begin{aligned}[E(\widehat{H}(\theta))][E(\widehat{H}(\theta))]^T &= \int\int s(x)\,g(x)s^T(y)\,g(y)\,dy\,dx\\ &+\frac{1}{4}\int\int s(x)\,B(\widehat{g}_n(x))s^T(y)\,B(\widehat{g}_n(y))dy\,dx\\ &+\frac{1}{2}\int\int s(x)\,s^T(y)\,g(y)\,B(\widehat{g}_n(x))dy\,dx\\ &+\frac{1}{2}\int\int s(x)\,s^T(y)\,g(x)B(\widehat{g}_n(y))dy\,dx\end{aligned}$$

$$-\frac{1}{8}\int\int s(x)\,r^T(y)\,g(x)\left(V(\widehat{g}_n(y))+B^2(\widehat{g}_n(y))\right)dy\,dx$$

$$-\frac{1}{8}\int\int r(x)\,s^T(y)\,g(y)\left(V(\widehat{g}_n(x))+B^2(\widehat{g}_n(x))\right)dy\,dx$$

$$+\frac{1}{16}\int\int s(x)\,t^T(y)\,g(x)E\left(\widehat{g}_n(y)-g(y)\right)^3dy\,dx$$

$$+\frac{1}{16}\int\int t(x)\,s^T(y)\,g(y)\,E\left(\widehat{g}_n(x)-g(x)\right)^3dy\,dx$$

$$-\frac{5}{128}\int\int s(x)\,g(x)\,u^T(y)\,E\left(\widehat{g}_n(y)-g(y)\right)^4dy\,dx$$

$$-\frac{5}{128}\int\int u(x)\,s^T(y)\,g(y)\,E\left(\widehat{g}_n(x)-g(x)\right)^4dy\,dx$$

$$-\frac{1}{16}\int\int s(x)\,r^T(y)\,B(\widehat{g}_n(x))\left(V(\widehat{g}_n(y))+B^2(\widehat{g}_n(y))\right)dy\,dx$$

$$-\frac{1}{16}\int\int r(x)\left(V(\widehat{g}_n(x))+B^2(\widehat{g}_n(x))\right)s^T(y)\,B(\widehat{g}_n(y))dy\,dx$$

$$+\frac{1}{32}\int\int s(x)\,B(\widehat{g}_n(x))t^T(y)\,E\left(\widehat{g}_n(y)-g(y)\right)^3dy\,dx$$

$$+\frac{1}{32}\int\int t(x)\,s^T(y)\,E\left(\widehat{g}_n(x)-g(x)\right)^3B(\widehat{g}_n(y))dy\,dx$$

$$+\frac{1}{64}\int\int r(x)\,r^T(y)\left(V(\widehat{g}_n(x))+B^2(\widehat{g}_n(x))\right)\left(V(\widehat{g}_n(y))+B^2(\widehat{g}_n(y))\right)dy\,dx+...$$

The variance of $\widehat{H}(\theta)$ is therefore

$$\begin{aligned}
&var\left(\widehat{H}(\theta)\right)\\
&= \frac{1}{4}\int\int s(x)\,s^T(y)\,cov(\widehat{g}_n(x),\widehat{g}_n(y))dy\,dx && \text{(A.1)}\\
&\quad -\frac{1}{16}\int\int r(x)\,s^T(y)\,E\left[\left(\widehat{g}_n(x)-g(x)\right)^2\left(\widehat{g}_n(y)-g(y)\right)\right]dy\,dx && \text{(A.2)}\\
&\quad -\frac{1}{16}\int\int s(x)\,r^T(y)\,E\left[\left(\widehat{g}_n(x)-g(x)\right)\left(\widehat{g}_n(y)-g(y)\right)^2\right]dy\,dx && \text{(A.3)}\\
&\quad +\frac{1}{16}\left[\int s(x)B(\widehat{g}_n(x))dx\right]\left[\int r(y)\left(V(\widehat{g}_n(y))+B^2(\widehat{g}_n(y))\right)dy\right]^T\\
&\quad +\frac{1}{16}\left[\int r(x)\left(V(\widehat{g}_n(x))+B^2(\widehat{g}_n(x))\right)dx\right]\left[\int s(y)\,B(\widehat{g}_n(y))dy\right]^T-...
\end{aligned}$$

where $B(\widehat{g}_n)=E(\widehat{g}_n-g)$, $V(\widehat{g}_n)=E(\widehat{g}_n-g)^2$, $\dot{f}_\theta=\frac{df_\theta}{d\theta}$, $s=\frac{\dot{f}_\theta}{\sqrt{f_\theta g}}$,

$r = \frac{\dot{f}_\theta}{\sqrt{f_\theta} g^{\frac{3}{2}}}$, $t = \frac{\dot{f}_\theta}{\sqrt{f_\theta} g^{\frac{5}{2}}}$ and $u = \frac{\dot{f}_\theta}{\sqrt{f_\theta} g^{\frac{7}{2}}}$.

Utilising the expressions for $cov(\widehat{g}_n(x), \widehat{g}_n(y))$, $\int K_h(x-z)\, s(x)\, dx$, $\int K_h(x-z)\, g(x)\, dx$ and $\int K_h(y-z)\, s(y)\, dy$ identified as equations $(A.4)$ to $(A.7)$ on pages 225 to 226 the first term $(A.1)$ in this expression becomes

$$
\begin{aligned}
& \frac{1}{4}\int\int s(x)\, s^T(y)\, cov(\widehat{g}_n(x), \widehat{g}_n(y)) dy\, dx \\
= \; & \frac{1}{4n}\int\int s(x)\, s^T(y) \int K_h(x-z) K_h(y-z) g(z) dz\, dy\, dx \\
& -\frac{1}{4n}\int\int s(x)\, s^T(y) \left[\int K_h(x-z) g(z) dz\right] \left[\int K_h(y-z) g(z) dz\right] dy\, dx \\
= \; & \frac{1}{4n}\int g(z) \left[\int s(x) K_h(x-z) dx\right] \left[\int s^T(y) K_h(y-z) dy\right] dz \\
& -\frac{1}{4n}\left[\int\int K_h(x-z) s(x) g(z) dx\, dz\right] \left[\int\int K_h(y-z) s^T(y) g(z)\, dy\, dz\right] \\
= \; & \frac{1}{4n}\int g(z) \left[s(z) + \frac{h^2}{2} s''(z) K_2\right] \left[s(z) + \frac{h^2}{2} s''(z) K_2\right]^T dz \\
& -\frac{1}{4n}\left[\int s(x)(g(x) + \frac{h^2}{2} g''(x) K_2) dx\right] \left[\int s(x)(g(x) + \frac{h^2}{2} g''(x) K_2) dx\right]^T \\
\approx \; & \frac{1}{4n}\left[\int s(x)\, s^T(x)\, g(x)\, dx - \left[\int s(x)\, g(x)\, dx\right] \left[\int s(x)\, g(x)\, dx\right]^T\right] \\
& + \frac{h^2 K_2}{4n}\left[\int g(x)\, s(x) [s''(x)]^T dx - \left[\int s(x)\, g(x)\, dx\right] \left[\int s(x)\, g''(x)\, dx\right]^T\right]
\end{aligned}
$$

where $K_2 = \int_{-\infty}^{\infty} K(u) u^2 du$, $s''(x) = \frac{d^2 s(x)}{dx^2}$ and $g''(x) = \frac{d^2 g(x)}{dx^2}$.

Utilising the relationship

$$cov((\widehat{g}_n - g)^2(x), (\widehat{g}_n - g)(y)) = cov(\widehat{g}_n^2(x), \widehat{g}_n(y)) - 2g(x) cov(\widehat{g}_n(x), \widehat{g}_n(y))$$

and equations A.4 to A.8 on pages 225 to 227 the second and third terms (expressions A.2 and A.3) can each be shown to be

$$\simeq -\frac{K_2}{16n^2h}\left[\int s(z)\,s^T(z)\,g(z)\,dz - \left[\int s(z)\,g(z)\,dz\right]\left[\int s(z)\,dz\right]^T\right]$$
$$-\frac{K_2h^2}{16n}\left[\int s(z)\,s^T(z)\,g''(z)\,dz - \left[\int s(x)\,g''(x)dx\right]\left[\int s(y)\,g(y)dy\right]^T\right]$$

The first few terms in the expression for the asymptotic variance of $\widehat{H}(\theta)$ are therefore

$$Var(\widehat{H}(\theta)) \approx \frac{1}{4n}\left[\int s(x)\,s^T(x)\,g(x)\,dx - \int\int s(x)\,s^T(y)\,g(x)\,g(y)\,dy\,dx\right]$$
$$+\frac{h^2K_2}{4n}\left[\int g(x)\,s(x)\,[s''(x)]^T\,dx - \int\int s(x)\,s^T(y)\,g(x)\,g''(y)\,dy\,dx\right]$$
$$-\frac{K_2}{8n^2h}\left[\int s(z)\,s^T(z)\,g(z)\,dz - \int\int s(z)\,s^T(y)\,g(z)\,dz\right]$$
$$-\frac{K_2h^2}{8n}\left[\int s(z)\,s^T(z)\,g''(z)\,dz - \int\int s(x)\,s^T(y)\,g''(x)g(y)\,dy\,dx\right]$$

where $K_2 = \int_{-\infty}^{\infty} K(u)u^2du$, $\dot{f}_\theta = \frac{df_\theta}{d\theta}$, $s = \frac{\dot{f}_\theta}{\sqrt{f_\theta g}}$, $s''(x) = \frac{d^2s(x)}{dx^2}$ and $g''(x) = \frac{d^2g(x)}{dx^2}$.

The central limit theorem applies to $\widehat{H}(\theta)$ so $\sqrt{n}(\hat{\theta}_h - \theta_1)$ has an asymptotically Normal distribution with $E(\sqrt{n}(\hat{\theta}_h - \theta_1))$ and $var(\sqrt{n}(\hat{\theta}_h - \theta_1))$ being $\sqrt{n}J^{-1}A$ and $J^{-1}KJ^{-1}$ respectively where

$$A = \frac{h^2}{4}\int s(x)\,g''(x)K_2\,dx$$

$$J = \int \dot{s}(x)\,g(x)\,dx$$

$$
\begin{aligned}
K &= \frac{1}{4}\left[\int s(x)\, s^T(x)\, g(x)\, dx - \int\int s(x)\, s^T(y)\, g(x)\, g(y)\, dy\, dx\right] \\
&+\frac{h^2 K_2}{4}\left[\int g(x)\, s(x)\, [s''(x)]^T\, dx - \int\int s(x)\, s^T(y)\, g(x)\, g(y)\, dy\, dx\right] \\
&-\frac{K_2}{8nh}\left[\int s(z)\, s^T(z)\, g(z)\, dz - \int\int s(z)\, s^T(y)\, g(z)\, g(y)\, dy\, dz\right] \\
&-\frac{K_2 h^2}{8}\left[\int s(z)\, s^T(z)\, g''(z)\; dz - \int\int s(x)\, s^T(y)\, g''(x) g(y)\, dy\, dx\right]
\end{aligned}
$$

and $K_2 = \int_{-\infty}^{\infty} K(u)u^2 du$, $\dot{f}_\theta = \frac{df_\theta}{d\theta}$, $s = \frac{\dot{f}_\theta}{\sqrt{f_\theta g}}$, $\dot{s} = \frac{ds}{d\theta}$, $s'' = \frac{d^2 s}{dx^2}$ and $g'' = \frac{d^2 g}{dx^2}$.

The first order approximation to the asymptotic mean and variance of $\sqrt{n}(\hat{\theta}_h - \theta_1)$ is obtained by ignoring the higher order terms in A and K to give

$$E(\sqrt{n}(\hat{\theta}_h - \theta_1) = 0$$

and

$$var(\sqrt{n}(\hat{\theta}_h - \theta_1) = J^{-1} K J^{-1}$$

where

$$J = \int \dot{s}(x)\, g(x)\, dx$$

and

$$K = \frac{1}{4}\left[\int s(x)\, s^T(x)\, g(x)\, dx - \int\int s(x)\, g(x)\, s^T(y)\, g(y)\, dy\, dx\right].$$

A.2.1 Some useful expressions

$$
\begin{aligned}
cov(\widehat{g}_h(x), \widehat{g}_h(y)) &= cov\left(\frac{1}{n}\Sigma_{i=1}^{n} K_h(x - X_i), \frac{1}{n}\Sigma_{j=1}^{n} K_h(y - X_j)\right) \quad \text{(A.4)} \\
&= n^{-2}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} cov(K_h(x - X_i), K_h(y - X_i)) \\
&= n^{-2}\Sigma_{i=1}^{n} cov(K_h(x - X_i), K_h(y - X_i)) \\
&= n^{-1} cov(K_h(x - X), K_h(y - X)) \\
&= n^{-1} \int K_h(x - z) K_h(y - z) g(z) dz \\
&\quad -n^{-1} \int K_h(x - z) g(z) dz \int K_h(y - z) g(z) dz
\end{aligned}
$$

$$
\int K_h(x - z)\, s(x) dx = \int \frac{1}{h} K\left(\frac{x - z}{h}\right) s(x)\, dx
$$

$$
\begin{aligned}
\text{Substituting} \quad u &= \frac{x - z}{h} \quad \text{and} \quad du = -\frac{dx}{h} \\
&= -\int_{\infty}^{-\infty} \frac{1}{h} K(u) s(z + uh) h\, du \\
&= \int_{-\infty}^{\infty} K(u)[s(z) + uhs'(z) + \frac{u^2 h^2}{2} s''(z) + ...]du \\
&\approx \int_{-\infty}^{\infty} K(u) s(z) du - hs'(z) \int_{-\infty}^{\infty} K(u) u du \\
&\qquad + \frac{h^2}{2} s''(z) \int_{-\infty}^{\infty} K(u) u^2 du \\
&\approx s(z) + \frac{h^2 s''(z)}{2} K_2 \qquad \text{(A.5)}
\end{aligned}
$$

$$E\left(K_h\left(x-z\right)\right) = \int K_h(x-z)g(z)dz$$
$$\approx \quad g(x) + \frac{h^2 g''(x)}{2} K_2 \qquad \text{as above.} \tag{A.6}$$

$$\text{Similarly,} \quad E\left(K_h\left(y-z\right)\right) \simeq g(y) + \frac{h^2 g''(y)}{2} K_2 \tag{A.7}$$

$$cov(\widehat{g}_n(x), \widehat{g}_n^2(y))$$
$$= \quad n^{-3}\Sigma_i\Sigma_j\Sigma_k cov(K_h(x-X_i), K_h(y-X_j), K_h(y-X_k))$$
$$= \quad n^{-3}\Sigma_i cov(K_h(x-X_i), K_h^2(y-X_i))$$
$$+\frac{2}{n^3}\Sigma_i\Sigma_{j\neq i} cov(K_h(x-X_i), K_h(y-X_i), K_h(y-X_j))$$
$$= \quad n^{-2} cov(K_h(x-X_i), K_h^2(y-X_i))$$
$$+\frac{2}{n^3}\Sigma_i\Sigma_{j\neq i} cov(K_h(x-X_i), K_h(y-X_i), K_h(y-X_j))$$

Now, taking second term only

$$\frac{2}{n^3}\Sigma_i\Sigma_{j\neq i} C(K_h(x-X_i), K_h(y-X_i), K_h(y-X_j))$$
$$= \quad \frac{2}{n^3}\Sigma_i\Sigma_{j\neq i} E(K_h(y-X_j)) C(K_h(x-X_i), K_h(y-X_i))$$
$$= \quad \frac{2}{n}\left(1-\frac{1}{n}\right)(K_h * g)(y) C(K_h(x-X_i), K_h(y-X_i))$$
$$= \quad \frac{2}{n}\left(1-\frac{1}{n}\right)\left(g(y) + \frac{h^2}{2}g''(y)K_2\right) C(K_h(x-z), K_h(y-z))$$
$$\approx \quad \frac{2}{n}\left(g(y) + \frac{h^2}{2}g''(y)K_2\right) C(K_h(x-z), K_h(y-z))$$

Therefore

$$cov(\widehat{g}_n(x), \widehat{g}_n^2(y)) = n^{-2} cov(K_h(x - X_i), K_h^2(y - X_i)) + \frac{2}{n}\left(g(y) + \frac{h^2}{2} g''(y) K_2\right) C(K_h(x - z), K_h(y - z)) \quad \text{(A.8)}$$

A.3 Derivation of the asymptotic mean and variance of Öztürk and Hettmansperger's criterion function estimator

The criterion function is defined as

$$d_F(\theta; p) = \int \left[G^p(x) - F_\theta^p(x)\right]^2 dx + \int \left[(1 - G(x))^p - (1 - F_\theta(x))^p\right]^2 dx$$

where $\theta = (\theta, \sigma, ..., \omega)^T$ is a vector of n parameters from the model distribution F_θ, G is the true distribution function and $p > 0$.

Differentiating this function with respect to θ gives the asymptotic estimating equation

$$\begin{aligned} \lambda(\theta; p) = {} & 2p \int G^p(x) F_\theta^{p-1}(x) \frac{dF_\theta(x)}{d\theta} dx \\ & -2p \int F_\theta^{2p-1}(x) \frac{dF_\theta(x)}{d\theta} dx \\ & -2p \int (1 - G(x))^p (1 - F_\theta(x))^{p-1} \frac{dF_\theta(x)}{d\theta} dx \\ & +2p \int (1 - F_\theta(x))^{2p-1} \frac{dF_\theta(x)}{d\theta} dx \end{aligned} \quad \text{(A.9)}$$

which can be set equal to zero and solved to give the parameter θ_p.

The true distribution function G is usually unknown however and is therefore estimated by the empirical distribution function which leads to the estimating equation

$$\begin{aligned}\widehat{\lambda}(\theta;p) &= 2p\int F_n^p(x)\,F_\theta^{p-1}(x)\,\frac{dF_\theta(x)}{d\theta}dx \\ &\quad -2p\int F_\theta^{2p-1}(x)\,\frac{dF_\theta(x)}{d\theta}dx \\ &\quad -2p\int (1-F_n(x))^p\,(1-F_\theta(x))^{p-1}\,\frac{dF_\theta(x)}{d\theta}dx \\ &\quad +2p\int (1-F_\theta(x))^{2p-1}\,\frac{dF_\theta(x)}{d\theta}dx \qquad \text{(A.10)}\end{aligned}$$

The solution of $\widehat{\lambda}(\theta;p)=0$ yields the OH estimator, $\widehat{\theta}_p$.

Subject to certain regularity conditions (see p.82 of Azzalini [2] for details), the asymptotic mean and variance of the OH estimator $\left(\widehat{\theta}_p\right)$ is obtained by expanding $\widehat{\lambda}\left(\widehat{\theta}_p;p\right)$ about $\widehat{\lambda}(\theta_p;p)$ using Taylor series as follows

$$\widehat{\lambda}\left(\widehat{\theta}_p;p\right)=\widehat{\lambda}(\theta_p;p)+(\widehat{\theta}_p-\theta_p)^T\widehat{\lambda}'(\theta_p;p)+\frac{1}{2}(\widehat{\theta}_p-\theta_p)^T\widehat{\lambda}''(\theta_p;p)(\widehat{\theta}_p-\theta_p)+\ldots$$

$$0=\widehat{\lambda}(\theta_p;p)+(\widehat{\theta}_p-\theta_p)^T\widehat{\lambda}'(\theta_p;p)+\ldots$$

where $\widehat{\lambda}'(\theta_p;p)$ is the vector of first derivatives and $\widehat{\lambda}''(\theta_p;p)$ is the matrix of second derivatives.

This leads to

$$\sqrt{n}(\widehat{\theta}_p-\theta_p)\simeq-\sqrt{n}\left[\widehat{\lambda}'(\theta_p;p)\right]^{-1}\widehat{\lambda}(\theta_p;p)$$

Replacing $\widehat{\lambda}'(\theta_p;p)$ with its expected value $E(\widehat{\lambda}'(\theta_p;p))$ (as justified by the "Law of Large Numbers") the asymptotic mean and variance of $\sqrt{n}(\widehat{\theta}_p-\theta_p)$ can be derived as follows

$$\begin{aligned} E(\sqrt{n}(\widehat{\theta}_p-\theta_p)) &\simeq -\sqrt{n}\left[E(\widehat{\lambda}'(\theta_p;p))\right]^{-1}E(\widehat{\lambda}(\theta_p;p)) \\ var(\sqrt{n}(\widehat{\theta}_p-\theta_p)) &\simeq n\left[E(\widehat{\lambda}'(\theta_p;p))\right]^{-1}var(\widehat{\lambda}(\theta_p;p))\left[E(\widehat{\lambda}'(\theta_p;p))\right]^{-1} \end{aligned}$$

The expected value of $\widehat{\lambda}(\theta;p)$ with respect to the true distribution is

$$\begin{aligned} E\left(\widehat{\lambda}(\theta;p)\right) &= E\left[2p\int\left[F_n^p(x)-F_\theta^p\right]F_\theta^{p-1}(x)\frac{dF_\theta(x)}{d\theta}dx\right] \\ &\quad -E\left[2p\int\left[(1-F_n(x))^p-(1-F_\theta)^p\right](1-F_\theta(x))^{p-1}\frac{dF_\theta(x)}{d\theta}dx\right] \\ &= 2p\int E\left[F_n^p(x)\right]F_\theta^{p-1}(x)\frac{dF_\theta(x)}{d\theta}dx-2p\int F_\theta^{2p-1}(x)\frac{dF_\theta(x)}{d\theta}dx \\ &\quad -2p\int E\left[(1-F_n(x))^p\right](1-F_\theta(x))^{p-1}\frac{dF_\theta(x)}{d\theta}dx \\ &\quad +2p\int(1-F_\theta(x))^{2p-1}\frac{dF_\theta(x)}{d\theta}dx \\ &\simeq 2p\int G^p(x)F_\theta^{p-1}(x)\frac{dF_\theta(x)}{d\theta}dx-2p\int F_\theta^{2p-1}(x)\frac{dF_\theta(x)}{d\theta}dx \\ &\quad -2p\int(1-G(x))^p(1-F_\theta(x))^{p-1}\frac{dF_\theta(x)}{d\theta}dx \\ &\quad +2p\int(1-F_\theta(x))^{2p-1}\frac{dF_\theta(x)}{d\theta}dx \end{aligned}$$

Thus, $E\left(\widehat{\lambda}(\theta_p;p)\right)\simeq\lambda(\theta_p;p)=0$.

Similarly, $E(\widehat{\lambda}'(\theta_p;p))=\lambda'(\theta_p;p)$ which is obtained by differentiating $\lambda(\theta;p)$

with respect to θ and evaluating at $\theta = \theta_p$.

$$\begin{aligned}
\lambda'(\theta;p) &= 2p(1-2p)\int F_\theta^{2p-2}(x)\,dx \\
&\quad +2p\int G^p(x)\,F_\theta^{p-1}(x)\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&\quad +2p(p-1)\int G^p(x)\,F_\theta^{p-2}(x)\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
&\quad -2p\int F_\theta^{2p-1}(x)\,\frac{d^2F_\theta(x)}{d\theta^2}dx \\
&\quad +2p(1-2p)\int (1-F_\theta(x))^{2p-2}\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
&\quad -2p\int (1-G(x))^p\,(1-F_\theta(x))^{p-1}\,\frac{d^2F_\theta(x)}{d\theta^2}dx
\end{aligned}$$

$$\begin{aligned}
&+2p(p-1)\int (1-G(x))^p\,(1-F_\theta(x))^{p-2}\left[\frac{dF_\theta(x)}{d\theta}\right]\left[\frac{dF_\theta(x)}{d\theta}\right]^T dx \\
&+2p\int (1-F_\theta(x))^{2p-1}\,\frac{d^2F_\theta(x)}{d\theta^2}dx
\end{aligned}$$

Thus $\lambda'(\theta_p;p) = \lambda'(\theta;p)|_{\theta=\theta_p}$.

The terms of equation A.10 which do not involve F_n have no effect on the variance of $\widehat{\lambda}(\theta;p)$ therefore

$$\begin{aligned}
var(\widehat{\lambda}(\theta;p)) &= 4p^2 var\left[\int \left[F_n^p(x)\,F_\theta^{p-1}(x) - [1-F_n(x)]^p\,[1-F_\theta(x)]^{p-1}\right]\frac{dF_\theta(x)}{d\theta}dx\right] \\
&= 4p^2\int\int v(s)\,v(t)\,Cov\left(F_n^p(s),F_n^p(t)\right)ds\,dt \qquad \text{(A.11a)} \\
&\quad +4p^2\int\int w(s)\,w(t)\,Cov\left((1-F_n(s))^p,(1-F_n(t))^p\right)ds\,dt \qquad \text{(A.11b)} \\
&\quad -4p^2\int\int v(s)w(t)Cov\left(F_n^p(s),(1-F_n(t))^p\right)ds\,dt \qquad \text{(A.11c)} \\
&\quad -4p^2\int\int w(s)v(t)Cov\left((1-F_n(s))^p,F_n^p(t)\right)ds\,dt \qquad \text{(A.11d)}
\end{aligned}$$

where $v(x) = F_\theta^{p-1}(x) \frac{dF_\theta(x)}{d\theta}$ and $w(x) = (1 - F_\theta(x))^{p-1} \frac{dF_\theta(x)}{d\theta}$.

Now since

$$\begin{aligned} Cov\left(F_n^p(s), F_n^p(t)\right) &\simeq p^2 G(s)^{p-1} G(t)^{p-1} Cov(F_n(s), F_n(t)) \\ &= \frac{p^2}{n} G(s)^{p-1} G(t)^{p-1} \left[G(\min(s,t)) - G(s)\, G(t)\right] \end{aligned}$$

The first term in this expression for the variance (A.11a) can therefore be simplified as follows

$$\begin{aligned} & 4p^2 \int\int F_\theta^{p-1}(s) \frac{dF_\theta(s)}{d\theta} F_\theta^{p-1}(t) \frac{dF_\theta(t)}{d\theta} Cov\left(F_n^p(s), F_n^p(t)\right) ds\, dt \\ &= \frac{4p^4}{n} \int\int r(s)\, r(t) \left[G(\min(s,t)) - G(s)\, G(t)\right] ds\, dt \\ &= \frac{4p^4}{n} \iint_{s<t} r(s) r(t) \left[G(s) - G(s)\, G(t)\right] ds\, dt \\ &+ \frac{4p^4}{n} \iint_{t<s} r(s) r(t) \left[G(t) - G(s)\, G(t)\right] ds\, dt \\ &= \frac{4p^4}{n} \iint_{s<t} r(s) r(t) G(s) \left[1 - G(t)\right] ds\, dt \\ &+ \frac{4p^4}{n} \iint_{t<s} r(s) r(t) G(t) \left[1 - G(s)\right] ds\, dt \\ &= 8\frac{p^4}{n} \iint_{s<t} r(s) r(t) G(s) \left[1 - G(t)\right] ds\, dt \end{aligned}$$

where $r(x) = F_\theta^{p-1}(x) \frac{dF_\theta(x)}{d\theta} G^{p-1}(x)$ and $u(x) = (1 - F_\theta(x))^{p-1} \frac{dF_\theta(x)}{d\theta} (1 - G(x))^{p-1}$.

Similarly,

$$\begin{aligned} & Cov\left((1 - F_n(s))^p, (1 - F_n(t))^p\right) \\ &\simeq p^2 (1 - G(s))^{p-1} (1 - G(t))^{p-1} Cov(F_n(s), F_n(t)) \\ &= \frac{p^2}{n} (1 - G(s))^{p-1} (1 - G(t))^{p-1} \left[G(\min(s,t)) - G(s)\, G(t)\right] \end{aligned}$$

$$Cov\left(F_n^p(s), (1-F_n(t))^p\right)$$

$$\simeq p^2 G^{p-1}(s)(1-G(t))^{p-1} Cov(F_n(s), F_n(t))$$

$$= \frac{p^2}{n} G^{p-1}(s)(1-G(t))^{p-1}\left[G(\min(s,t)) - G(s)G(t)\right]$$

and

$$Cov\left((1-F_n(s))^p, F_n^p(t)\right)$$

$$\simeq p^2 (1-G(s))^{p-1} G^{p-1}(t) Cov(F_n(s), F_n(t))$$

$$= \frac{p^2}{n}(1-G(s))^{p-1} G^{p-1}(t)\left[G(\min(s,t)) - G(s)G(t)\right]$$

which means that the second to fourth terms in the expression for the variance (A.11b-A.11d) can also be simplified to give

$$4p^2 \int\int w(s)w(t) Cov\left((1-F_n(s))^p, (1-F_n(t))^p\right)$$
$$= 8\frac{p^4}{n} \iint_{s<t} \left[u(s)u(t)G(s)(1-G(t))\right] ds\, dt,$$

$$4p^2 \int\int v(s)w(t) Cov\left(F_n^p(s), (1-F_n(t))^p\right)$$
$$= 4\frac{p^4}{n} \iint_{s<t} \left[r(s)u(t)G(s)(1-G(t)) + r(t)u(s)G(s)(1-G(t))\right] ds\, dt$$

and

$$4p^2 \int\int w(s)v(t) Cov\left((1-F_n(s))^p, F_n^p(t)\right)$$
$$= 4\frac{p^4}{n} \iint_{s<t} \left[u(s)r(t)G(s)[1-G(t)] + u(t)r(s)G(s)(1-G(t))\right] ds\, dt$$

where $v(x) = F_\theta^{p-1}(x) \frac{dF_\theta(x)}{d\theta}$, $w(x) = (1 - F_\theta(x))^{p-1} \frac{dF_\theta(x)}{d\theta}$,

$r(x) = G^{p-1}(x) \frac{dF_\theta(x)}{d\theta} F_\theta^{p-1}(x)$ and $u(x) = (1 - G(x))^{p-1} \frac{dF_\theta(x)}{d\theta} (1 - F_\theta(x))^{p-1}$.

Therefore,

$$\begin{aligned}
var(\widehat{\lambda}(\theta;p)) &= 8\frac{p^4}{n}\iint_{s<t} r(s)r(t)^T G(s)[1-G(t)]\,ds\,dt \\
&\quad +8\frac{p^4}{n}\iint_{s<t} u(s)\,u(t)^T G(s)[1-G(t)]\,ds\,dt \\
&\quad -4\frac{p^4}{n}\iint_{s<t} r(s)\,u(t)^T G(s)(1-G(t))\,ds\,dt \\
&\quad -4\frac{p^4}{n}\iint_{s<t} r(t)\,u(s)^T G(s)(1-G(t))\,ds\,dt \\
&\quad -4\frac{p^4}{n}\iint_{s<t} u(s)r(t)^T G(s)(1-G(t))\,ds\,dt \\
&\quad -4\frac{p^4}{n}\iint_{s<t} u(t)r(s)^T G(s)[1-G(t)]\,ds\,dt \\
&= 8\frac{p^4}{n}\iint_{s<t} r(s)r(t)^T G(s)[1-G(t)]\,ds\,dt \\
&\quad +8\frac{p^4}{n}\iint_{s<t} u(s)\,u(t)^T G(s)[1-G(t)]\,ds\,dt \\
&\quad -8\frac{p^4}{n}\iint_{s<t} r(t)\,u(s)^T G(s)(1-G(t))\,ds\,dt \\
&\quad -8\frac{p^4}{n}\iint_{s<t} u(t)r(s)^T G(s)[1-G(t)]\,ds\,dt \\
&= 8\frac{p^4}{n}\iint_{s<t} (r(s)+u(s))(r(t)+u(t))^T G(s)(1-G(t))\,ds\,dt
\end{aligned}$$

where $r(x) = G^{p-1}(x) \frac{dF_\theta(x)}{d\theta} F_\theta^{p-1}(x)$ and $u(x) = (1 - G(x))^{p-1} \frac{dF_\theta(x)}{d\theta} (1 - F_\theta(x))^{p-1}$.

The central limit theorem applies to $\widehat{\lambda}(\theta;p)$ (because the integrals may be replaced by sums of order statistics as detailed in Appendix C, p.246) so $\sqrt{n}(\widehat{\theta}_p - \theta_p)$ has an asymptotically Normal distribution with

$$E(\sqrt{n}(\widehat{\theta}_p - \theta_p)) \simeq 0$$

and

$$var(\sqrt{n}(\widehat{\theta}_p - \theta_p)) \simeq J^{-1} K J^{-1}$$

where

$$K = 2p^2 \iint_{s<t} (r(s) + u(s))(r(t) + u(t))^T G(s)(1 - G(t))\, ds\, dt$$

and

$$\begin{aligned}
J &= (1-2p)\int F_\theta^{2p-2}(x)\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&+ \int G^p(x) F_\theta^{p-1}(x) \frac{d^2 F_\theta(x)}{d\theta^2} dx \\
&+ (p-1)\int G^p(x) F_\theta^{p-2}(x)\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&- \int F_\theta^{2p-1}(x) \frac{d^2 F_\theta(x)}{d\theta^2} dx \\
&+ (1-2p)\int (1 - F_\theta(x))^{2p-2}\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&- \int (1-G(x))^p (1-F_\theta(x))^{p-1} \frac{d^2 F_\theta(x)}{d\theta^2} dx \\
&+ (p-1)\int (1-G(x))^p (1-F_\theta(x))^{p-2}\left(\frac{dF_\theta(x)}{d\theta}\right)\left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&+ \int (1-F_\theta(x))^{2p-1} \frac{d^2 F_\theta(x)}{d\theta^2} dx.
\end{aligned}$$

When $f_\theta = N(\theta, \sigma^2)$ the expression for J can be greatly simplified because the terms $-2p\int F_\theta^{2p-1}(x)\frac{dF_\theta(x)}{d\theta}dx + 2p\int (1-F_\theta(x))^{2p-1}\frac{dF_\theta(x)}{d\theta}dx$ in the estimating equation sum to zero for the location parameter and are constants

for the dispersion parameter. This leads to

$$\begin{aligned}
J &= \int G^p(x) F_\theta^{p-1}(x) \frac{d^2 F_\theta(x)}{d\theta^2} dx \\
&\quad + (p-1) \int G^p(x) F_\theta^{p-2}(x) \left(\frac{dF_\theta(x)}{d\theta}\right) \left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&\quad - \int (1-G(x))^p (1-F_\theta(x))^{p-1} \frac{d^2 F_\theta(x)}{d\theta^2} dx \\
&\quad + (p-1) \int (1-G(x))^p (1-F_\theta(x))^{p-2} \left(\frac{dF_\theta(x)}{d\theta}\right) \left(\frac{dF_\theta(x)}{d\theta}\right)^T dx \\
&= \int G^p(x) \frac{d}{d\theta} \left[F_\theta^{p-1}(x) \frac{dF_\theta(x)}{d\theta}\right] dx \\
&\quad - \int (1-G(x))^p \frac{d}{d\theta} \left[(1-F_\theta(x))^{p-1} \frac{dF_\theta(x)}{d\theta}\right] dx \\
&= \frac{d}{d\theta} \left[\int G^p(x) F_\theta^{p-1}(x) \frac{dF_\theta(x)}{d\theta} dx\right] \\
&\quad - \frac{d}{d\theta} \left[\int (1-G(x))^p (1-F_\theta(x))^{p-1} \frac{dF_\theta(x)}{d\theta} dx\right]
\end{aligned}$$

Now when $f_\theta = N(\theta, \sigma^2)$, $\frac{dF_\theta}{d\theta} = -\frac{1}{\sigma} f_\theta(x)$ and $\frac{dF_\theta}{d\sigma} = -\frac{1}{\sigma}\left(\frac{x-\theta}{\sigma}\right) f_\theta(x)$ so making the substitution $z = \left(\frac{x-\theta}{\sigma}\right)$ leads to the general expression

$$\begin{aligned}
J &= \frac{d}{d\theta} \left[\int G^{p-1}(z\sigma + \theta) \Phi^{p-1}(z) \psi(z) dz\right] \\
&\quad - \frac{d}{d\theta} \left[\int [1 - G(z\sigma + \theta)]^{p-1} [1 - \Phi(z)]^{p-1} \psi(z) dz\right]
\end{aligned}$$

where ψ represents either $-\sigma \frac{dF_\theta}{d\theta}$ or $-\sigma \frac{dF_\theta}{d\sigma}$ as required. After differentiating with respect to θ and reverting to the original parameterisation, it can be shown that

$$\begin{aligned}
J &= p \int [1 - G(x)]^{p-1} [1 - F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} g(x) dx \\
&\quad + p \int G^{p-1}(x) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} g(x) dx.
\end{aligned}$$

Thus, when $f_\theta = N(\theta, \sigma^2)$, the asymptotic variance of $\sqrt{n}(\widehat{\theta}_p - \theta_p)$ is given by $J^{-1}KJ^{-1}$ where

$$J = \int [r(s) + u(s)]\, g(s) ds$$

$$K = 2 \iint_{s<t} (r(s) + u(s))\, (r(t) + u(t))^T\, G(s)\, (1 - G(t))\, ds\, dt$$

$r(x) = G^{p-1}(x)\, \frac{dF_\theta(x)}{d\theta} F_\theta^{p-1}(x)$ and $u(x) = (1 - G(x))^{p-1}\, \frac{dF_\theta(x)}{d\theta}\, (1 - F_\theta(x))^{p-1}$.

Appendix B

Influence Functions

B.1 Influence function for the BHHJ estimator

The estimating equation for the BHHJ estimator is defined as

$$0 = \int f_\theta^{\alpha+1}(x)\, u_\theta(x)\, dx - \int f_\theta^{\alpha}(x)\, u_\theta(x)\, g(x)\, dx$$

where f_θ is the model, g the true density and u_θ the score function.

Substituting the density function $g_\varepsilon(x) = (1-\varepsilon)g(x)+\varepsilon\delta_\xi(x)$ for $g(x)$ enables the behaviour of the estimating procedure when the data contains outliers to be investigated. This dependence on ε affects the parameter estimates

so θ_ε replaces θ and the estimating equation becomes

$$\begin{aligned} 0 &= \int f_{\theta_\varepsilon}^{\alpha+1}(x)\, u_{\theta_\varepsilon}(x)\, dx - \int f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x)\, g(x)\, dx \\ &\quad + \varepsilon \int f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x)\, g(x)\, dx - \varepsilon f_{\theta_\varepsilon}^{\alpha}(\xi)\, u_{\theta_\varepsilon}(\xi) \end{aligned}$$

Differentiating with respect to ε gives

$$\begin{aligned} 0 &= (\alpha+1) \int f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x) \frac{df_\varepsilon}{d\varepsilon} dx + \int \frac{du_{\theta_\varepsilon}}{d\varepsilon} f_{\theta_\varepsilon}^{\alpha+1}(x)\, dx \\ &\quad - \int f_{\theta_\varepsilon}^{\alpha}(x) \frac{du_{\theta_\varepsilon}}{d\varepsilon} g(x)\, dx - \alpha \int f_{\theta_\varepsilon}^{\alpha-1}(x)\, u_{\theta_\varepsilon}(x) \frac{df_{\theta_\varepsilon}}{d\varepsilon} g(x)\, dx \\ &\quad + \varepsilon \left[\int \alpha f_{\theta_\varepsilon}^{\alpha-1}(x) \frac{df_{\theta_\varepsilon}}{d\varepsilon} u_{\theta_\varepsilon}(x)\, g(x)\, dx + \int f_{\theta_\varepsilon}^{\alpha}(x) \frac{du_{\theta_\varepsilon}}{d\varepsilon} g(x)\, dx \right] \\ &\quad + \int f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x)\, g(x)\, dx - \varepsilon f_{\theta_\varepsilon}^{\alpha}(\xi) \frac{du_{\theta_\varepsilon}}{d\varepsilon} \\ &\quad - \varepsilon \alpha f_{\theta_\varepsilon}^{\alpha-1}(\xi)\, u_{\theta_\varepsilon}(\xi) \frac{df_{\theta_\varepsilon}}{d\varepsilon} - f_{\theta_\varepsilon}^{\alpha}(\xi)\, u_{\theta_\varepsilon}(\xi) \end{aligned}$$

Now using the relationships $\frac{dU_{\theta_\varepsilon}}{d\varepsilon} = i_{\theta_\varepsilon} \frac{d\theta_\varepsilon}{d\varepsilon}$ and $\frac{df_\varepsilon}{d\varepsilon} = u_{\theta_\varepsilon}^T \frac{d\theta_\varepsilon}{d\varepsilon}$ leads to

$$\begin{aligned} 0 &= (\alpha+1) \int f_{\theta_\varepsilon}^{\alpha+1}(x)\, U_{\theta_\varepsilon}(x)\, U_{\theta_\varepsilon}^T(x) \frac{d\theta_\varepsilon}{d\varepsilon} dx \\ &\quad - \int i_{\theta_\varepsilon}(x)\, f_{\theta_\varepsilon}^{\alpha+1}(x) \frac{d\theta_\varepsilon}{d\varepsilon} dx \\ &\quad + \int f_{\theta_\varepsilon}^{\alpha}(x)\, i_{\theta_\varepsilon}(x) \frac{d\theta_\varepsilon}{d\varepsilon} g(x)\, dx \\ &\quad - \alpha \int f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x)\, u_{\theta_\varepsilon}^T(x) \frac{d\theta_\varepsilon}{d\varepsilon} g(x)\, dx \\ &\quad - \varepsilon \int \alpha f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x)\, u_{\theta_\varepsilon}^T(x) \frac{d\theta_\varepsilon}{d\varepsilon} g(x)\, dx \\ &\quad - \varepsilon \int f_{\theta_\varepsilon}^{\alpha}(x)\, i_\theta(x) \frac{d\theta_\varepsilon}{d\varepsilon} g(x)\, dx \\ &\quad + \int f_{\theta_\varepsilon}^{\alpha}(x)\, u_{\theta_\varepsilon}(x)\, g(x)\, dx - \varepsilon f_{\theta_\varepsilon}^{\alpha}(\xi)\, i_\theta(\xi) \frac{d\theta_\varepsilon}{d\varepsilon} \\ &\quad - \varepsilon \alpha f_{\theta_\varepsilon}^{\alpha}(\xi)\, u_{\theta_\varepsilon}(\xi)\, u_{\theta_\varepsilon}^T(\xi) \frac{d\theta_\varepsilon}{d\varepsilon} - f_{\theta_\varepsilon}^{\alpha}(\xi)\, u_{\theta_\varepsilon}(\xi) \end{aligned}$$

Letting $\varepsilon \to 0$, yields the influence function, $\frac{d\theta_\varepsilon}{d\varepsilon}$, as follows

$$\frac{d\theta_\varepsilon}{d\varepsilon} = J^{-1}\left(f_\theta^\alpha(\xi)\,u_\theta(\xi) - \int f_\theta^\alpha(x)\,u_\theta(x)\,g(x)\,dx\right)$$

where

$$\begin{aligned} J &= \int f_\theta^{\alpha+1}(x)\,u_\theta(x)\,u_\theta^T(x)\,dx \\ &\quad + \int \left(i_\theta(x) - \alpha u_\theta(x)\,u_\theta^T(x)\right)\left(g(x) - f_\theta(x)\right) f_\theta^\alpha(x)\,dx \end{aligned}$$

When $f_\theta = g = N(\theta, \sigma^2)$ this reduces to

$$\frac{d\theta_\varepsilon}{d\varepsilon} = \left[\int f_\theta^{\alpha+1}(x)\,u_\theta(x)\,u_\theta^T(x)\,dx\right]^{-1} f_\theta^\alpha(\xi)\,u_\theta(\xi).$$

B.2 Influence function for the Hellinger distance estimator

The estimating equation for the Hellinger distance estimator is as follows

$$0 = \int g^{\frac{1}{2}}(x) f_\theta^{-\frac{1}{2}}(x) \frac{df_\theta}{d\theta} dx$$

Substituting the density function $g_\varepsilon(x) = (1-\varepsilon)g(x) + \varepsilon\delta_\xi(x)$ for $g(x)$ enables the behaviour of the estimating procedure when the data contains outliers to be investigated. This dependence on ε affects the parameter estimates so θ_ε replaces θ and the estimating equation becomes

$$0 = \int g_{\varepsilon}^{\frac{1}{2}}(x) f_{\theta_\varepsilon}^{-\frac{1}{2}}(x) \frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon} dx$$

Differentiating this revised estimating equation with respect to ε gives

$$\begin{aligned} 0 &= \int \frac{dg_{\varepsilon}^{\frac{1}{2}}}{d\varepsilon} f_{\theta_\varepsilon}^{-\frac{1}{2}}(x) \frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon} dx \\ &\quad + \int g_{\varepsilon}^{\frac{1}{2}}(x) \left[f_{\theta_\varepsilon}^{-\frac{1}{2}}(x) \frac{d^2 f_{\theta_\varepsilon}}{d\varepsilon d\theta_\varepsilon} - \frac{1}{2} f_{\theta_\varepsilon}^{-\frac{3}{2}}(x) \left[\frac{df_{\theta_\varepsilon}}{d\varepsilon} \right] \left[\frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon} \right] \right] dx. \end{aligned}$$

Using the relationships $\frac{dg_\varepsilon}{d\varepsilon} = \delta_\xi - g$, $\frac{df_{\theta_\varepsilon}}{d\varepsilon} = \frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon}^T \frac{d\theta_\varepsilon}{d\varepsilon}$ and $\frac{d^2 f_{\theta_\varepsilon}}{d\varepsilon d\theta_\varepsilon} = \frac{d^2 f_{\theta_\varepsilon}}{d\theta_\varepsilon^2} \frac{d\theta_\varepsilon}{d\varepsilon}$

$$\begin{aligned} 0 &= \frac{1}{2} \int g_{\varepsilon}^{-\frac{1}{2}}(x) \left(\delta_\xi(x) - g(x) \right) f_{\theta_\varepsilon}^{-\frac{1}{2}}(x) \frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon} dx + \int g_{\varepsilon}^{\frac{1}{2}}(x) f_{\theta_\varepsilon}^{-\frac{1}{2}}(x) \frac{d^2 f_{\theta_\varepsilon}}{d\theta_\varepsilon^2} \left(\frac{d\theta_\varepsilon}{d\varepsilon} \right) dx \\ &\quad - \frac{1}{2} \int g_{\varepsilon}^{\frac{1}{2}}(x) f_{\theta_\varepsilon}^{-\frac{3}{2}}(x) \left[\frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon} \right] \left[\frac{df_{\theta_\varepsilon}}{d\theta_\varepsilon} \right]^T \left(\frac{d\theta_\varepsilon}{d\varepsilon} \right) dx \end{aligned}$$

Letting $\varepsilon \to 0$ and rearranging to give the influence function, $\frac{d\theta_\varepsilon}{d\varepsilon}$ gives

$$\frac{d\theta_\varepsilon}{d\varepsilon} = J^{-1} \left(\frac{1}{2} \int g(x)^{\frac{1}{2}}(x) f_\theta^{-\frac{1}{2}}(x) \frac{df_\theta}{d\theta} dx - \frac{1}{2} g^{-\frac{1}{2}}(\xi) f_\theta^{-\frac{1}{2}}(\xi) \frac{df_\theta(\xi)}{d\theta} \right)$$

where

$$J = \int g^{\frac{1}{2}}(x) f_\theta^{-\frac{1}{2}}(x) \frac{d^2 f_\theta}{d\theta^2} dx - \frac{1}{2} \int g^{\frac{1}{2}}(x) f_\theta^{-\frac{3}{2}}(x) \left[\frac{df_\theta}{d\theta} \right] \left[\frac{df_\theta}{d\theta} \right]^T dx.$$

When $f_\theta = g = N(\theta, 1)$ this reduces to

$$\frac{d\theta_\varepsilon}{d\varepsilon} = (\xi - \theta).$$

This is the same as the influence function for the maximum likelihood estimator.

B.3 Influence function for the OH estimator

The estimating equation for the OH estimator is as follows

$$\begin{aligned} 0 &= \int [1-G(x)]^p [1-F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} dx - \int [1-F_\theta(x)]^{2p-1} \frac{dF_\theta}{d\theta} dx \\ &\quad - \int G^p(x) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} dx + \int F_\theta^{2p-1}(x) \frac{dF_\theta}{d\theta} dx. \end{aligned} \tag{B.1}$$

Substituting the distribution function $G_\varepsilon(x) = (1-\varepsilon)G(x)+\varepsilon\Delta_\xi(x)$ for $G(x)$ enables the behaviour of the estimating procedure when the data contains outliers to be investigated. This dependence on ε affects the parameter estimates so θ_ε replaces θ and the estimating equation becomes

$$\begin{aligned} 0 &= \int [1-G_\varepsilon(x)]^p [1-F_{\theta_\varepsilon}(x)]^{p-1} \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} dx - \int [1-F_{\theta_\varepsilon}(x)]^{2p-1} \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} dx \\ &\quad - \int G_\varepsilon^p(x) F_{\theta_\varepsilon}^{p-1}(x) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} dx + \int F_{\theta_\varepsilon}^{2p-1}(x) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} dx. \end{aligned}$$

Differentiating with respect to ε then gives

$$\begin{aligned} 0 &= \int [1-G_\varepsilon(x)]^p [1-F_{\theta_\varepsilon}(x)]^{p-1} \frac{d^2F_{\theta_\varepsilon}}{d\theta_\varepsilon d\varepsilon} dx \\ &\quad - (p-1) \int [1-G_\varepsilon(x)]^p [1-F_{\theta_\varepsilon}(x)]^{p-2} \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} \frac{dF_{\theta_\varepsilon}}{d\varepsilon} dx \\ &\quad - p \int [1-G_\varepsilon(x)]^{p-1} [1-F_{\theta_\varepsilon}(x)]^{p-1} \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} \frac{dG_\varepsilon}{d\varepsilon} dx \\ &\quad - \int [1-F_{\theta_\varepsilon}(x)]^{2p-1} \frac{d^2F_{\theta_\varepsilon}}{d\theta_\varepsilon d\varepsilon} dx \\ &\quad + (2p-1) \int [1-F_{\theta_\varepsilon}(x)]^{2p-2} \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} \frac{dF_{\theta_\varepsilon}}{d\varepsilon} dx \end{aligned}$$

$$
\begin{aligned}
&- \int G_{\varepsilon}^{p}(x) F_{\theta_\varepsilon}^{p-1}(x) \frac{d^2 F_{\theta_\varepsilon}}{d\theta_\varepsilon d\varepsilon} dx \\
&- (p-1) \int G_{\varepsilon}^{p}(x) F_{\theta_\varepsilon}^{p-2}(x) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} \frac{dF_{\theta_\varepsilon}}{d\varepsilon} dx \\
&- p \int G_{\varepsilon}^{p-1}(x) F_{\theta_\varepsilon}^{p-1}(x) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} \frac{dG_{\varepsilon}}{d\varepsilon} dx \\
&+ \int F_{\theta_\varepsilon}^{2p-1}(x) \frac{d^2 F_{\theta_\varepsilon}}{d\theta_\varepsilon d\varepsilon} dx \\
&+ (2p-1) \int F_{\theta_\varepsilon}^{2p-2}(x) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} \frac{dF_{\theta_\varepsilon}}{d\varepsilon} dx.
\end{aligned}
$$

Now using the relationships $\frac{dF_\theta}{d\varepsilon} = \frac{dF_\theta}{d\theta}^T \frac{d\theta}{d\varepsilon}$, $\frac{d^2 F_\theta}{d\theta d\varepsilon} = \frac{d^2 F_\theta}{d\theta^2} \frac{d\theta}{d\varepsilon}$ and $\frac{dG_\varepsilon}{d\varepsilon} = \Delta_\xi - G$ gives

$$
\begin{aligned}
0 \;=\; & \int [1 - G_\varepsilon(x)]^p [1 - F_{\theta_\varepsilon}(x)]^{p-1} \frac{d^2 F_{\theta_\varepsilon}}{d\theta_\varepsilon^2} \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& - (p-1) \int [1 - G_\varepsilon(x)]^p [1 - F_{\theta_\varepsilon}(x)]^{p-2} \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right] \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right]^T \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& - p \int [1 - G_\varepsilon(x)]^{p-1} [1 - F_{\theta_\varepsilon}(x)]^{p-1} (\Delta_\xi(x) - G(x)) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} dx \\
& - \int [1 - F_{\theta_\varepsilon}(x)]^{2p-1} \frac{d^2 F_{\theta_\varepsilon}}{d\theta_\varepsilon^2} \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& + (2p-1) \int [1 - F_{\theta_\varepsilon}(x)]^{2p-2} \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right] \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right]^T \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& - \int G_\varepsilon^p(x) F_{\theta_\varepsilon}^{p-1}(x) \frac{d^2 F_{\theta_\varepsilon}}{d\theta_\varepsilon^2} \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& - (p-1) \int G_\varepsilon^p(x) F_{\theta_\varepsilon}^{p-2}(x) \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right] \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right]^T \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& - p \int G_\varepsilon^{p-1}(x) F_{\theta_\varepsilon}^{p-1}(x) (\Delta_\xi(x) - G(x)) \frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon} dx \\
& + \int F_{\theta_\varepsilon}^{2p-1}(x) \frac{d^2 F_{\theta_\varepsilon}}{d\theta_\varepsilon^2} \frac{d\theta_\varepsilon}{d\varepsilon} dx \\
& + (2p-1) \int F_{\theta_\varepsilon}^{2p-2}(x) \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right] \left[\frac{dF_{\theta_\varepsilon}}{d\theta_\varepsilon}\right]^T \frac{d\theta_\varepsilon}{d\varepsilon} dx.
\end{aligned}
$$

Letting $\varepsilon \to 0$ yields the influence function, $\frac{d\theta}{d\varepsilon}$, as follows

$$\frac{d\theta}{d\varepsilon} = [t(x;p)]^{-1} s(x;p)$$

where

$$s(x;p) = p \int \left[(1 - G(x))^{p-1} (1 - F_\theta(x))^{p-1} + G^{p-1}(x) F_\theta^{p-1}(x) \right] (\Delta_\xi(x) - G(x)) \frac{dF_\theta}{d\theta} dx$$

and

$$\begin{aligned} t(x;p) \;=\; & \int [1 - G(x)]^p [1 - F_\theta(x)]^{p-1} \frac{d^2 F_\theta}{d\theta^2} dx \\ & - (p-1) \int [1 - G(x)]^p [1 - F_\theta(x)]^{p-2} \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx \\ & - \int [1 - F_\theta(x)]^{2p-1} \frac{d^2 F_\theta}{d\theta^2} dx \\ & + (2p-1) \int [1 - F_\theta(x)]^{2p-2} \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx \\ & - \int G^p(x) F_\theta^{p-1}(x) \frac{d^2 F_\theta}{d\theta^2} dx \\ & - (p-1) \int G^p(x) F_\theta^{p-2}(x) \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx \\ & + \int F_\theta^{2p-1}(x) \frac{d^2 F_\theta}{d\theta^2} dx \\ & + (2p-1) \int F_\theta^{2p-2}(x) \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx. \end{aligned}$$

When $f_\theta = N(\theta, \sigma^2)$ $t(x;p)$ can be simplified greatly because the terms $-\int F_\theta^{2p-1}(x) \frac{dF_\theta(x)}{d\theta} dx + \int (1 - F_\theta(x))^{2p-1} \frac{dF_\theta(x)}{d\theta} dx$ in the estimating equation $[B.1]$ sum to zero for the location parameter and are constants for the

dispersion parameter. Thus,

$$\begin{aligned} t(x;p) &= \int [1-G(x)]^p [1-F_\theta(x)]^{p-1} \frac{d^2F_\theta}{d\theta^2} dx \\ &\quad - (p-1) \int [1-G(x)]^p [1-F_\theta(x)]^{p-2} \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx \\ &\quad - \int G^p(x) F_\theta^{p-1}(x) \frac{d^2F_\theta}{d\theta^2} dx \\ &\quad - (p-1) \int G^p(x) F_\theta^{p-2}(x) \left[\frac{dF_\theta}{d\theta}\right] \left[\frac{dF_\theta}{d\theta}\right]^T dx \\ &= \int [1-G(x)]^p \frac{d}{d\theta} \left[[1-F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta}\right] dx \\ &\quad - \int G^p(x) \frac{d}{d\theta} \left[F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta}\right] dx \\ &= \frac{d}{d\theta} \left[\int [1-G(x)]^p [1-F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} dx\right] \\ &\quad - \frac{d}{d\theta} \left[\int G^p(x) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} dx\right] \end{aligned}$$

Now when $f_\theta = N(\theta, \sigma^2)$, $\frac{dF_\theta}{d\theta} = -\frac{1}{\sigma} f_\theta(x)$ and $\frac{dF_\theta}{d\sigma} = -\frac{1}{\sigma}\left(\frac{x-\theta}{\sigma}\right) f_\theta(x)$ so making the substitution $z = \left(\frac{x-\theta}{\sigma}\right)$ leads to the general expression

$$\begin{aligned} J &= \frac{d}{d\theta} \left[\int [1-G(z\sigma+\theta)]^{p-1} [1-\Phi(z)]^{p-1} \psi(z)\, dz\right] \\ &\quad - \frac{d}{d\theta} \left[\int G^{p-1}(z\sigma+\theta) \Phi^{p-1}(z) \psi(z)\, dz\right] \end{aligned}$$

where ψ represents either $-\sigma\frac{dF_\theta}{d\theta}$ or $-\sigma\frac{dF_\theta}{d\sigma}$ as required. After differentiating with respect to θ and reverting to the original parameterisation, it can be shown that in this case

$$\begin{aligned} J &= p \int [1-G(x)]^{p-1} [1-F_\theta(x)]^{p-1} \frac{dF_\theta}{d\theta} g(x)\, dx \\ &\quad + p \int G^{p-1}(x) F_\theta^{p-1}(x) \frac{dF_\theta}{d\theta} g(x)\, dx. \end{aligned}$$

The influence function for the location parameter when $f_\theta = N(\theta, 1)$ is therefore

$$\frac{d\theta}{d\varepsilon} = \frac{\int \left[(1 - G(x))^{p-1} (1 - F_\theta(x))^{p-1} + G^{p-1}(x) F_\theta^{p-1}(x)\right] (\Delta_\xi(x) - G(x)) f_\theta(x)\, dx}{\int \left[(1 - G(x))^{p-1} (1 - F_\theta(x))^{p-1} + G^{p-1}(x) F_\theta^{p-1}(x)\right] g(x) f_\theta(x)\, dx}$$

Appendix C

Estimating equation for the OH estimator as the sum of order statistics

The estimating equation for the OH estimator with $f_\theta = N(\theta, \sigma^2)$ is

$$
\begin{aligned}
0 \;=\; & \int F_n^p(x)\,\Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,\frac{dz}{d\theta}dx && \text{(C.1)}\\
& -\int \Phi_\sigma^{2p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,\frac{dz}{d\theta}dx && \text{(C.2)}\\
& -\int [1-F_n(x)]^p\,[1-\Phi_\theta(x)]^{p-1}\,\phi_\sigma(x-\theta)\,\frac{dz}{d\theta}dx && \text{(C.3)}\\
& +\int [1-\Phi_\theta(x)]^{2p-1}\,\phi_\sigma(x-\theta)\,\frac{dz}{d\theta}dx && \text{(C.4)}
\end{aligned}
$$

where F_n is the empirical distribution function and $z = \left(\frac{x-\theta}{\sigma}\right)$.

The empirical distribution $F_n(x)$ is calculated using the formula $F_n(x) = \frac{1}{n}\sum I(X_i \leq x)$ where I is an indicator variable and n is the sample size. This is a step function with values as follows

$$F_n^p(x) = \begin{cases} 0 & I\left(x < X_{(1)}\right) \\ +\left(\frac{1}{n}\right)^p & I\left(X_{(1)} \leq x < X_{(2)}\right) \\ +\left(\frac{2}{n}\right)^p & I\left(X_{(2)} \leq x < X_{(3)}\right) \\ \vdots & \vdots \\ +\left(\frac{n-1}{n}\right)^p & I\left(X_{(n-1)} \leq x < X_{(n)}\right) \\ +1 & I\left(X_{(n)} \leq x\right) \end{cases}$$

where $X_{(i)}$ denotes the i^{th} order statistic.

Therefore $F_n^p(x) = \frac{1}{n^p}\sum_{i=0}^{n} i^p\, I\left(X_{(i)} \leq x < X_{(i+1)}\right)$ where $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$.

Thus for a generic function $h(x)$, $\int h(x)F_n^p(x)\,dx = \frac{1}{n^p}\sum_{i=0}^{n} i^p \int_{X(i)}^{X_{(i+1)}} h(x)dx$.

Applying this to the first term of the estimating equation (C.1) for the location parameter gives

$$\begin{aligned} & \int F_n^p(x)\,\Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,\frac{dz}{d\theta}dx \\ = & -\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^p \int_{X_{(i)}}^{X_{(i+1)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,dx \\ = & -\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^p \left[\frac{\sigma}{p}\Phi_\sigma^p(x-\theta)\right]_{X_{(i)}}^{X_{(i+1)}} \end{aligned}$$

$$= -\frac{1}{p}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^{p}\left[\Phi_{\sigma}^{p}\left(X_{(i+1)}-\theta\right)-\Phi_{\sigma}^{p}\left(X_{(i)}-\theta\right)\right]$$

$$= -\frac{1}{p}\left\{\sum_{j=1}^{n+1}\left(\frac{j-1}{n}\right)^{p}\Phi_{\sigma}^{p}\left(X_{(j)}-\theta\right)-\sum_{i=0}^{n}\left(\frac{i}{n}\right)^{p}\Phi_{\sigma}^{p}\left(X_{(i)}-\theta\right)\right\}$$

$$= -\frac{1}{p}\left\{1-\sum_{i=1}^{n}c_{i}\Phi_{\sigma}^{p}\left(X_{(i)}-\theta\right)\right\}$$

where $c_i = \left(\frac{i}{n}\right)^p - \left(\frac{i-1}{n}\right)^p$.

The second term (C.2) is

$$\frac{1}{\sigma}\int \Phi_{\sigma}^{2p-1}\left(x-\theta\right)\phi_{\sigma}\left(x-\theta\right)dx = \frac{1}{2p}.$$

Using the same approach for the third term (C.3) as with the first leads to

$$-\int\left[1-F_{n}\left(x\right)\right]^{p}\left[1-\Phi_{\sigma}\left(x-\theta\right)\right]^{p-1}\phi_{\sigma}\left(x-\theta\right)\frac{dz}{d\theta}dx$$

$$= \frac{1}{p}\left\{1-\sum_{i=1}^{n}c_{i}^{*}\left[1-\Phi_{\sigma}\left(X_{(i)}-\theta\right)\right]^{p}\right\}$$

where $c_i^* = \left(\frac{n+i-1}{n}\right)^p - \left(\frac{n-i}{n}\right)^p$.

and the last term (C.4) is

$$-\frac{1}{\sigma}\int\left[1-\Phi_{\theta}\left(x\right)\right]^{2p-1}\phi_{\sigma}\left(x-\theta\right)dx = -\frac{1}{2p}.$$

The estimating equation for the location parameter in the *Normal* family of models can therefore be written as

$$
\begin{aligned}
0 &= -\frac{1}{p}\left\{1-\sum_{i=1}^{n} c_i \Phi_\sigma^p\left(X_{(i)}-\theta\right)\right\}+\frac{1}{p}\left\{1-\sum_{i=1}^{n} c_i^*\left[1-\Phi_\sigma\left(X_{(i)}-\theta\right)\right]^p\right\} \\
&= \sum_{i=1}^{n}\left\{c_i \Phi_\sigma^p\left(X_{(i)}-\theta\right)-c_i^*\left[1-\Phi_\sigma\left(X_{(i)}-\theta\right)\right]^p\right\}
\end{aligned}
$$

where $c_i = \left(\frac{i}{n}\right)^p - \left(\frac{i-1}{n}\right)^p$ and $c_i^* = \left(\frac{n+i-1}{n}\right)^p - \left(\frac{n-i}{n}\right)^p$.

Similarly, for the dispersion parameter $\frac{dz}{d\theta} = -\frac{1}{\sigma}\left(\frac{x-\theta}{\sigma}\right)$ the first term (C.1) is

$$
\begin{aligned}
&\int F_n^p(x)\,\Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\,\frac{dz}{d\theta}dx \\
&= -\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^p \int_{X_{(i)}}^{X_{(i+1)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx \\
&= -\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^p \left[\int_{-\infty}^{X_{(i+1)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&\quad +\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^p \left[\int_{-\infty}^{X_{(i)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&= -\frac{1}{\sigma}\sum_{j=1}^{n+1}\left(\frac{j-1}{n}\right)^p \left[\int_{-\infty}^{X_{(j)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&\quad +\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{i}{n}\right)^p \left[\int_{-\infty}^{X_{(i)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&= -\frac{1}{\sigma}\sum_{j=1}^{n}\left(\frac{j-1}{n}\right)^p \left[\int_{-\infty}^{X_{(j)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&\quad -\frac{1}{\sigma}\int_{-\infty}^{\infty} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx \\
&\quad +\frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{i}{n}\right)^p \left[\int_{-\infty}^{X_{(i)}} \Phi_\sigma^{p-1}(x-\theta)\,\phi_\sigma(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right]
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{\sigma}\int_{-\infty}^{\infty}\Phi_{\sigma}^{p-1}(x-\theta)\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx \\
&\quad +\frac{1}{\sigma}\sum_{i=1}^{n}c_i\int_{-\infty}^{X_{(i)}}\Phi_{\sigma}^{p-1}(x-\theta)\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx \\
&= -\frac{1}{\sigma}K(\infty,p-1)+\frac{1}{\sigma}\sum_{i=1}^{n}c_iK(X_{(i)},p-1)
\end{aligned}
$$

where $K(t,m)=\int_{-\infty}^{t}\Phi_{\sigma}^{m}(x-\theta)\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx$.

The second term (C.2) is

$$
\begin{aligned}
-\int\Phi_{\sigma}^{2p-1}(x-\theta)\,\phi_{\sigma}(x-\theta)\frac{dz}{d\theta}dx &= \frac{1}{\sigma}\int\Phi_{\sigma}^{2p-1}(x-\theta)\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx \\
&= \frac{1}{\sigma}K(\infty,2p-1)
\end{aligned}
$$

The third term (C.3) is

$$
\begin{aligned}
&-\int(1-F_n(x))^{p}(1-\Phi_{\sigma}(x-\theta))^{p-1}\,\phi_{\sigma}(x-\theta)\frac{dz}{d\theta}dx \\
&= \frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{n-i}{n}\right)^{p}\int_{X_{(i)}}^{X_{(i+1)}}(1-\Phi_{\sigma}(x-\theta))^{p-1}\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx \\
&= \frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{n-i}{n}\right)^{p}\left[\int_{-\infty}^{X_{(i+1)}}(1-\Phi_{\sigma}(x-\theta))^{p-1}\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&\quad -\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{n-i}{n}\right)^{p}\left[\int_{-\infty}^{X_{(i)}}(1-\Phi_{\sigma}(x-\theta))^{p-1}\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&= \frac{1}{\sigma}\sum_{j=1}^{n+1}\left(\frac{n-j+1}{n}\right)^{p}\left[\int_{-\infty}^{X_{(ij)}}(1-\Phi_{\sigma}(x-\theta))^{p-1}\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&\quad -\frac{1}{\sigma}\sum_{i=0}^{n}\left(\frac{n-i}{n}\right)^{p}\left[\int_{-\infty}^{X_{(i)}}(1-\Phi_{\sigma}(x-\theta))^{p-1}\,\phi_{\sigma}(x-\theta)\left(\frac{x-\theta}{\sigma}\right)dx\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\sigma}\sum_{j=1}^{n}\left(\frac{n-j+1}{n}\right)^{p}\left[\int_{-\infty}^{X_{(j)}}\left(1-\Phi_{\sigma}\left(x-\theta\right)\right)^{p-1}\phi_{\sigma}\left(x-\theta\right)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&\quad -\frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{n-i}{n}\right)^{p}\left[\int_{-\infty}^{X_{(i)}}\left(1-\Phi_{\sigma}\left(x-\theta\right)\right)^{p-1}\phi_{\sigma}\left(x-\theta\right)\left(\frac{x-\theta}{\sigma}\right)dx\right] \\
&= \frac{1}{\sigma}\sum_{i=1}^{n}c_{i}^{*}\int_{-\infty}^{X_{(i)}}\left(1-\Phi_{\sigma}\left(x-\theta\right)\right)^{p-1}\phi_{\sigma}\left(x-\theta\right)\left(\frac{x-\theta}{\sigma}\right)dx \\
&= \frac{1}{\sigma}\sum_{i=1}^{n}c_{i}^{*}K^{*}\left(X_{(i)},p-1\right)
\end{aligned}
$$

where $K^{*}(t,m)=\int_{-\infty}^{t}\left[1-\Phi_{\sigma}\left(x-\theta\right)\right]^{p-1}\phi_{\sigma}\left(x-\theta\right)\left(\frac{x-\theta}{\sigma}\right)dx$.

The last term (C.4) is now

$$
\begin{aligned}
\int\left[1-\Phi_{\theta}\left(x\right)\right]^{2p-1}\phi_{\sigma}\left(x-\theta\right)\frac{dz}{d\theta}dx &= -\frac{1}{\sigma}\int\left[1-\Phi_{\theta}\left(x\right)\right]^{2p-1}\phi_{\sigma}\left(x-\theta\right)\left(\frac{x-\theta}{\sigma}\right)dx \\
&= -\frac{1}{\sigma}K^{*}\left(\infty,2p-1\right)
\end{aligned}
$$

Therefore the estimating equation for the dispersion parameter in the *Normal* family of models can be written as

$$
\begin{aligned}
0 &= -K(\infty,p-1)+\sum_{i=1}^{n}c_{i}K(X_{(i)},p-1)+K(\infty,2p-1) \\
&\quad +\sum_{i=1}^{n}c_{i}^{*}K^{*}\left(X_{(i)},p-1\right)-K^{*}\left(\infty,2p-1\right).
\end{aligned}
$$

Appendix D

Asymptotic mean squared error

The multi-parameter mean squared error (MSE) is $E\left[\left(\hat{\theta}-\theta_*\right)\left(\hat{\theta}-\theta_*\right)^T\right]$ where $\widehat{\theta}=\left(\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_k\right)^T$ is the vector of parameter estimates and $\theta_*=(\theta_{1*},\theta_{2*},...,\theta_{k*})^T$ the vector of true parameters. Thus when there are k parameters to be estimated the MSE is a $k\times k$ matrix with element $[i,j]$ given by $E\left[\left(\hat{\theta}_i-\theta_{i*}\right)\left(\widehat{\theta}_j-\theta_{j*}\right)\right]$ for $1\leq i,j\leq k$.

Assuming that $E\left(\widehat{\theta}\right)=\theta$ then this reduces to

$$E\left[\left(\hat{\theta}-\theta_*\right)\left(\hat{\theta}-\theta_*\right)^T\right]=\left(\theta-\theta_*\right)\left(\theta-\theta_*\right)^T+var(\widehat{\theta})$$

where $\theta-\theta_*$ is the $k\times 1$ bias vector and $var\left(\widehat{\theta}\right)$ the $k\times k$ dimensional $variance-covariance$ matrix of the estimators.

In the two parameter case, when the location and scale of a Normal distribution are to be estimated for example, $\widehat{\theta} = \left(\widehat{\theta}, \widehat{\sigma}\right)^T$, $\theta_* = (\theta_*, \sigma_*)^T$ and the mean squared error is

$$MSE\left(\widehat{\theta}, \widehat{\sigma}\right) = \begin{bmatrix} E\left[\left(\hat{\theta} - \theta_*\right)^2\right], & E\left[\left(\hat{\theta} - \theta_*\right)(\widehat{\sigma} - \sigma_*)\right] \\ E\left[\left(\hat{\theta} - \theta_*\right)(\widehat{\sigma} - \sigma_*)\right], & E\left[(\widehat{\sigma} - \sigma_*)^2\right] \end{bmatrix}.$$

Assuming $E\left(\hat{\theta}\right) = \theta$, $E\left[\left(\hat{\theta} - \theta_*\right)^2\right] = var(\widehat{\theta}) + (\theta - \theta_*)^2$, $E\left[(\widehat{\sigma} - \sigma_*)^2\right] = var(\widehat{\sigma}) + (\sigma - \sigma_*)^2$ and $E\left[\left(\hat{\theta} - \theta_*\right)(\widehat{\sigma} - \sigma_*)\right] = cov(\widehat{\theta}; \widehat{\sigma}) + (\theta - \theta_*)(\sigma - \sigma_*)$ which leads to

$$MSE(\hat{\theta}, \widehat{\sigma}) = \begin{bmatrix} var(\widehat{\theta}) + (\theta - \theta_*)^2, & cov(\widehat{\theta}; \widehat{\sigma}) + (\theta - \theta_*)(\sigma - \sigma_*) \\ cov(\widehat{\theta}; \widehat{\sigma}) + (\theta - \theta_*)(\sigma - \sigma_*), & var(\widehat{\sigma}) + (\sigma - \sigma_*)^2 \end{bmatrix}.$$

In order to obtain a single expression for the MSE (so the minimiser can be found), it is common practice to take either the trace or determinant. Using the simplest of these two approaches, the trace, leads to the following

$$MSE(\hat{\theta}) \simeq var(\widehat{\theta}) + var(\widehat{\sigma}) + (\theta - \theta_*)^2 + (\sigma - \sigma_*)^2$$

The asymptotic mean squared error ($AMSE$) can be obtained by replacing the variance and bias components with their asymptotic equivalents which gives the multi-parameter $AMSE$ as

$$As.E\left[\left(\hat{\theta}-\theta_*\right)\left(\hat{\theta}-\theta_*\right)^T\right]=\left(\theta-\theta_*\right)\left(\theta-\theta_*\right)^T+As.var(\widehat{\theta}). \qquad \text{(D.1)}$$

Taking the trace, for the two-parameter case, then gives

$$AMSE(\hat{\theta}) \simeq As.var(\widehat{\theta})+As.var(\widehat{\sigma})+(\theta-\theta_*)^2+(\sigma-\sigma_*)^2 \qquad \text{(D.2)}$$

where $As.var$ denotes an asymptotic variance and $\widehat{\theta}$ is an asymptotically unbiased estimate of θ.

Bibliography

[1] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton, New Jersey, 1972.

[2] A. Azzalini. *Statistical Inference: Based on the Likelihood.* Chapman and Hall, 1996.

[3] A. Basu, I. R. Harris, and S. Basu. Minimum distance estimation: The approach using density-based distances. In G.S.Maddala and C. R. Rao, editors, *Robust Inference*, volume 15 of *Handbook of Statistics*, pages 21–48. Elsevier Science B. V., 1997.

[4] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559, 1998.

[5] A. Basu and B. G. Lindsay. Minimum disparity estimation for con-

tinuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705, 1994.

[6] R. Beran. Minimum distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463, 1977.

[7] Panagiotis Besbeas. *Parameter Estimation Based on Empirical Transforms.* PhD thesis, University of Kent at Canterbury, 1999.

[8] N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B*, 46(3):440–464, 1984.

[9] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View.* Wiley, New York, 1985.

[10] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[11] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics : The Approach based on Influence Functions.* Probability and Mathematical Statistics. John Wiley & Sons, 1986.

[12] Ian Harris and Ayanendranath Basu. Hellinger distance as a penalized likelihood and the impact of empty cells. *Communications in Statistics:Simulation and Computation*, 23:1097–1113, 1994.

[13] C. R. Heathcote and M. J. Silvapulle. Minimum mean squared error estimation of location and scale parameters under misspecification of the model. *Biometrika*, 68(2):501–514, 1981.

[14] T. P. Hettmansperger, I. Heuter, and J. Hüsler. Minimum distance estimators. *Journal of Statistical Planning and Inference*, 41:291–302, 1994.

[15] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

[16] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[17] M. C. Jones. Rough-and-ready assessment of the degree and importance of smoothing in functional estimation. *Statistica Neerlandica*, 54:37–46, 2000.

[18] V. N. LaRiccia. Asymptotic properties of weighted L^2 quantile distance estimators. *The Annals of Statistics*, 10:621–624, 1982.

[19] B. G. Lindsay. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, 22(2):1081–1114, 1994.

[20] R. von Mises. On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348, 1947.

[21] Ömer Öztürk. A robust and almost fully efficient M-estimator. *Australian & New Zealand Journal of Statistics*, 40(4):415–424, 1998.

[22] Ömer Öztürk and T. P. Hettmansperger. Generalised weighted Cramér-von Mises distance estimators. *Biometrika*, 84(2):283–294, 1997.

[23] Ömer Öztürk and Thomas P Hettmansperger. Almost fully efficient and robust simultaneous estimation of location and scale parameters: A minimum distance approach. *Statistics & Probability Letters*, 29:233–244, 1996.

[24] Ömer Öztürk and Thomas P Hettmansperger. Simultaneous robust estimation of location and scale parameters: A minimum-distance approach. *The Canadian Journal of Statistics*, 26(2):217–229, 1998.

[25] W. C. Parr. Minimum distance estimation: A bibliography. *Communications in Statistics: Theory and Methods*, A10(12):1205–1224, 1981.

[26] W. C. Parr and T. de Wet. On minimum Cramér-von Mises-norm parameter estimation. *Communications in Statistics: Theory and Methods*, A10(12):1149–1166, 1981.

[27] W. C. Parr and W. R. Schucany. Minimum distance and robust estima-

tion. *Journal of the American Statistical Association*, 75(371):616–624, 1980.

[28] P. J. Rousseeuw. A new infinitesimal approach to robust estimation. *Z. Wahsch. verw. Geb.*, 56:127–132, 1981.

[29] P. J. Rousseeuw. *New Infinitesimal Methods in Robust Statistics.* PhD thesis, Vrije Universiteit, Brussels, Belgium, 1981.

[30] David W. Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43:274–285, 2001.

[31] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society (Series B)*, 53(3):683–690, 1991.

[32] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

[33] Jeffrey S. Simonoff. *Smoothing Methods in Statistics.* Springer, 1996.

[34] D. G. Simpson. Hellinger deviance tests: Efficiency, breakdown points and examples. *Journal of the American Statistical Association*, 84(405):107–113, 1989.

[35] R. G. Staudte and S. J. Sheather. *Robust Estimation and Testing.* John Wiley & Sons, 1990.

[36] R. Tamura and D. Boos. Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

[37] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

[38] J. Wolfowitz. The minimum distance method. *Annals of Mathematical Statistics*, 28:75–88, 1957.