NTNU
Norwegian University of
Science and Technology

Bandwidth Selection in Kernel Density Estimation

Håkon Kile

Master of Science in Physics and Mathematics
Submission date: May 2010
Supervisor: Nikolai Ushakov, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

Problem 1 E|X|

An explicit formula that unites the L2 and L1 mean integrated errors could reveal interesting information. Actually, one could try to find a common expression for all 0 < L < 2 mean integrated errors. Such an expression might be found by applying characteristic functions. This expression could then be used to calculate the exact mean integrated errors; and the corresponding optimal bandwidths can be found fairly easy by minimizing these expressions. One could also use these expressions to compare different bandwidth selectors in terms of best approximation of the exact optimal bandwidth.

Problem 2 cur(f)

The curvature of the target functions f is an important measure, as it is used in many different bandwidth selectors. As the target density is unknown, we usually have to rely on some estimate of the total curvature. New and better estimators of the curvature would be beneficial in many settings and could improve the bandwidth selection procedures.
It is also possible to use this curvature measure as a bandwidth selection procedure. The main idea is to estimate the curvature of the target density and then choose the bandwidth that corresponds to the density estimate with the same curvature as the curvature estimate. To explore this approach, and compare it with other methods, might results in new knowledge about the bandwidth selection problem.

Assignment given: 13. January 2010
Supervisor: Nikolai Ushakov, MATH

Preface

"Do not put your faith in what statistics say until you have carefully considered what they do not say. "
- William W. Watt

In this report we study one of the most popular non-parametric density estimators, the kernel density estimator. While the idea behind the kernel density estimator is simple, the practical implications are many. We can change the parameters involved, and by that severely change the features of the resulting density estimate. Finding the optimal parameters of the kernel density estimator is therefore extremely important in order to obtain a good estimate. Although several parameters are involved, the bandwidth is the dominant parameter. This is the motivation behind this project, where we study different aspects of bandwidth selection.

I want to thank to my advisor, professor Nikolai Ushakov, for all the help during this project.

Trondheim, Spring 2010
Håkon Kile

Abstract

In kernel density estimation, the most crucial step is to select a proper bandwidth (smoothing parameter). There are two conceptually different approaches to this problem: a subjective and an objective approach. In this report, we only consider the objective approach, which is based upon minimizing an error, defined by an error criterion.

The most common objective bandwidth selection method is to minimize some squared error expression, but this method is not without its critics. This approach is said to not perform satisfactory in the tail(s) of the density, and to put too much weight on observations close to the mode(s) of the density. An approach which minimizes an absolute error expression, is thought to be without these drawbacks. We will provide a new explicit formula for the mean integrated absolute error. The optimal mean integrated absolute error bandwidth will be compared to the optimal mean integrated squared error bandwidth. We will argue that these two bandwidths are essentially equal.

In addition, we study data-driven bandwidth selection, and we will propose a new data-driven bandwidth selector. Our new bandwidth selector has promising behavior with respect to the visual error criterion, especially in the cases of limited sample sizes.

Contents

Chapter 1

Introduction

Density estimation usually means fitting a parametric density model to a data sample. To use such a parametric estimation, we need an assumption about which density family the data is generated from. Often this assumption is based on little or no evidence, and if our assumption is wrong we end up with a false result. In non-parametric density estimation we do not need to make such an assumption. Non-parametric density estimation can be used as an initial analysis, or as an analysis tool itself.

The idea behind non-parametric density estimation is old, but many of the methods are quite compute-intensive. As the computational power has increased rapidly the last 30-40 years, non-parametric methods have received more attention and great improvements have been made.

1.1 Non-parametric density estimation

There exists many different non-parametric density estimators, like the histogram, the kernel density estimator and projection estimators. This report takes a closer look at the kernel density estimator, but problems and issues that arises in this context are related to problems and issues associated with other non-parametric density estimators as well.

The kernel density estimator at x, based upon an iid sequence $\mathbf{X} = (X_1, ...X_n)$ generated from a continuous, univariate target density $f_X(x)$, is

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

where $\int K(x)dx = 1$. There are three parameters in this estimator; the sample size n, the kernel function $K(\cdot)$ and the bandwidth h. Quite typically we cannot do anything about the sample size, and we have to make the best out of the situation by choosing an appropriate kernel and a suitable bandwidth. It is well known that the bandwidth selection is the most crucial step in order to obtain a good estimate, e.g. see chapter 2 in Wand and Jones (1995), [26]. Unfortunately, bandwidth selection is the most difficult

problem in kernel density estimation, and there exist no definite and unique solution to this problem. This report will look into different aspects of bandwidth selection methods.

It is rather surprising that the most effective bandwidth selection method is a visual assessment by the analyzer. The analyzer (visually) compares different density estimates, based upon a variety of bandwidths, and then chooses the bandwidth that corresponds to the (subjectively) optimal estimate. The unfortunate part is that such bandwidths are non-unique; this method will (probably) yield different bandwidths when performed by different analyzers. This method can also be very time consuming.

A more mathematical approach is to quantify the discrepancy between the estimate and the target density by some error criterion. The optimal bandwidth will then be the bandwidth value that minimizes the error measured by the error criterion. Such a method is objective, and can be time-efficient as computers can solve it numerically. However, if the error criterion does not reflect the discrepancy between the estimate and the target density, the error criterion is worthless.

In this report, we will take a closer look at three different objective error criteria; the squared error (SE), the absolute error (AE) and the visual error (VE). The first two lead to the following global error measures

- Mean integrated absolute error: $MIAE = E[\int |\hat{f}_n(x;h) - f(x)|dx]$
- Mean integrated squared error: $MISE = E[\int (\hat{f}_n(x;h) - f(x))^2 dx]$

which we will handle in more detail in chapter 2.

One attraction toward the subjective bandwidth selection method is that the two error criteria above only measure vertical distance. The human eye takes both horizontal and vertical distance into account when assessing the distance between two curves, it can therefore be some discrepancy between these approaches. An error criteria that better captures the visual perception of distance will be beneficial in many situations. Such a criterion was proposed by Marron and Tsybakov (1995), [18], and is denoted the visual error. This error criterion will be handled in detail in chapter 2.

While the above error criteria are interesting from a theoretical point of view, they do not do much good in practice, since the target function is unknown. In practice we have to rely on methods that approximates these errors, and the corresponding optimal bandwidths, based on the data at hand. Data-driven bandwidth selectors aim to estimate the optimal bandwidth using different techniques, e.g. cross validation and asymptotic expansions. We will discuss data-driven bandwidth selection in details later in this report.

1.2 Background reading

For a general introduction to non-parametric density estimation, we will refer to Silverman (1986), [23], and Wand and Jones (1995), [26]. These books cover a broad specter of methods and results within non-parametric density estimation.

There are countless papers on the MISE as an error criterion in density estimation. Marron and Wand (1992), [19], is a central paper in this report, and is a cornerstone for research related to the MISE criterion. In Marron and Wand (1992), [19], they found an exact formula for the MISE, when the target density is a member of the normal mixture family. Alternatively, we can study the asymptotic behavior of the MISE, see chapter 2 in Wand and Jones (1995), [26].

The monograph by Devroye and Györfi (1985), [6], is a a general introduction to the absolute error criterion in density estimation, and contains many useful results. Devroye and Györfi presented some compelling arguments in favor of the L_1 approach and some interesting and useful results concerning its asymptotic behavior. Hall and Wand (1988), [9, 10], derived an expression for the general asymptotic mean integrated absolute error. They showed that the asymptotically optimal bandwidth can be found by numerically solving a specific equation.

The visual error criterion is fairly new, and was first introduced in Marron and Tsybakov (1995), [18]. In Marron (1998), [17], further analysis of the VE was presented. As of today, many aspects of the VE criterion remains unexplored.

Data-driven bandwidth selection has been investigated in many papers. Two thorough and extensive papers, regarding modern data-driven bandwidth selectors, are Cao, Cuevas and Manteiga (1994), [3], and Jones, Marron and Sheather (1996), [12].

1.3 Outline of the report

In chapter 2 we present some of the background information we need in this study. We will discuss the kernel density estimator, the normal mixture family, the MISE, the MIAE and the VE in detail.

In chapter 3 we will present new results, as we will derive an explicit formula for the mean L^p error and the mean integrated L^p error, where $0 < p < 2$. We will use these formulas in the context of kernel density estimation and investigate the relationship between the MISE and the MIAE.

Chapter 4 will deal with data-driven bandwidth selection, and optimal density estimation with the respect to the VE criterion. Here we will present a new data-driven bandwidth selector with promising behavior.

In Appendix A there is a summary of the numerical methods we use in chapter 3 and 4. Appendix B contains some additional tables and figures related to the results in chapter 3 and 4. In Appendix C there is an overview of the notation and definitions used in the report.

Chapter 2

Survey

This report will only deal with univariate density estimation. There are no obstacles to use the kernel density estimator in multivariate density estimation, but not all results from the univariate setting are directly applicable in a multivariate setting. To first study the properties of the kernel density estimator in the univariate setting, and later on find the multivariate extensions, is therefore quite natural.

Our sample will throughout this report be an iid sequence $\mathbf{X} = (X_1, ...X_n)$ from a known continuous, univariate target density $f_X(x)$. Unqualified integral sign, $\int$, will mean integration over the whole real line, $\mathbb{R}$.

2.1 The kernel density estimator

The kernel density estimator, or the Parzen-Rosenblatt estimator (see Parzen (1962) [20]), at location x is:

$$\hat{f}_n(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{2.1}$$

where $K_h(x) = \frac{1}{h}K(\frac{x}{h})$, and $K(\cdot)$ is a function satisfying $\int K(x)dx = 1$ and $K(x) \geq 0$ $\forall x$. $\hat{f}$ will be a density as long as these conditions are satisfied and $K(\cdot)$ and h do not depend on x.

2.1.1 The kernel, $K(x)$

A usual requirement is that the kernel function is a continuous, unimodal, symmetric density function. This ensures easy interpretation of the density estimate, the estimate is a density itself, and it is easy to understand the estimation process.

It is possible to construct kernels with asymptotically better performance than these conventional kernels. We can remove some of the restrictions and obtain kernels like higher order kernels, infinite order kernels and the sinc kernel. In chapter 2 in Wand and Jones (1995), [26], and in Marron and Wand (1992), [19], there is a thorough discussion

regarding different types of kernels. The main result is that these alternative kernels often need very large sample sizes before they outperform the conventional kernels. In addition, these kernels distort the density estimates as the estimates are no longer guaranteed to be densities. A method for correcting density estimates that are not densities can be found in Glad, Hjort and Ushakov (2003), [7], but this problem will not be addressed in this report. We need to consider if the gain in performance is worth the sacrifices and disadvantages before we use such kernels.

In this report we focus on different error criteria and bandwidth selection methods, we will not look into optimal kernel selection. Therefore, we will use a simple kernel which ensures easily interpretable estimates, namely the $N(0,1)$ density. It also simplifies the mean integrated squared error calculation when combined with a target density that belongs to the normal mixture family, which we will see later.

2.1.2 The bandwidth, h

As mentioned in the introduction, the bandwidth is the most dominant parameter in the kernel density estimator. This parameter controls the amount of smoothing, and is analogous to the binwidth in a histogram. Even though the kernel estimator depends on the kernel and the bandwidth in a rather complicated way, a graphical representation clearly illustrates the difference in importance between these two parameters, see figure 2.3 page 13 and 2.6a page 29 in Wand and Jones (1995), [26]. Alternatively, we can see the dominant role the bandwidth has in a Taylor expansion of the MISE for a kernel density estimator, see chapter 2 in Wand and Jones (1995), [26].

A problem in bandwidth selection is what we define as the optimal bandwidth. Actually, there is no obvious and unique definition of the optimal bandwidth, as it depends on our objective. One analyzer might be interested in the precise location of the modes of the target density, like in spectroscopy, while other analyzers might be more interested in the number of modes, like in finance. These two situations calls for different bandwidths. As we will see in section 2.3, different error criteria, and their optimal bandwidths, are used for different objectives.

Another issues is whether we want to focus on one global bandwidth or if the bandwidth should be allowed to vary according to the location. The argument in favor of a bandwidth that varies, is that the amount of smoothing should reflect the amount of smoothing needed at each location. E.g. around the modes one needs little smoothing, while in the tail of a distribution one can use much smoothing. But if we do not know anything about the qualitative features of the target density, we cannot use such local bandwidths. In order to effectively use such an approach, we first need to do a preliminary analysis, which in practice means an analysis with a global bandwidth. Another problem with the local bandwidth is that the resulting density estimate is not a density. This situation is outside the scope of this study. For a quick discussion, see chapter 2.10 in Wand and Jones (1995), [26].

2.1.3 The target density

Some densities are harder to estimate than others. This is quite clear from a quick look at the plot of our target densities in figure 2.1 at page 15. We see that some of the densities have very complicated structures. For such densities, we need very large sample sizes before we can expect our density estimate to capture all the features of the target density. There have been several attempts to classify how hard a density is to estimate, e.g. see Wand and Devroye (1993), [25], or chapter 2.9 in Wand and Jones (1995), [26].

Another issues in kernel density estimation is the estimation of densities with boundaries, e.g. the exponential distribution. This problem is raised as a consequence of our choice of a continuous kernels, see chapter 2.11 in Wand and Jones (1995), [26], for more details. All our target densities are continuous, and we will not deal with this problem in this study.

2.2 Normal mixture family

The normal mixture family is very useful due to the fact that any density can be approximated arbitrarily well by a member of this family. A probability density function within this family is defined:

$$f(x) = \sum_{l=1}^{k} \omega_l \phi_{\sigma_l}(x - \mu_l) \tag{2.2}$$

where $\phi_\sigma(x)$ denotes the $N(0, \sigma^2)$ density, k is a positive integer, $\omega_1, ..., \omega_k$ is a set of positive numbers that sums to one, and for each l, $-\infty < \mu_l < \infty$ and $\sigma_l^2 > 0$.

We will use 15 different target densities in this report. They are defined in table 2.1 and can be seen in figure 2.1 on page 15. These densities were constructed by Marron and Wand (1992), [19], to: "...typify many different types of challenges to curve estimation." See section 3 in that paper for more details.

Many later studies have used these 15 densities as target densities. This is useful to us, since we can easily compare our results with other research studies within the field of kernel density estimation.

2.3 Error criteria

We mentioned in the introduction that the most effective bandwidth selection method is a visual assessment by the analyzer. The analyzer (visually) compares different density estimates, based upon a variety of bandwidths, and then chooses the bandwidth that corresponds to the (subjectively) optimal estimate. This method suffers from non-unique bandwidths, and it is time consuming. Further we mentioned that a more mathematical approach is to quantify the discrepancy between the estimate and the target density by some error criterion. In this section we discuss three different objective error criteria.

Table 2.1: The normal mixture densities

Density	Parameters
#1 Gaussian	$N(0,1)$
#2 Skewed unimodal	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{12}{13},(\frac{5}{9})^2)$
#3 Strongly skewed	$\sum_{l=0}^{7} \frac{1}{8}N(3\{(\frac{2}{3})^l - 1\},(\frac{2}{3})^{2l})$
#4 Kurtotic Unimodal	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2))$
#5 Outlier	$\frac{1}{10}N(0,1) + \frac{9}{10}N(0,(\frac{1}{10})^2)$
#6 Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2)) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
#7 Separated Bimodal	$\frac{1}{2}N(-\frac{3}{2}),(\frac{1}{2})^2)) + \frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
#8 Skewed Bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$
#9 Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5}(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$
#10 Claw	$\frac{1}{2}N(0,1) + \sum_{l=0}^{4} \frac{1}{10}N((\frac{l}{2} - 1),(\frac{1}{10})^2)$
#11 Double claw	$\frac{49}{100}N(-1,(\frac{2}{3})^2) + \frac{49}{100}N(1,(\frac{2}{3})^2)$
	$+ \sum_{l=0}^{6} \frac{1}{350}N(\frac{(l-3)}{2},(\frac{1}{100})^2)$
#12 Asymmetric claw	$\frac{1}{2}N(0,1) + \sum_{l=-2}^{2} \frac{2^{1-l}}{31}N(l + \frac{1}{2},(2^{-l}/10)^2$
#13 Asymmetric double claw	$\sum_{l=0}^{1} \frac{46}{100}N(2l-1,(\frac{2}{3})^2) + \sum_{l=1}^{3} \frac{1}{300}N(\frac{-l}{2},(\frac{1}{100})^2)$
	$+ \sum_{l=1}^{3} \frac{7}{300}N(\frac{l}{2},(\frac{7}{100})^2)$
#14 Smooth Comb	$\sum_{l=0}^{5} \frac{2^{5-l}}{63}N(\{65 - 96(\frac{1}{2})^l\},(\frac{35}{63})^2 7(2^{2l}))$
#15 Discrete Comb	$\sum_{l=0}^{2} \frac{2}{7}N((12l-15)/7,(\frac{2}{7})^2) + \sum_{l=8}^{10} \frac{1}{21}N(2l/7,(\frac{1}{21})^2)$

The first two should be quite familiar; the mean integrated squared error (MISE) and the mean integrated absolute error (MIAE). As error criteria in density estimation, they measure the vertical distance between the density estimate and target density, using the L_2 or L_1 norm. The optimal density estimates in terms of the MISE and MIAE are quite often far away from the visually optimal estimate, especially in the case of smaller sample sizes. MISE and MIAE optimal estimates are often undersmoothed for smaller sample sizes.

An error criterion that aims to produce estimates that are closer to the visually optimal estimates, is the visual error (VE) criterion. The VE is linked to the idea of qualitative smoothing, see Mammen (1993), [16], and was first introduced in Marron and Tsybakov (1995), [18]. We will take a closer look at the VE in this section.

For limited sample sizes, it is quite obvious that we cannot extract all features of the target density. At least for target densities with complicated structures, as discussed earlier. How many of the features we can extract will depend on different factors, e.g. how many false features we are willing to accept. An oversmooth estimate will smooth out a multimodal structure, while an undersmooth estimate will results in too many modes. We need to carefully weigh our options depending on our objective.

Data-driven bandwidth selectors aims to minimize an error, defined by some error criterion. Modern data-driven bandwidth selectors have a relatively good performance for large sample sizes, but often fail for limited sample sizes. This is mainly due to the fact that the MISE and MIAE optimal bandwidths often fail to produce good estimates for limited sample sizes. An error criteria with better performance for limited sample sizes might provide a better basis for data-driven bandwidth selectors. We will see that the VE often has better performance for limited sample sizes, and at the same time converges to the same optimal estimate as the MISE/MIAE optimal estimate for large sample sizes. On this basis, the VE has promising features.

2.3.1 The L2 Loss

The most common error criterion in estimation procedures is based upon the L_2-norm, and is here denoted SE; the squared error. The SE of the kernel density estimator at x is:

$$SE(\hat{f}_n(x;h)) = [\hat{f}_n(x;h) - f(x)]^2$$

The global error measure is found by integration, the integrated squared error:

$$ISE(\hat{f}_n(\cdot;h)) = \int [\hat{f}_n(x;h) - f(x)]^2 dx$$

These errors are random variables since they are functions of the sample. In order to minimize these errors we need deterministic numbers that represents their "natural values". The expectation yields such numbers.

A substantial amount of work has been devoted to the E(ISE) as an error criterion in kernel density estimation. One reason is the appealing variance-bias decomposition. This

variance-bias trade-off situation is well known in all statistical estimation procedures, and the interpretation of this trade-off situation should be familiar to the analyzer. Another attraction toward the E(ISE) is the easy exact calculation formula from Marron and Wand (1992), [19].

This section will summarize the results from chapter 2 in Wand and Jones (1995), [26], and Marron and Wand (1992), [19].

The MISE and AMISE

The expected value of the SE will be denoted MSE:

$$\begin{aligned} MSE(\hat{f}_n(x;h)) &= E(\hat{f}_n(x;h) - f(x))^2 \\ &= Var(\hat{f}_n(x;h) + (E\hat{f}_n(x;h) - f(x))^2 \\ &= Var(\hat{f}_n(x;h) + [bias(\hat{f}_n(x;h))]^2 \end{aligned}$$

where the variance-bias decomposition is represented after the last equality sign. The bias of the kernel density estimator is:

$$\begin{aligned} bias(\hat{f}_n(x;h))) &= E(\hat{f}_n(x;h)) - f(x) \\ &= E[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)] - f(x) \\ &= n \cdot \frac{1}{n} E[K_h(x - X)] - f(x) \\ &= \int K_h(x - y) f(y) dy - f(x) \\ &= (K_h * f)(x) - f(x) \end{aligned}$$

where $(K_h * f)(x)$ is the convolution. We use a similar method to find the variance of the kernel estimator:

$$\begin{aligned} Var(\hat{f}_n(x;h)) &= Var[\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)] \\ &= n \cdot \frac{1}{n^2} Var[K_h(x - X)] \\ &= \frac{1}{n}[\int K_h^2(x - y) f(y) - [\int K_h(x - y) f(y) dy]^2] \\ &= \frac{1}{n}[(K_h^2 * f)(x) - ((K_h * f)^2(x))] \end{aligned}$$

Combining these two yield the mean squared error at x. Taking the expectation of the ISE we obtain the mean integrated squared error, MISE:

$$MISE(\hat{f}_n(\cdot;h)) = E[ISE(\hat{f}_n(\cdot;h))] = E\int [\hat{f}_n(x;h) - f(x)]^2 dx \tag{2.3}$$

$$= \int E[\hat{f}_n(x;h) - f(x)]^2 dx = \int MSE(\hat{f}_n(x;h)) \tag{2.4}$$

The MISE depends on the bandwidth and the kernel in a rather complicated way, but through an asymptotic expansion one can see that the bandwidth is the dominant parameter. The asymptotic MISE (AMISE) is

$$AMISE(\hat{f}_n(\cdot;h)) = (nh)^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)R(f'') + o((nh)^{-1} + h^4) \tag{2.5}$$

where $R(K) = \int K^2(x)dx$ and $\mu_k(K) = \int x^k K(x)dx$. The minimizer of AMISE is h_{AMISE}:

$$h_{AMISE} = \left[\frac{R(K)}{\mu_2(K)R(f'')n}\right]^{1/5} \tag{2.6}$$

See chapter 2.5 in Wand and Jones (1995), [26], for more details.

MISE for normal mixture densities

When the target density belongs to the normal mixture family, the MISE can be calculated exactly. Marron and Wand (1992), [19], provided an explicit formula for calculating the exact MISE for arbitrary order kernels, see Theorem 2.1 in Marron and Wand (1992), [19]. When the kernel is the $N(0,1)$ density, their formula simplifies and can be expressed as in chapter 2.6 in Wand and Jones (1995), [26].

$$MISE(\hat{f}_n(\cdot;h)) = (2\pi^{\frac{1}{2}}nh)^{-1} + \mathbf{w^T}[(\mathbf{1} - \mathbf{n^{-1}})\mathbf{\Omega_2} - \mathbf{\Omega_1} + \mathbf{\Omega_0}]\mathbf{w} \tag{2.7}$$

where $\mathbf{w} = (\omega_\mathbf{1}, ..., \omega_\mathbf{k})^\mathbf{T}$ and $\mathbf{\Omega}_\alpha$ is the $k \times k$ matrix with (l, l') entry equal to

$$\phi_{(\alpha h + \sigma_l^2 + \sigma_{l'}^2)^{\frac{1}{2}}}(\mu_l - \mu_{l'})$$

From (2.7) it is fairly easy to determine the MISE optimal bandwidth, h_{MISE}.

Data driven bandwidth selection

The majority of modern data-driven bandwidth selectors are motivated by minimizing the MISE or the AMISE. The cross validation techniques, Least squares cross validation (LSCV) and Smoothed cross validation (SCV), aims to minimize the MISE. See chapter 3 in Wand and Jones (1995), [26], for more details.

Other bandwidth selectors aim to minimize the AMISE. The most popular data-driven bandwidth selector is the one of Sheather and Jones (1991), [22], which we will call the Direct plug-in (DPI). In addition, we have the Biased cross-validation, which also aims to minimize the AMISE. We will take a closer look at the DPI in chapter 4. Again, for more details, and alternative bandwidth selectors, see chapter 3 in Wand and Jones (1995), [26].

2.3.2 The L1 Loss

An alternative to the MISE criterion, is an error measure that makes use of the L_1-norm, the absolute error (AE). The AE for the kernel density estimator at x is:

$$AE = |\hat{f}_n(x;h) - f(x)|$$

Once again we obtain the global error by integration; the integrated absolute error:

$$IAE(\hat{f}_n(\cdot;h)) = \int |\hat{f}_n(x;h) - f(x)|dx$$

This error criterion received a lot of attention due to Devroye and Györfi (1985), [6]. Devroye and Györfi stated these arguments in favor of the IAE as an error criterion:

- L_1 is the natural space for all densities.
- The IAE is invariant under monotone transformation of the coordinate axis. The IAE is a universal measure of the error, e.g. one can compare different error values without any scaling. (This is not the case for the ISE.)
- The IAE has the appealing interpretation of being the area between the curves, which makes it easy to visualize.
- IAE is proportional to the total variation: $\int |f_n - f|/2$
- If $IAE \to 0$ as $n \to \infty$, then $\hat{f}$ is a consistent estimator of f

MAE and MIAE

The AE and IAE are random variables and again we need deterministic numbers that represent their "natural value". We can argue in favor of both the median and the expectation as the appropriate procedure to obtain these values. For an extensive discussion, see Devroye (1987), [4]. We will use the expected absolute error and the expected integrated absolute error:

$$MAE(\hat{f}_n(x;h)) = E[AE(\hat{f}_n(x;h))] = E|\hat{f}_n(x;h) - f(x)| \tag{2.8}$$

$$MIAE(\hat{f}_n(\cdot;h)) = E[IAE(\hat{f}_n(\cdot;h))] = \int E|\hat{f}_n(x;h) - f(x)|dx \tag{2.9}$$

A lot of research has been done within this field, e.g. Devroye and Györfi (1985), [6], Hall and Wand (1988), [9, 10], and Scott and Wand (1991), [21]. All this research investigates asymptotic expansions and upper and lower bounds for the quantities in (2.8) and (2.9). For details regarding the asymptotic behavior of the MAE and MIAE, we refer to the articles mentioned in this paragraph. In chapter 3 we will show that we can calculate both the MAE and the MIAE explicitly.

Data driven bandwidth selection

There are very few bandwidth selectors that aim to minimize the absolute error. One of the very few, is the Double kernel method of Devroye (1989), [5]. We will not discuss this further here, and only refer to that paper for more details.

2.3.3 The Visual Error Criterion

The MISE and MIAE optimal estimates are based upon minimizing the integrated vertical distance between the density estimate and the target density. We have previously mentioned the idea of qualitative smoothing, and referred to Mammen (1993), [16]. In some situations we are more interested in the shape of the density, rather than the point values. In such a situation we are willing to increase the ISE/IAE in order to extract the qualitative features of the target density.

While doing a visual inspection of a density estimate, we use many of the qualitative features as indicators of how good the estimate is; how smooth it is, number of modes etc. The VE criterion can been seen as a weighted average between the quantitative and qualitative features of the density estimate. This idea was first introduces in Marron and Tsybakov (1995), [18], and further developed in Marron (1998), [17]. In this section we will give a summary of the methods and results of those papers.

The visual error criterion

Much of the inadequacy of the MISE and the MIAE lies in the fact that they only measure the vertical distance between the density estimate and the target density. From a visual point of view, both vertical and horizontal distance is important when we compare a density estimate with a target function. The question is how can we construct a distance measure that accounts for both vertical and horizontal distance?

Any continuous function, $f : [a : b] \to \mathbb{R}$, can be represented by its graph:

$$G_f = \{(x, y) : x \in [a, b], y = f(x)\} \subset \mathbb{R}^2$$

The shortest distance from a point, (x, y), to a graph, G, is

$$d((x, y), G) = \inf_{(x',y') \in G} ||(x, y) - (x', y')||_2,$$

where $|| \cdot ||_2$ is the euclidian distance. We can use other norms as well, see Marron and Tsybakov (1995), [18], but in this report we will only use the euclidian distance. This way of measuring the distance between a point and a function is an improvement over the L_1 and L_2 norm, as it accounts for both horizontal and vertical distance. The distance between a kernel density estimate, $\hat{f}(x; h)$, and a target function, $f_X(x)$, we now define as:

$$d((x, \hat{f}(x; h)), G_{f_X}) = \inf_{(x',y') \in G_{f_X}} ||(x, \hat{f}(x; h)) - (x', y')||_2,$$

We can use this distance measure to find a global performance measure defined as:

$$VE_2(\hat{f} \to f_X) = \left[\int_a^b d((x, \hat{f}(x;h)), G_{f_X})^2 dx \right]^{1/2}, \tag{2.10}$$

which is denoted the asymmetric visual error. It is the integrated distance from the kernel density estimate to the target function. A symmetrized version of this criterion follows as:

$$SE_2(\hat{f}, f_X) = \left[VE_2(\hat{f} \to f_X)^2 + VE_2(f_X \to \hat{f})^2 \right]^{1/2}.$$

Marron and Tsybakov (1995), [18], recommended these two error criteria for two different situations:

- If the goal is to duplicate an experienced analyst's choice, i.e. recover the features of the target density that are well documented trough the sample, use $VE_2(\hat{f} \to f)$.
- If the goal is to recover as many of the qualitative features of the target density as possible, use $SE_2(\hat{f}, f)$

Marron and Tsybakov (1995), [18], showed that both $E[VE_2(\hat{f} \to f)]^2$ and $E[VE_2(f \to \hat{f})]^2$ converges to MISE, but with faster convergence for $E[VE_2(\hat{f} \to f)]^2$. The SE_2 converges to twice the MISE. We will in the rest of this report only work with $E[VE_2(\hat{f} \to f)]^2$, which we will denote EVE2.

Discretizing and Scaling

The minimizer of EVE2, h_{EVE2}, cannot be found explicitly; we need to find it numerically. As computers cannot handle continuous functions, we need to discretize the graphs of the kernel density estimate and the target function. We do this by dividing our x-interval into a grid of equally spaced points, χ, where the distance between points in χ is Δx.

$$G_f^{\text{discr}} = \{(x, f(x)) : x \in \chi\} \subset \mathbb{R}^2$$

We can use this discretized version in (2.10), and use for instance Riemann approximations or Simpson's rule to approximate the integral.

A well known fact about norms is that they are not scale invariant, except for the L_1-norm in some special cases. Therefore we need to scale before we calculate the VE. We will scale [a,b] and $[\inf_{x\in[a,b]} f(x), \sup_{x\in[a,b]} f(x)]$ onto $[0,1]$, to avoid that the horizontal or vertical distance dominates the error measure.

In this report we scale x onto $[0,1]$ by:

$$x_{\text{scaled}} = \frac{x - \min_{x\in[a,b]}(x)}{\max_{x\in[a,b]}(x) - \min_{x\in[a,b]}(x)}.$$

Further will we scale $\hat{f}$ and f_X onto $[0, 1]$ by:

$$f_{\text{scaled}}(x_{\text{scaled}}) = f(x)/(\max_{x \in [a,b]}(f_X(x)))$$

This will not map $\hat{f}$ onto $[0, 1]$ when $\max_{x \in [a,b]}(\hat{f}) > \max_{x \in [a,b]}(f_X)$, but it is rare that $\max(\hat{f}) \gg \max(f_X)$, and we want to use the same scaling factor every time.

Bandwidth selection

To our knowledge there exist no bandwidth selector that specifically aims to minimize the VE. The VE is a quite difficult criterion to handle, but it is a better error criterion than MISE/MIAE for most situations where we are faced with a limited sample sizes. A bandwidth selector that aims to minimize the VE might end up with an all over better performance than today's bandwidth selectors.

A starting point for such a bandwidth selector might be the asymptotic expansion of the EVE2, which can be found in Marron and Tsybakov (1995), [18], but this is not considered in this report.

For a review of many of today's data-driven bandwidth selectors in terms of their VE2 performance, see Marron (1998), [17].

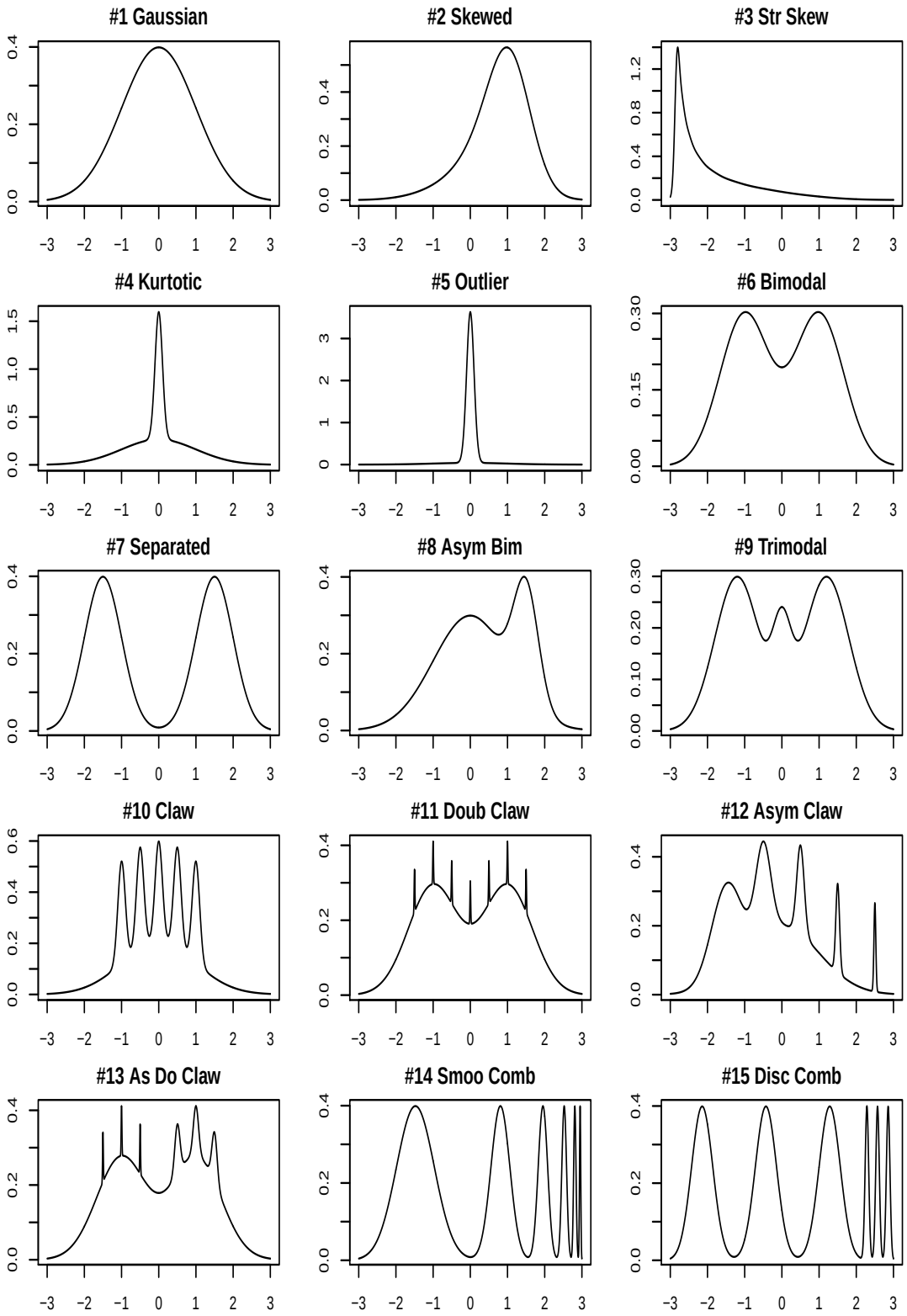

Figure 2.1: The 15 target densities in report. All are from Marron and Wand (1992), [19].

Chapter 3

Mean Lp Error for Kernel Density Estimators

In the preceding chapter we discussed different objective error criteria. The mean integrated squared error is the most popular, in particular because of its mathematical simplicity, but it is not without its flaws. In this chapter we will take a closer look at an alternative error criterion; the mean integrated L^p error, where L^p means $|\cdot|^p$, $0 < p < 2$. We will derive explicit formulas for the MLPE and the MILPE of the kernel density estimator. With $p = 1$ as the most interesting case; it results in MAE and MIAE.

3.1 The MLPE and MILPE

One of the important topics in density estimation is the analysis of the performance of the density estimator, and such an analysis needs an error measurement. An error criterion can quantify the discrepancy between the density estimate and the target density. In chapter 2 we discussed three different error criteria.

A major drawback for the mean absolute error, compared to the mean squared error, is the lack of an explicit formula. In this chapter we will overcome this problem by deriving explicit formulas for the MLPE and the MILPE.

3.1.1 Some general results

We will start by deriving some general results we will need later. Two well known trigonometric relations are:

$$1 = \sin^2(x/2) + \cos^2(x/2)$$
$$\cos x = \cos^2(x/2) - \sin^2(x/2)$$

which combined results in:

$$1 - \cos x = 2\sin(x/2).$$

The function inside the integral below is an even function, which makes it possible to change the integration limits; as we have done after the first equality sign. The trigonometric relation above is used after the first equality sign as well. We get:

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1-\cos t}{|t|^{(p+1)}}dt &= 2\int_{0}^{\infty} \frac{2\sin^2(t/2)}{t^{(p+1)}}dt \\
&= 2\int_{0}^{\infty} \frac{2\sin^2 u}{(2u)^{(p+1)}}2du \\
&= \frac{8}{2^{(p+1)}}\int_{0}^{\infty} \frac{\sin^2 u}{u^{(p+1)}}du \\
&= \frac{4}{2^p}\left[\frac{-\pi}{2^{3-(p+1)}\Gamma(p+1)\cos(\frac{(p+1)\pi}{2})}\right] \\
&= \frac{4}{2^p}\left[\frac{-\pi}{2^{(2-p)}\Gamma(p+1)\cos(\frac{(p+1)\pi}{2})}\right] \\
&= \left[\frac{-\pi}{\Gamma(p+1)\cos(\frac{(p+1)\pi}{2})}\right] \\
&= c(p).
\end{aligned}$$

In figure 3.1 we can see a plot of $c(p)$ for $0 < p < 2$. The following lemma will be the basis for our explicit formulas for the MLPE and the MILPE.

Lemma 1. *For any* $0 < p < 2$,

$$|x|^p = \tfrac{1}{c(p)}\textstyle\int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|t|^{(p+1)}}dt$$

where $c(p) = \frac{-\pi}{\Gamma(p+1)\cos(\frac{(p+1)\pi}{2})}$

Proof. We define a function $g(x)$.

$$g(x) \equiv \int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|t|^{(p+1)}}dt,\ 0 < p < 2 \tag{3.1}$$

Now, let $x > 0$. Then it follows that:

$$\begin{aligned}
g(x) &= \int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|t|^{(p+1)}}dt \\
&= \int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|\frac{xt}{x}|^{(p+1)}}\frac{1}{x}d(xt) \\
&= x^p\int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|xt|^{(p+1)}}d(xt) \\
&= x^p c(p)
\end{aligned}$$

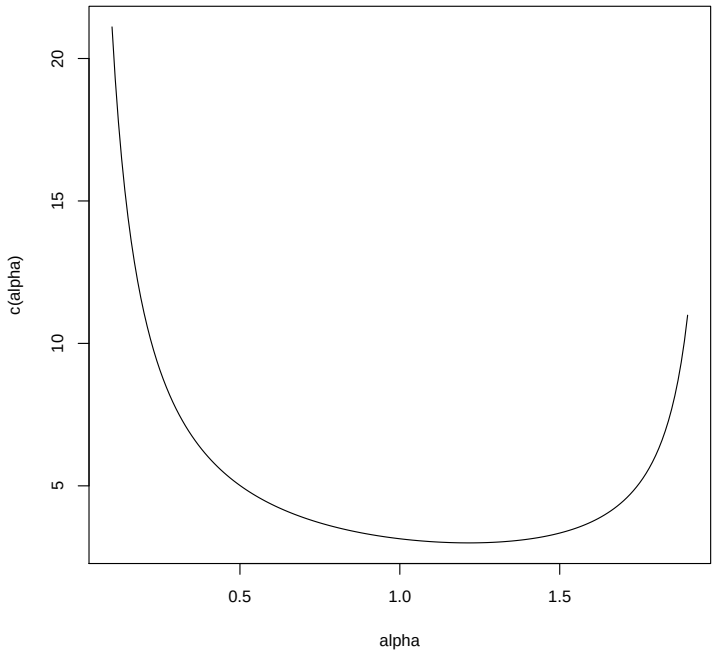

Figure 3.1: A plot of the $c(p)$ for $0 < p < 2$

where $c(p)$ is the same constant as earlier. $g(x)$ is an even function; $g(x) = g(-x)$.

$$\int_{-\infty}^{\infty} \frac{1-\cos(-xt)}{|t|^{(p+1)}} dt = \int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|t|^{(p+1)}} dt$$

It follows that

$$g(-x) = x^p c(p)$$

and we get

$$|x|^p = \frac{1}{c(p)} g(x).$$

□

The characteristic functions of a random variable x, with distribution function $F_X(x)$, is:

$$\begin{aligned} \varphi(t) = Ee^{ixt} &= \int_{-\infty}^{\infty} e^{ixt} dF_X(x) \\ &= \int_{-\infty}^{\infty} [\cos(xt) + i\sin(xt)] dF_X(x) \end{aligned}$$

which means that

$$\Re\varphi(t) = \int_{-\infty}^{\infty} \cos(xt) dF_X(x)$$

3.1.2 The Mean LP Error (MLPE)

Now we can go on to find the explicit formula for the MLPE.

Theorem 1. *Let X be a random variable with a finite absolute moment of order p, $0 < p < 2$. Then*

$$E|X|^p = \frac{1}{c(p)} \int_{-\infty}^{\infty} \frac{1-\Re\varphi(t)}{|t|^{(p+1)}} dt$$

where $\Re\varphi(t) = \int_{-\infty}^{\infty} \cos(xt) dF_X(x)$ *and* $c(p) = \frac{-\pi}{\Gamma(p+1)\cos(\frac{(p+1)\pi}{2})}$

Proof.

$$\begin{aligned}
E|X|^p &= \int_{-\infty}^{\infty} |X|^p dF_X(x) \\
&= \int_{-\infty}^{\infty} \frac{1}{c(p)} g(x) dF_X(x) \\
&= \frac{1}{c(p)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1-\cos(xt)}{|t|^{(p+1)}} dt dF_X(x) \\
&= \frac{1}{c(p)} \int_{-\infty}^{\infty} \frac{1}{|t|^{(p+1)}} \int_{-\infty}^{\infty} [1-\cos(xt)] dF_X(x) dt \\
&= \frac{1}{c(p)} \int_{-\infty}^{\infty} \frac{1-\Re\varphi(t)}{|t|^{(p+1)}} dt
\end{aligned}$$

□

In lemma 2 we state the characteristic function of the difference between the kernel density estimate, $\hat{f}_n(x;h)$, and the target function, $f_X(x)$.

Lemma 2. *Define $Z \equiv \hat{f}(x,h) - f_X(x)$. The characteristic function of Z will then be:*

$$\varphi_Z(t) = e^{-itf_X(x)} \left[\int_{-\infty}^{\infty} e^{\frac{it}{nh} K(\frac{x-y}{h})} f_X(y) dy \right]^n$$

Proof.

$$\begin{aligned}
\varphi_Z(t) &= Ee^{it(\hat{f}(x,h)-f_X(x))} \\
&= e^{-itf_X(x)}Ee^{\frac{it}{nh}\sum_{i=1}^{n}K(\frac{x-X_i}{h})} \\
&= e^{-itf_X(x)}E\prod_{i=1}^{n}e^{\frac{it}{nh}K(\frac{x-X_i}{h})} \\
&= e^{-itf_X(x)}\prod_{i=1}^{n}Ee^{\frac{it}{nh}K(\frac{x-X_i}{h})} \\
&= e^{-itf_X(x)}\left[Ee^{\frac{it}{nh}K(\frac{x-X}{h})}\right]^n \\
&= e^{-itf_X(x)}\left[\int_{-\infty}^{\infty}e^{\frac{it}{nh}K(\frac{x-y}{h})}f_X(y)dy\right]^n
\end{aligned}$$

□

Finally we can find the mean LP error of the kernel density estimator.

Theorem 2. *Let $|Z|^p = |\hat{f}(x,h) - f_X(x)|^p$, then the expectation of $|Z|^p$ is:*

$$E|\hat{f}(x,h) - f_X(x)|^p = \frac{1}{c(p)}\int_{-\infty}^{\infty}\frac{1}{|t|^{(p+1)}}\left[1-\Re\left(e^{-itf_X(x)}\left[\int_{-\infty}^{\infty}e^{\frac{it}{nh}K(\frac{x-y}{h})}f_X(y)dy\right]^n\right)\right]dt$$

Proof.

$$\begin{aligned}
E|\hat{f}(x,h) - f_X(x)|^p &= \frac{1}{c(p)}\int_{-\infty}^{\infty}\frac{1-\Re\varphi(t)}{|t|^{(p+1)}}dt \\
&= \frac{1}{c(p)}\int_{-\infty}^{\infty}\frac{1-\Re\left(e^{-itf_X(x)}\left[\int_{-\infty}^{\infty}e^{\frac{it}{nh}K(\frac{x-y}{h})}f_X(y)dy\right]^n\right)}{|t|^{(p+1)}}dt \\
&= \frac{1}{c(p)}\int_{-\infty}^{\infty}\frac{1}{|t|^{(p+1)}}\left[1-\Re\left(e^{-itf_X(x)}\left[\int_{-\infty}^{\infty}e^{\frac{it}{nh}K(\frac{x-y}{h})}f_X(y)dy\right]^n\right)\right]dt
\end{aligned}$$

□

3.1.3 The Mean Integrated LP Error

To find a global error measure, we integrate the point error measure over the desired interval, i.e. integration over $(-\infty, \infty)$.

Theorem 3. *Let $|Z|^p = |\hat{f}(x,h) - f_X(x)|^p$, then will the expected integrated value of this be:*

$$\int_{-\infty}^{\infty}E|\hat{f}(x,h) - f_X(x)|^p = \frac{1}{c(p)}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\left[1-\Re\left(e^{-itf_X(x)}\left[\int_{-\infty}^{\infty}e^{\frac{it}{nh}K(\frac{x-y}{h})}f_X(y)dy\right]^n\right)\right]}{|t|^{(p+1)}}dtdx$$

Proof. This follows directly from theorem 2 after integrating with respect to x. □

3.2 Calculation of MAE and MIAE

We now have explicit formulas for the MAE and the MIAE, by setting $p = 1$, of the kernel density estimator. In order to determine these values we need to do some fairly complicated integrations. We need to solve the integrals in theorem 2 and theorem 3 numerically. We will start by analyzing the MAE and then later move on to the MIAE. All the numerical techniques and methods used in this section are described in appendix A. These methods are essential in our discussion below, and a good understanding of these methods will be beneficial in the following.

Throughout this section we will work under the following assumption.

- Our kernel function, $K_h(x) = \frac{1}{h}K(\frac{x}{h})$, will be the standard normal density.
- Our target function, $f_X(y)$, will always be one of the 15 MarronWand densities from Marron and Wand (1992), [19]. These densities are defined in table 2.1 on page 7.

3.2.1 The Mean Absolute Error

To ease our notation we will define some functions that are included in theorem 2.

$$\begin{aligned}
g(y,t,x,h,n) &\equiv e^{\frac{it}{n}K_h(x-y)}f_X(y) \\
&= \cos\left(\frac{t}{n}K_h(x-y)\right)f_X(y) + i\sin\left(\frac{t}{n}K_h(x-y)\right)f_X(y) \\
&\equiv a(y,t,x,h,n) + ib(y,t,x,h,n) \\
w(y,t,x,h,n) &\equiv \frac{t}{n}K_h(x-y) \\
\varphi(t,x,h,n)) &\equiv e^{-itf_X(x)}\left[\int_{-\infty}^{\infty} g(y,t,x,h,n)dy\right]^n \\
f(t,x,h,n) &\equiv \frac{1-\Re(\varphi(t,x,h,n))}{t^2} \\
MAE(x,h,n) &= \int_{-\infty}^{\infty} f(t,x,h,n)dt
\end{aligned}$$

Before we start a numerical integration, it is important to understand the behavior of the functions involved. We will start by doing some general observations regarding the functions defined above. The kernel function, $K_h(\cdot)$, is a part of $w(y,t,x,h,n)$.

$$\lim_{y\to\pm\infty} K_h(x-y) = 0$$

with a convergence rate of e^{-Cy^2}, where C is a constant. Now let y and t be linear functions of s; i.e. $y(s) = y_0 \cdot s$ and $t(s) = t_0 \cdot s$, where $|y_0| < \infty$ and $|t_0| < \infty$. Then it follows that

$$\lim_{s\to\pm\infty} \frac{t}{n}K_h(x-y) = 0.$$

It follows directly that

$$\lim_{s\to\pm\infty} \cos\left(\frac{t}{n}K_h(x-y)\right) f_X(y) \to f_X(y)$$
$$\lim_{s\to\pm\infty} \sin\left(\frac{t}{n}K_h(x-y)\right) f_X(y) \to 0$$

since

$$\lim_{s\to\pm\infty} \cos(\frac{t}{n}K_h(x-y)) = 1$$
$$\lim_{s\to\pm\infty} \sin(\frac{t}{n}K_h(x-y)) = 0.$$

The h and n will be constants during the integration process, and not effect our limit considerations above.

The magnitude of t/n will determine the frequency of the trigonometric functions above, $\cos(\cdot)$ and $\sin(\cdot)$. When $t \gg n$, we will have a large frequency; the trigonometric functions will oscillate with a high frequency.

Our bandwidth h will control the standard deviation of our kernel function; a decrease in h will result in a faster convergence to zero for $K_h(x)$. That is to say, the rate of convergence of $w(y,t,x,h,n)$ will to some degree depend on the magnitude of the bandwidth.

For our target densities we assume that:

$$f_X(x) = 0 \ \forall |x| > 6.$$

This is not true in general, but our target densities are constructed in such a way that they meet this requirement. It is easy to find densities that invalidates this assumption, e.g. heavy tail densities.

Integration of $g(y,t,x,h,n)$

Before trying to determine the value of $\int_{-\infty}^{\infty} g(y,t,x,h,n)dy$ we split $g(y,t,x,h,n)$ into a real part and an imaginary part: $a(\cdot)$ and $b(\cdot)$, which we have defined earlier. We will integrate these functions separately. In order to choose the appropriate integration method, we need to study the involved functions in more detail than we have done so far.

The structure of the target density is of great importance. If the target density has a complicated structure, e.g. the Asymmetric Claw density in table 2.1, it complicates the numerical integration. It is clear from figure 2.1 that some of our target densities have a very complicated structure, and we need to take this into consideration when we do a numerical integration.

Our integration limits are $[-\infty,\infty]$, but we cannot use these limits in a numerical integration. We will integrate $g(y,t,x,h,n)$, with respect to y, over the interval $[-6,6]$.

These limits are easily found by noting that:

$$f_X(x) = 0 \forall\ |x| > 6$$
$$|\cos(\cdot)| \leq 1$$
$$|\sin(\cdot)| \leq 1.$$

This makes it natural to assume $a(y,t,x,h,n)$ and $b(y,t,x,h,n)$ to be zero outside $[-6,6]$.

Earlier we discussed the role of t/n in the trigonometric functions; it determines the frequency. In figure 3.2 and 3.3, at page 28 and 29, we have plotted $a(y,t,x,h,n)$ and $b(y,t,x,h,n)$ for different values of t, and at the same time kept the values of x, h and n constant. We can see that as t increases, the structure of the function gets more complicated. To do an accurate numerical integration when t/n increases in magnitude is very difficult; it will require an integration grid with an almost infinite number of points. We can use some properties of the characteristic function to see that the accuracy of the integral is not that important for large values of t. These results are from chapter 6 in Karr (1993), [14]. For a characteristic function, $\varphi(t)$, we have the following basic properties:

- $\varphi(t)$ is uniformly continuous
- $\varphi(0) = 1$
- $|\varphi(t)| \leq 1$ for all real t

These results follows immediately:

$$|1 - \Re\varphi(t)| < 2$$
$$\frac{|1 - \Re\varphi(t)|}{|t|^{p+1}} \ll 1\ \forall t \gg 1$$

We can allow some error in the numerical evaluation of $\varphi(t)$ when t is large since it in practice will have little impact on our results.

When t approaches zero, we have a curve which is as smooth as the target density. We can do a numerical integration very accurate as long as the structure of the target density is not too complicated.

The bandwidth h determines the spread of the oscillating area we find in figure 3.2 and 3.3, at pages 28 and 29. The location parameter x determines the location of the noisy area. The oscillating area is within the interval $(x-3h, x+3h)$, which can be seen in figure 3.4 at page 30. In figure 3.4 we have changed the bandwidth and the location parameters to illustrate this point; t and n are kept constant.

The important difference between $a(y,t,x,h,n)$ and $b(y,t,x,h,n)$ is that $a(y,t,x,h,n)$ converges to $f_X(x)$ when y increases, while $b(y,t,x,h,n)$ converges to 0 when y increases. If we compare the plots in figure 3.2 and 3.3 this difference in convergence is clearly illustrated. Even though there is a difference between these functions, it will not affect

our numerical integration, as both functions converge fast to 0, and we can use the same integration technique on both of them.

Adaptive quadrature is usually preferred over the Composite Simpson's rule as a method of numerical integration. In adaptive quadrature we can to some degree control the error, and potentially do fewer function evaluations. However, adaptive quadrature requires that $f^{(4)}(x)$, where $f(x)$ is the function we want to integrate, is approximately of the same magnitude over the whole integration interval. It is clear from our plots in figure 3.2 and 3.3 that this is not the case. To numerically integrate $a(y,t,x,h,n)$ and $b(y,t,x,h,n)$ we will use the Composite Simpson's rule.

The number of subintervals in the Composite Simpson's rule should be chosen so it reflects the structure of the target function, and to some degree the magnitude of t. A complicated structure requires a large number of subintervals in order to obtain a good approximation. The larger t gets, the larger number of subintervals is required to keep the approximation error at a reasonable level. This should be taken into consideration when we implement the Composite Simpson's Rule in algorithm 3 .

Integration of $f(t,x,h,n)$

Above we have described how we can evaluate $\int g(y,t,x,h,n)dx$. If this is done accurate, it is straight forward to determine the values of $\varphi(t,x,h,n)$ and $f(t,x,h,n)$.

However, we see that there is a problem in $f(t,x,h,n)$ when $t=0$. This problem is overcome by doing a Taylor expansion for $\varphi(t,x,h,n)$.

$$\Re\varphi_X(t) = \Re Ee^{itX} = cos(E(tX)) = 1 - \frac{E(tX)^2}{2} + ...$$

If we put this into the function $f(t,x,h,n)$ we get

$$\int_{-\epsilon}^{\epsilon} \frac{1-\Re\varphi_X(t)}{t^2}dt \approx \int_{-\epsilon}^{\epsilon} \frac{1-1+\frac{E(tX)^2}{2}}{t^2}dt = \int_{-\epsilon}^{\epsilon} \frac{1}{2}EX^2dt,$$

which is valid in a close neighborhood of $t=0$. EX^2 will in kernel density estimation be the well known MSE. The MSE of a kernel density estimator can be found in chapter 2.3 in Wand and Jones 1995, [26].

$$\begin{aligned} MSE\left(\hat{f}_n(x;h)\right) = n^{-1}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\} \\ + \{(K_h * f)^2(x) - f(x)^2\} \end{aligned}$$

The first term is the variance and the second term is the bias. When we know the target denisty and the kernel, it is easy to calculate the MSE.

Further we have a similar problem as encountered earlier, we cannot integrate over the interval $[-\infty,\infty]$. In this situation it is quite difficult to decide the integration limits, since $f(t,x,h,n)$ converges slowly to zero as $|t| \to \infty$. Through a study of different graphical representations of $f(t,x,h,n)$ and comparing the calculated MAE

with simulated values of MAE, it seems to be more than sufficient to integrate over $[-10.000, 10.000]$.

It is hard to say anything about the magnitude of $f^{(4)}(t)$ over the integration region, and therefore it is hard to say whether or not an adaptive quadrature method is appropriate. But if we plot $f(t, x, h, n)$ for different values of the parameters, it is a smooth function when $|t| > \epsilon > 0$. This is natural since $\varphi(t, x, h, n)$ is a uniformly continuos function, and therefore $f(t, x, h, n)$ will also be a continuous function in this interval. Our practical experience is that the adaptive quadrature approach works excellent for this integral.

If we implement an adaptive quadrature for the evaluation of $\int f(t, ...)dt$, we can find the MAE(x, h, n).

3.2.2 The Mean Integrated Absolute Error

To find the MIAE we need to determine the value of:

$$MIAE(n, h) = \int_{-\infty}^{\infty} MAE(x, h, n)dx.$$

It is obvious that the MAE(x, h, n) is larger in areas where $f_X(x)$ changes rapidly, than in areas where the $f_X(x)$ is smoother. So we can already conclude that for "complicated structure" target densities, the $\frac{d^4}{dx^4}MAE(x, h, n)$ is far from constant over our integration interval, and an adaptive quadrature is not appropriate. For nice target density, as the standard normal density, adaptive quadrature is successful when the bandwidth is not too far from the the optimal MIAE bandwidth. As the adaptive quadrature often fails, we will not go into details of this approach here, and again we will use the Composite Simpson's rule.

To compute the MIAE can be quite computer intensive, as it may require many function evaluations of $g(y, t, x, h, n)$ and $f(t, x, h, n)$. We again need to do an analysis of the problem before we start our numerical integration.

First we return to the assumption that

$$f_X(x) = 0 \ \forall |x| > 6,$$

which makes it natural to only consider x values in the interval $[-6, 6]$.

Further it is natural that the smoother the target density is, the fewer subintervals we need in the evaluation of MIAE. Some densities are easier to estimate than others, and for these densities we expect the MAE to be little, and fairly evenly distributed over the integration interval.

We can divide our integration interval into subintervals. Remember that our target densities are constructed such that $f_X(x) \approx 0$ when $|x| > 3$. Therefore we can expect the MAE to be very little outside $[-3, 3]$, and we can use fewer subintervals in the Composite Simpson's rule in $[-6, 3]$ and $[3, 6]$ than in $[-3, 3]$.

To find the MIAE(h,n), we will integrate MAE(x,h,n) over the intervals, $[-6,-3]$, $[-3,3]$ and $[3,6]$, where we choose the number of subintervals in each integration interval depending on the structure of the target function.

3.2.3 Results

After taking the considerations above into account, we have implemented the numerical integration methods on a computer and done the numerical integration. In table 3.1 we have found the optimal MIAE bandwidth for the 15 target densities, see chapter 2.2 for definitions and figures.

In table 3.1 the sample size is set to 100. In addition have we calculated h_{MISE}, as it is done in Marron and Wand (1992), [19], for easy comparison with a familiar bandwidth. The MIAE optimal bandwidth can also be found through a simulation study, see Kile (2009) [15], and we have reported this bandwidth as well to verify our numerical calculations. The minimizers of the MIAE and the MISE are found by searching over a grid of h values, ranging from 0.03 to 0.90. When we do a grid search, we avoid the possible pitfall of local minima we might end up in if we use a minimization algorithm like Newton's Methods.

Table 3.1: Optimal bandwidths, $n = 100$

Density	h_{MIAE}	h_{MISE}	Simulated h_{mIAE}
1	0.42	0.45	0.43
2	0.30	0.31	0.30
3	0.12	0.08	0.12
4	0.10	0.08	0.10
5	0.05	0.05	0.05
6	0.36	0.39	0.36
7	0.25	0.26	0.25
8	0.31	0.32	0.32
9	0.33	0.36	0.33
10	0.10	0.10	0.10
11	0.36	0.39	0.36
12	0.21	0.20	0.21
13	0.34	0.36	0.34
14	0.14	0.14	0.14
15	0.15	0.16	0.15

Further we have plotted the MIAE(h) and MISE(h) for each target density in figure 3.5, 3.6, 3.7 and 3.8 . The MIAE(h) follows the scale on the left, and the MISE(h) follows the scale to the right. The vertical lines are the corresponding optimal bandwidths.

3.2.4 Other values of p

After implementing algorithms to solve MAE and MIAE, it is quite easy to modify them to calculate MLPE and MILPE. We have done this, but the MILPE optimal bandwidth results in basically the same bandwidths as the MIAE optimal bandwidth. In appendix B we have included some tables and plots of these values.

It seems the MILPE converges to the MISE as $p \to 2$, and diverges from MISE when $p \to 0$. This is quite natural, and is illustrated in the plots in the appendix.

Perhaps a more thorough investigation of the MILPE could reveal interesting results, but this is outside the scope of this report.

3.3 Summary

We have seen that the MIAE and MISE optimal density estimates are essentially equal. In other words, the optimal MIAE bandwidth is no better than the MISE optimal bandwidth for limited sample sizes. But our new formula might lead to a new data-driven bandwidth selectors that might outperform today's bandwidth selectors, and in that way give a new dimension to bandwidth selection.

There is an old controversy between which of the MISE and MIAE is the better error criterion in kernel density estimation. Our results indicate that these are essentially equal, and for further development of kernel density estimation, we should consider looking at other error criteria, like the visual error criterion.

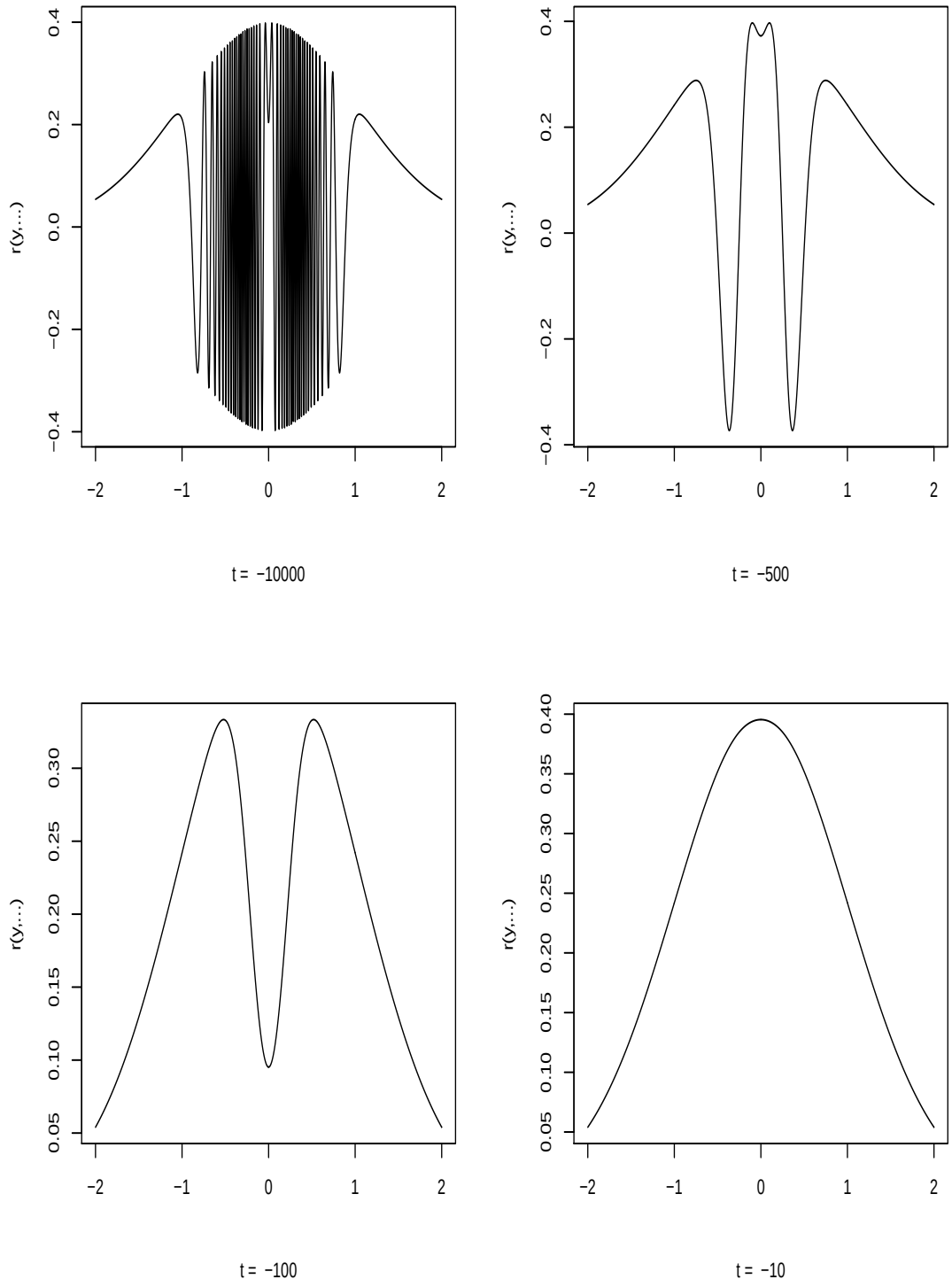

Figure 3.2: Plot of $a(y, t, x, h, n)$ for different values of t, when h=0.30, x=0, n=100. The target function is the standard normal density.

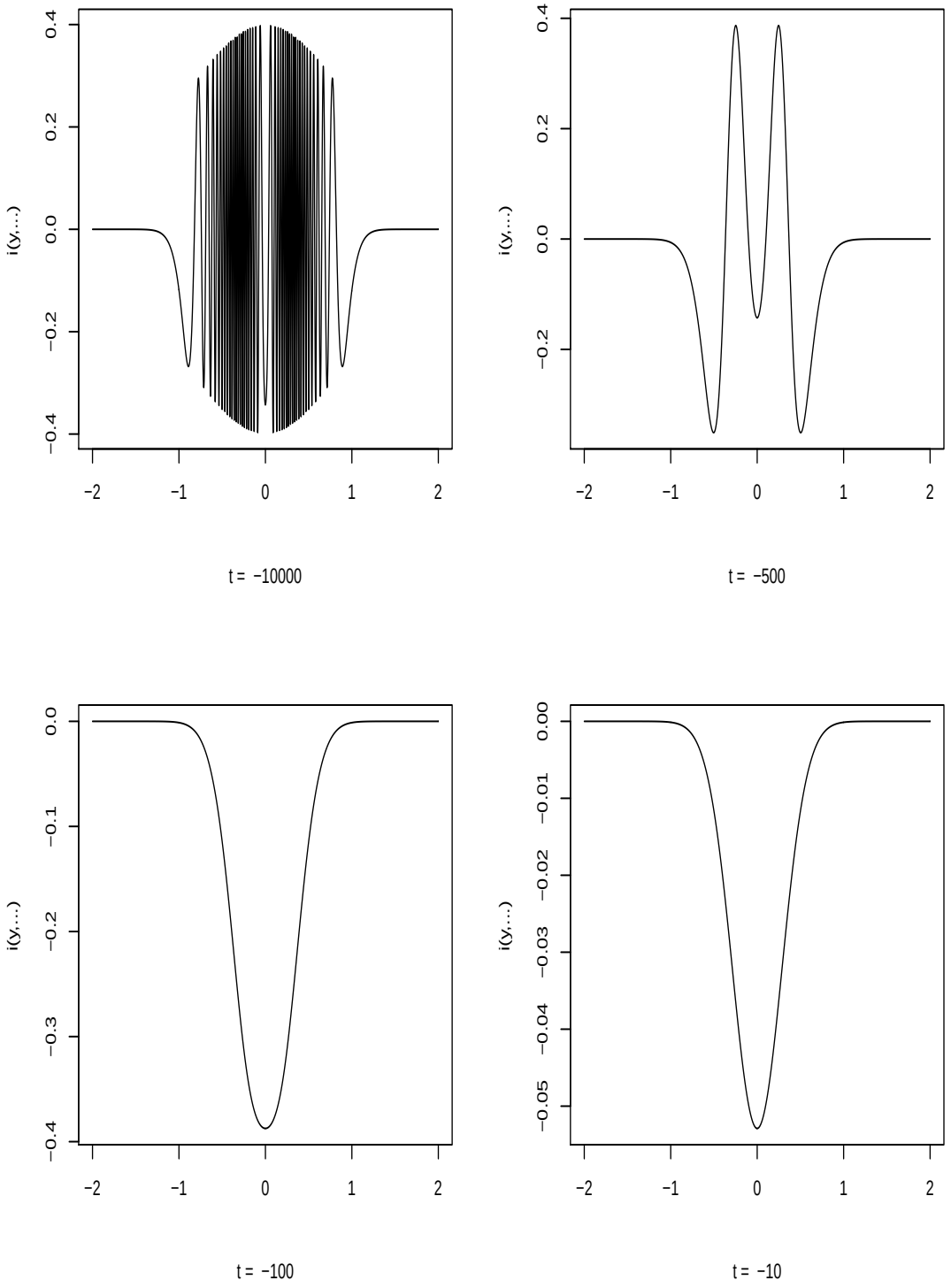

Figure 3.3: Plot of $b(y, t, x, h, n)$ for different values of tm when h=0.30, x=0, n=100. The target function is the standard normal density.

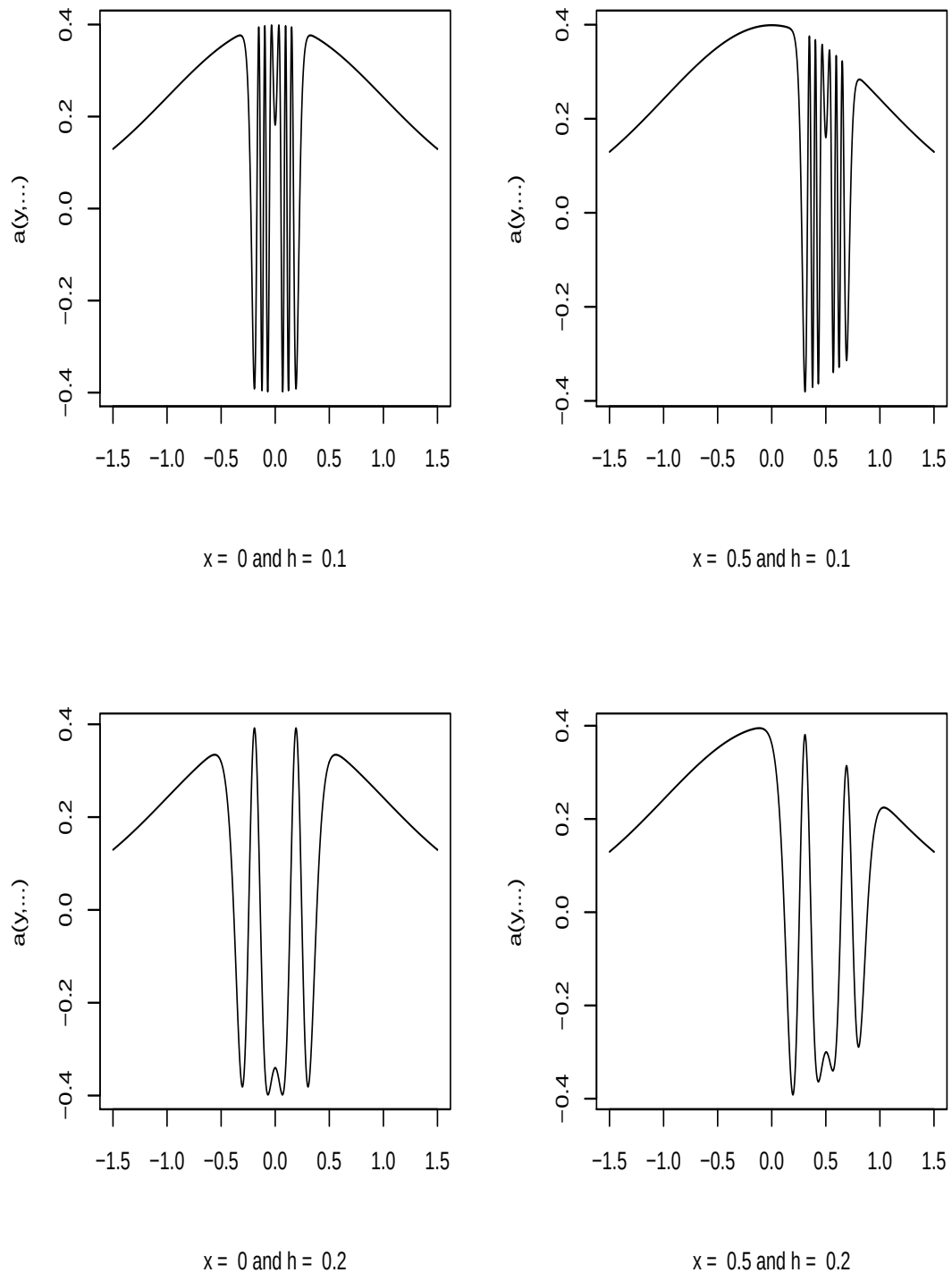

Figure 3.4: Plot of $a(y, t, x, h, n)$ for different values of h and x. t and n keep are kept constant; t = -500 and n = 100. The target function is the standard normal density.

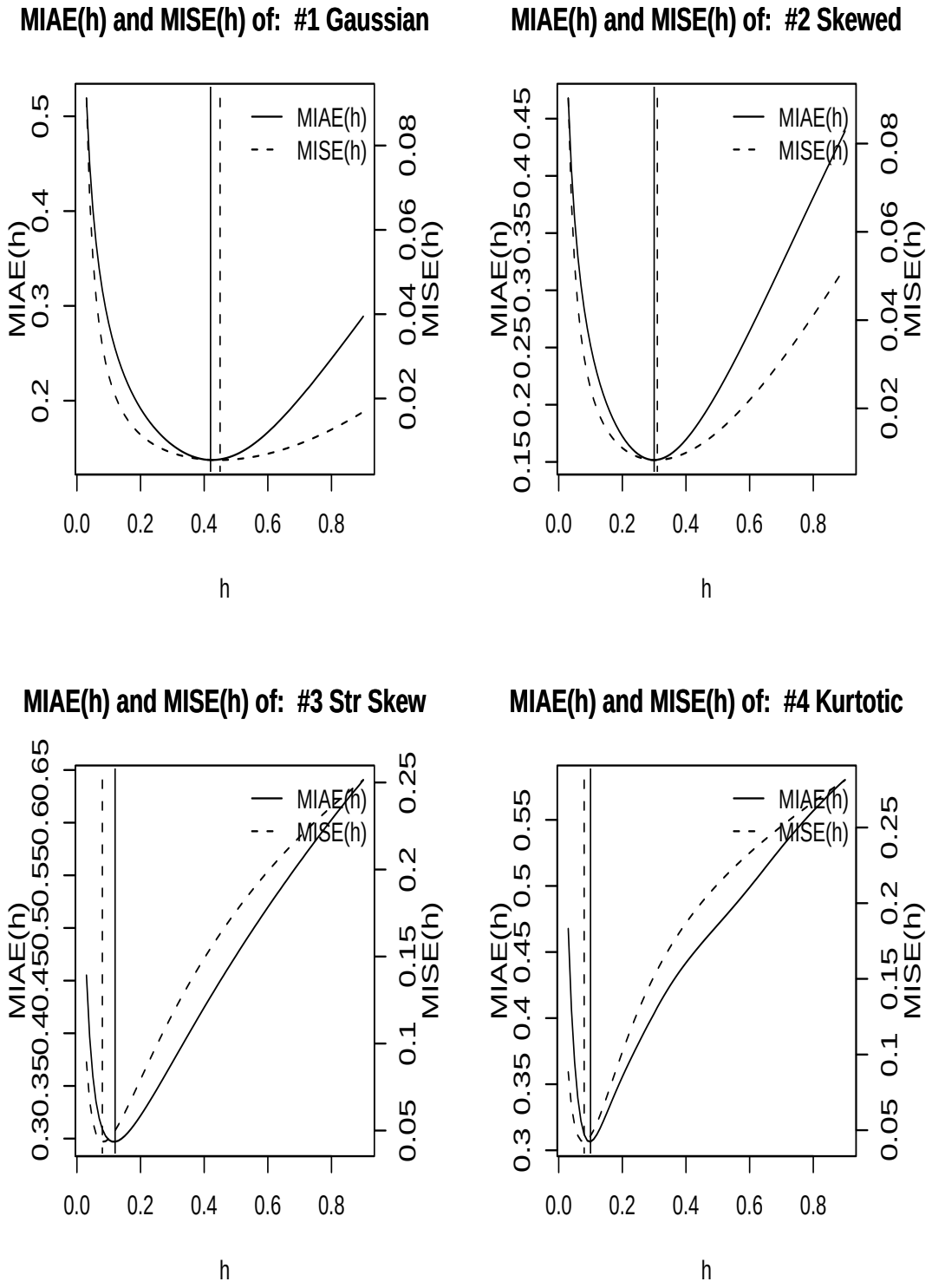

Figure 3.5: MIAE(h) and MISE(h) of the target densities from #1 to #4. The sample size is 100.

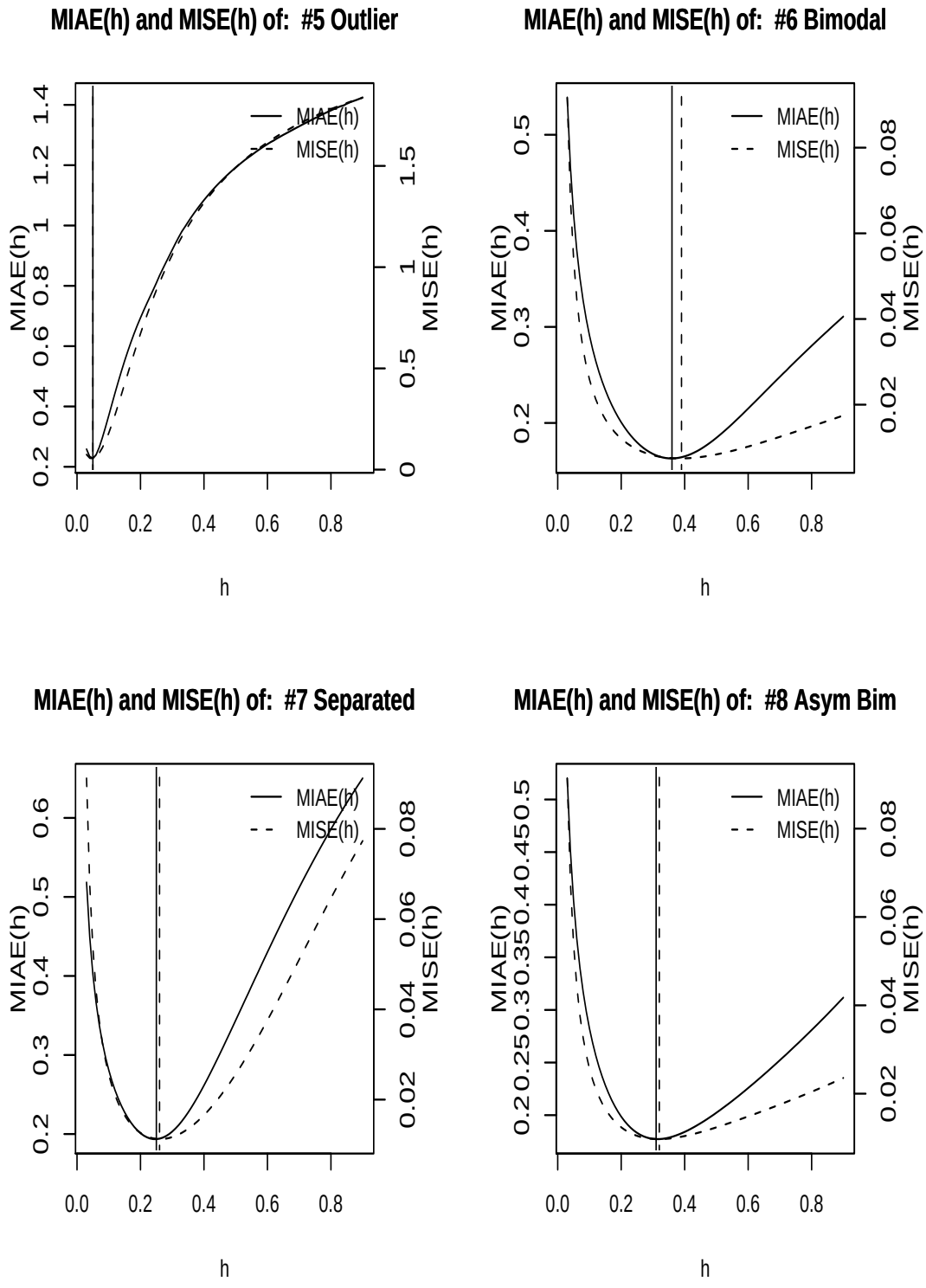

Figure 3.6: MIAE(h) and MISE(h) of the target densities from #5 to #8. The sample size is 100.

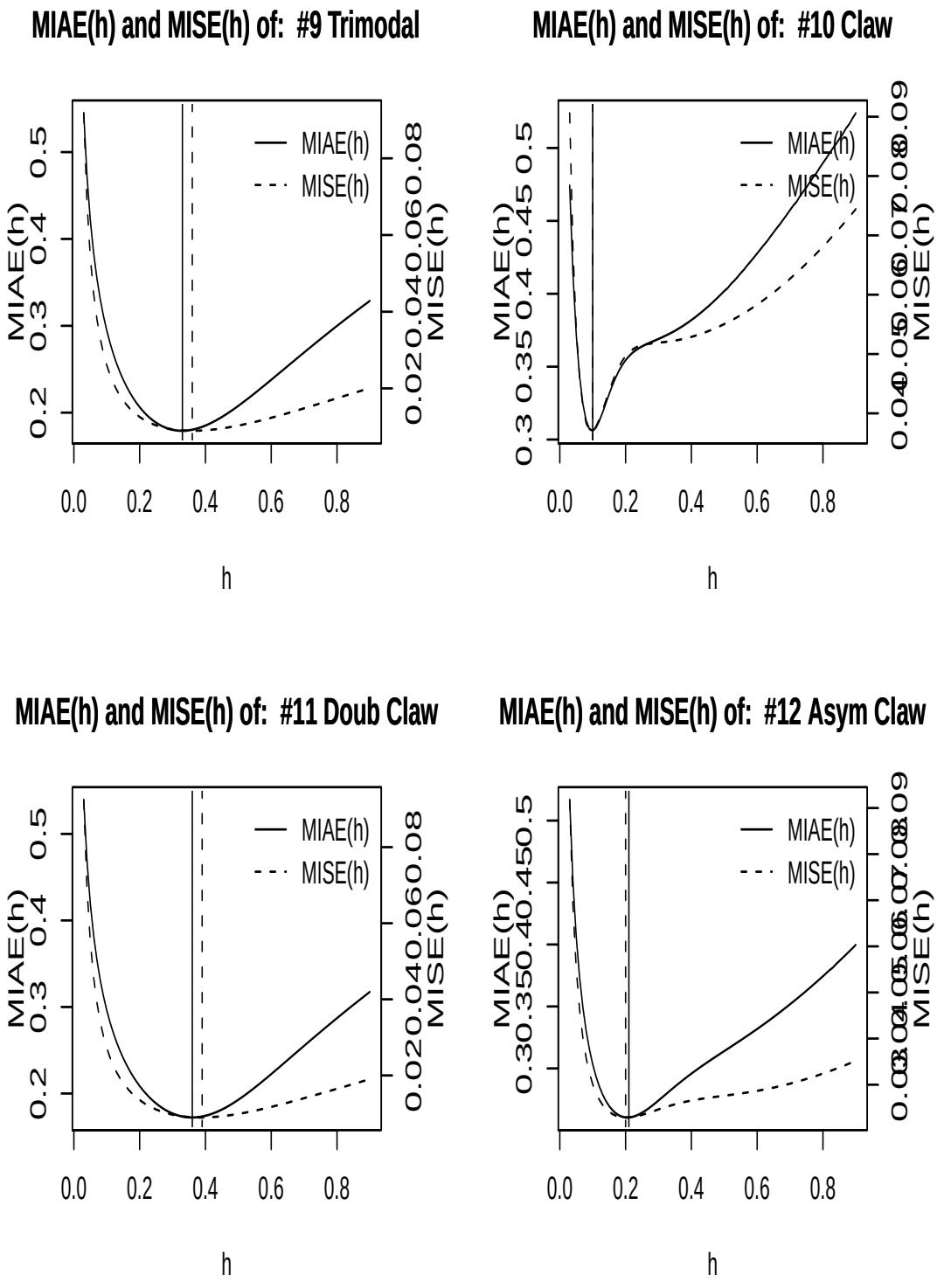

Figure 3.7: MIAE(h) and MISE(h) of the target densities from #9 to #12. The sample size is 100.

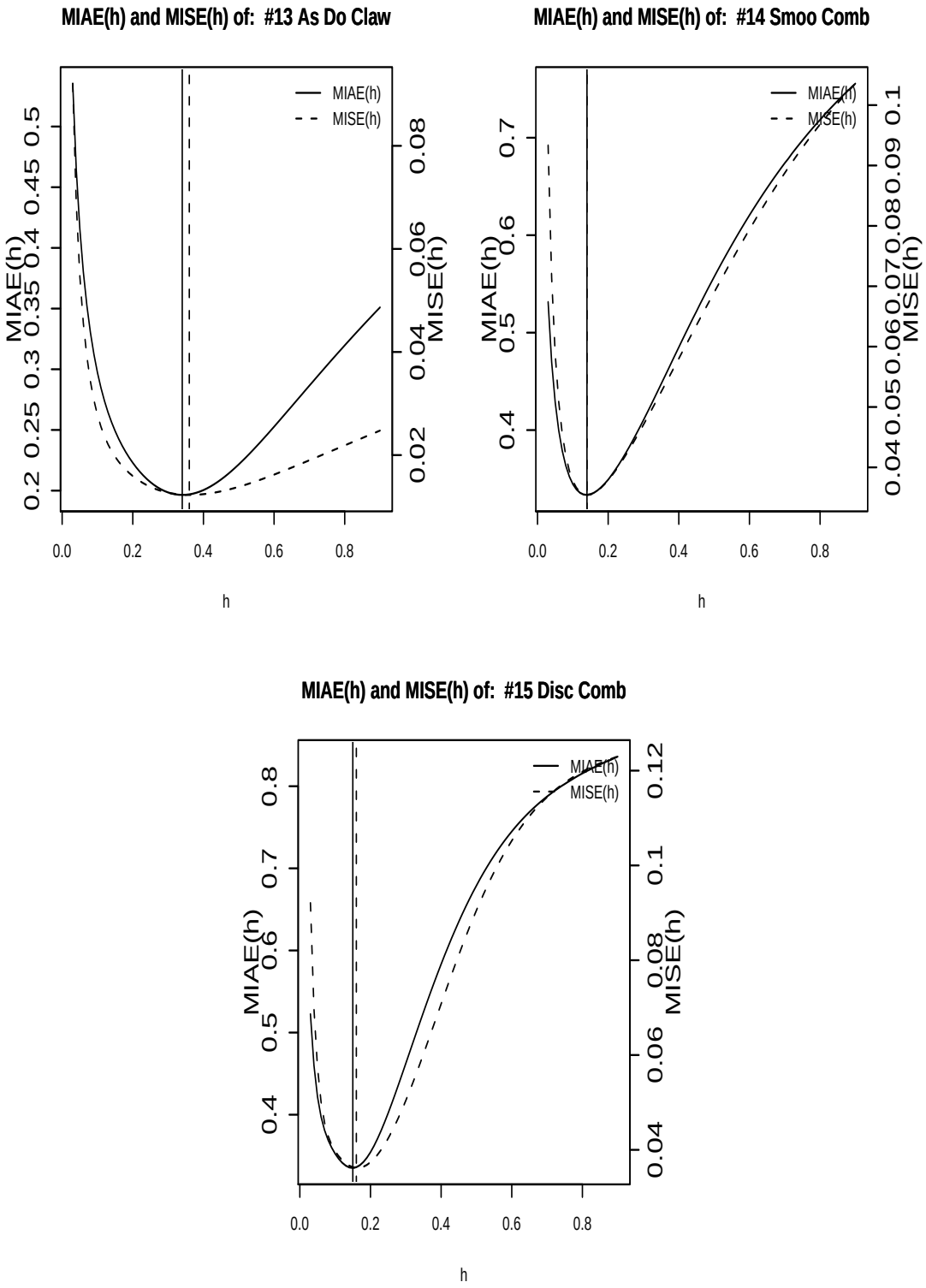

Figure 3.8: MIAE(h) and MISE(h) of the target densities from #13 to #15. The sample size is 100.

Chapter 4

Bandwidth Selection

In practice it is not possible to find the exact minimizers of the error criteria discussed in section 2.3, since the target density is unknown. Data-driven bandwidth selectors try to estimate the minimizers by using different techniques, e.g. cross validation and asymptotic approximation. There exists several different bandwidth selectors, but all of them have their limitations. In this chapter we suggest a new data-driven bandwidth selector, which is a modified version of the direct-plug in selector from Hall and Marron (1987), [8], and Sheather and Jones (1991), [22]. We will also use some results from Jones and Sheather (1991), [13]. Our new estimator will produces smoother density estimates, and in some situations outperform the direct plug-in selector.

4.1 Data-driven bandwidth selection

We have previously discussed different aspects of bandwidth selection, and some of the limitations of today's methods. The best bandwidth selection method is by visual assessment, but this can be very time consuming. Today's data-driven bandwidth selectors cannot replace the visual assessment, but can provide useful information. For an inexperienced analyzer it can provide a final bandwidth, while for an experienced analyzer it might be a good starting point for further analysis. The same can be said about the method of local varying bandwidth; we need a starting point, and data-driven bandwidth selectors can provide that.

In section 2.3 we mentioned some of the data-driven bandwidth selectors that aim to minimize both the MISE and the MIAE. For extensive discussions and simulation studies, we refer to other sources.

- For a simulation study of different bandwidth selectors, see Cao, Cuevas and Manteiga (1994), [3].
- For both asymptotic rate of convergence of bandwidth selectors and a simulation study, see Jones, Marron and Sheather (1996), [11] and [12].

- In Marron (1998), [17], different bandwidth selectors are compared with respect to their EVE2 performance.
- In chapter 3 in Wand and Jones (1995), [26], there is a thorough discussion of several different bandwidth selectors.

The all around favorite from the above sources is the estimator of Sheather and Jones (1991), from here on out denoted DPI; short for direct plug-in. But there is still room for improvement; especially for limited sample sizes, and improved performance in terms of the VE criterion. Maybe an approach to minimize the MIAE formula in chapter 3 will be more successful than the current bandwidth selectors..

4.2 Optimal EVE2 bandwidth

For limited sample sizes the h_{MISE} and h_{MIAE} will often result in undersmoothed density estimates. Often twice the MISE optimal bandwidth is a better choice when the resulting density estimates are compared visually. This of course, depends on the structure of the target density; is it an "easy to estimate" or "hard to estimate" density? Especially in the class of "hard to estimate" densities, the inadequacy of the MISE and the MIAE is clear.

To see that the VE optimal bandwidths are often different from the MISE and the MIAE optimal bandwidths for limited sample sizes, we do a simulations study to find the EVE2 optimal bandwidths. If we add these bandwidths to table 3.1, we get table 4.1. We can see that the EVE2 optimal bandwidths are larger than the MISE and the MIAE optimal bandwidths, i.e. the EVE2 optimal estimates are smoother than the MISE and MIAE optimal estimates. We can also see a plot of the MISE versus the EVE2 in figure 4.1.

To save space we have not plotted the resulting density estimates, but some examples can be found in appendix B.

We have essentially done the same simulation study as in Marron (1998), [17], but our optimal bandwidths seems to differ a little from what is obtained in that paper. The reason might be that we have chosen to do our simulations on a finer grid than Marron; we have 1200 points versus Marron's 400. We have also chosen do to it over a larger x grid, $[-5, 5]$, and to use Composite Simpson's Rule, instead of Riemann approximation, to numerically approximate the integral in (2.10). But our bandwidths are essentially equal, and small differences in the bandwidths are not important.

4.3 Density Functionals and Direct plug-in

Many of the modern data-driven bandwidth selectors are based upon an estimate of the integrated squared density derivative. A squared density derivative functional of order s is defined as:

$$\psi_{2s} = \int [f^{(s)}(x)]^2 dx = (-1)^s \int f^{(2s)}(x) f(x) dx.$$

Table 4.1: Optimal bandwidths, $n = 100$

Density	h_{MIAE}	h_{MISE}	h_{EVE2}
1	0.42	0.45	0.48
2	0.30	0.31	0.35
3	0.12	0.08	0.30
4	0.10	0.08	0.29
5	0.05	0.05	0.11
6	0.36	0.39	0.47
7	0.25	0.26	0.34
8	0.31	0.32	0.47
9	0.33	0.36	0.47
10	0.10	0.10	0.29
11	0.36	0.39	0.41
12	0.21	0.20	0.32
13	0.34	0.36	0.39
14	0.14	0.14	0.30
15	0.15	0.16	0.25

In the following we let $r = 2s$.

In chapter 3.5 in Wand and Jones (1995), [26], there is a thorough discussion regarding estimation of density functionals; here we briefly summarize the results from that chapter. An estimator of ψ_r is:

$$\hat{\psi}_r(g) = n^{-1} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i; g) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} L_g^{(r)}(X_i - X_j)$$

where g is a bandwidth and $L(\cdot)$ is a kernel function. The asymptotic MSE optimal bandwidth for this estimator is:

$$g_{\text{AMSE}} = \left[\frac{k! L^r(0)}{-\mu_k(L) \psi_{r+k} n} \right]^{1/(r+k+1)}$$

where k is the order of the kernel. We can see that the optimal bandwidth depends on ψ_{r+k}, and this dependence on higher order density functionals will continue. This means that a density functional estimate will always depend on the value of a higher order density functional, which is unknown.

The normal way of solving this problem is to choose to do an l-stage procedure. We choose an initial estimate of ψ by assuming that the sample is generated from a normal density, and then use the results from chapter 3.6 in Wand and Jones (1995), [26]. If the sample is generated from a normal density, with standard deviation σ, ψ_r will be:

$$\psi_r = \frac{(-1)^{r/2} r!}{(2\sigma)^{r+1} (r/2)! \pi^{1/2}} \tag{4.1}$$

We can find an estimate $\hat{\psi}_r$, of ψ_r, by replacing σ by $\hat{\sigma}$ in (4.1). We call such an estimate the normal scale estimate, denoted $\hat{\psi}_{r,NS}$. Further we plug this estimate into the estimate of $g_{AMSE,(r-2)}$ and obtain the ψ_{r-2} estimate. We then plug this estimate into the $g_{AMSE,r-4}$ and so on. See algorithm 1.

Algorithm 1: l stage Direct-Plug in

Input: number of stages l, kernel function $K()$ of order 2, a data sample X
Output: Approximation of ψ_r

begin
  $\hat{\sigma} \leftarrow [\mathrm{Var}(X)]^{1/2}$
  $c \leftarrow r + 2l$
  $\psi_c \leftarrow \frac{(-1)^{c/2} c!}{(2\hat{\sigma})^{(c+1)}(c/2)!\pi^{1/2}}$
  $c \leftarrow c - 2$
  while $c >= r$ **do**
    $g \leftarrow \left[\frac{-2K^{(c)}(0)}{\mu_2(K)\hat{\psi_{c+2}} n}\right]^{1/(c+2+1)}$
    $\psi_c = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K_g^{(r)}(X_i - X_j)$
    $c \leftarrow c - 2$
  end
  RETURN(ψ_c)
end

To choose the number of stages is not straight forward, but Wand and Jones (1995), [26], suggest to take l to be greater or equal to 2, with the most common value to be $l = 2$. For a more thorough discussion regarding the selection of number of stages, see Chacon and Tenreiro (2009), [2].

To find the Direct plug-in bandwidth, we use the asymptotically optimal bandwidth from (2.6),

$$h_{DPI} = \left[\frac{R(K)}{\mu_2(K)\hat{\psi}_4 n}\right]^{(1/5)} \tag{4.2}$$

where we have replaced $R(f'')$ in (2.6) by $\hat{\psi}_4$.

4.4 New data-driven bandwidth selector

We can now find an asymptotical MSE optimal estimate of the integrated squared derivative density functional of arbitrary order, and the question is now how we can use this estimate in the most efficient way. We define the second order integrated derivative density functional of the kernel density estimator as the total curvature (TC). TC will

be a function of the bandwidth h.

$$TC(h) = \int \left| \hat{f}^{(2)}(x;h)) \right]^2 dx \tag{4.3}$$

We use a standard normal density as our kernel, and the TC(h) can be expressed as:

$$TC(h) = \int \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{(x-X_i)^2}{h^4} - \frac{1}{h^2} \right) K_h(x - X_i) \right]^2 dx \tag{4.4}$$

We obtain this by noting that the second derivative of our kernel density estimator is

$$\hat{f}''(x;h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{(x-X_i)^2}{h^4} - \frac{1}{h^2} \right) \cdot K_h(x - X_i)$$

This follows directly from some features of the normal density. For a normal density $N(\mu, \sigma)$, denoted as $\phi_\sigma(x, \mu)$, we have the following equalities

$$\begin{aligned}
\phi_\sigma(x,\mu) &= \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\
\phi'_\sigma(x,\mu) &= \frac{-(x-\mu)}{\sqrt{2\pi}\sigma^3} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{-(x-\mu)}{\sigma^2} \phi_\sigma(x,\mu) \\
\phi''_\sigma(x,\mu) &= \frac{-1}{\sigma^2} \phi_\sigma(x,\mu) + \frac{(x-\mu)^2}{\sigma^4} \phi_\sigma(x,\mu)
\end{aligned}$$

If we plug the h_{DPI} into (4.3), we get

$$TC(h_{DPI}) > \hat{\psi}_4$$

This is not easy to prove explicitly, but we have done a simulation study and this is always true for our 15 target densities in our simulations. We can interpret this as the h_{DPI} results in an undersmoothed density estimate compared to our estimated smoothness, $\hat{\psi}_4$. Especially for limited sample sizes, this undersmoothing is not optimal in terms of the VE.

An alternative approach to bandwidth selection is based on an idea of Ushakov and Ushakov (2009), [24]. Here they investigate the total variation of the kernel density estimator, and suggest to choose the bandwidth in such a way that the first order integrated squared density derivative of the kernel density estimate equals the estimate we can obtain for ψ_2. In this section we will use a slightly different approach, and choose the bandwidth of the kernel density estimate in such a way that the second order integrated squared density derivative will equal the estimate of ψ_4.

$$TC_{\hat{f}}(h) = \psi_4 \tag{4.5}$$

The solution of this equation is purely a numerical problem. We denote the bandwidth chosen in this way h_{NBW}.

To see the effect of the bandwidth chosen in this way, versus the DPI bandwidth, we have done a simulation study. The simulation procedure is described in algorithm 2.

Algorithm 2: Simulation study of h_{NBW} and h_{DPI}

```
begin
    sims ← Number of simulations
    for i = 1 to sims do
        X ← A sample from the target density
        ψ̂_4 ← from algorithm 1
        h_NBW ← from (4.5)
        h_DPI ← from (2.6)
        kdeFitOne ← f̂(x, h_NBW)
        kdeFitTwo ← f̂(x, h_DPI)
        Calculate VE2 and MISE for both kdeFitOne and kdeFitTwo
        Store all values
    end
end
```

4.4.1 Results

In figure 4.2 we can see the distributions of $h_{NBW} - h_{MISE}$ and $h_{DPI} - h_{MISE}$. The distributions are estimated by a kernel density estimator. The vertical line at 0 represents the h_{MISE}, while the solid vertical line is the $h_{EVE2} - h_{MISE}$. Even though the densities are not strictly separated, it is important to note that $h_{NBW} > h_{DPI}$ in all our simulations. Our new density estimate is always smoother than the DPI estimate.

In figure 4.3 we can see the visual error of the h_{NBW} and h_{DPI} estimates plotted against each other. For most densities the h_{NBW} results in a smaller VE than the DPI. In figure 4.4 we can see the ISE of h_{NBW} versus h_{DPI} . Here we can see that the DPI tends to have a lower ISE than h_{NBW}.

From our simulation study we see that the h_{NBW} is always larger than h_{DPI}, so our new approach is consistently smoother than the DPI approach for our 15 target densities. This agrees with our goal to obtain smoother estimates than the MISE optimal estimate, but our new approach leads to an oversmooth estimate compared to he EVE2 optimal estimate for some densities. This is especially clear for the densities 14 and 15, which have a very complicated structure.

A summary table of the h_{NBW} and h_{DPI} is included in the appendix in table B.2.

4.5 Conclusion

We see that our new bandwidth selector is smoother than the DPI, but it still fails in many situations with respect to VE optimal estimate. One reason might be that we use the MSE optimal bandwidth in the estimation of the ψ, which again leaves us with the problems related to the squared error. Further, the bandwidth selector is not designed to minimize the VE, but instead just uses the estimate of ψ in a different manner. Therefore, we cannot expect a VE optimal estimate in all cases, when we in fact do not

try to minimize it.

For further development in bandwidth selection one should try to develop bandwidth selectors which specifically aim to minimize the VE. Marron and Tsybakov (1995), [18], suggested to look at a bandwidth selector based on the asymptotical expansion of the VE. However, this problem is outside the scope of this report.

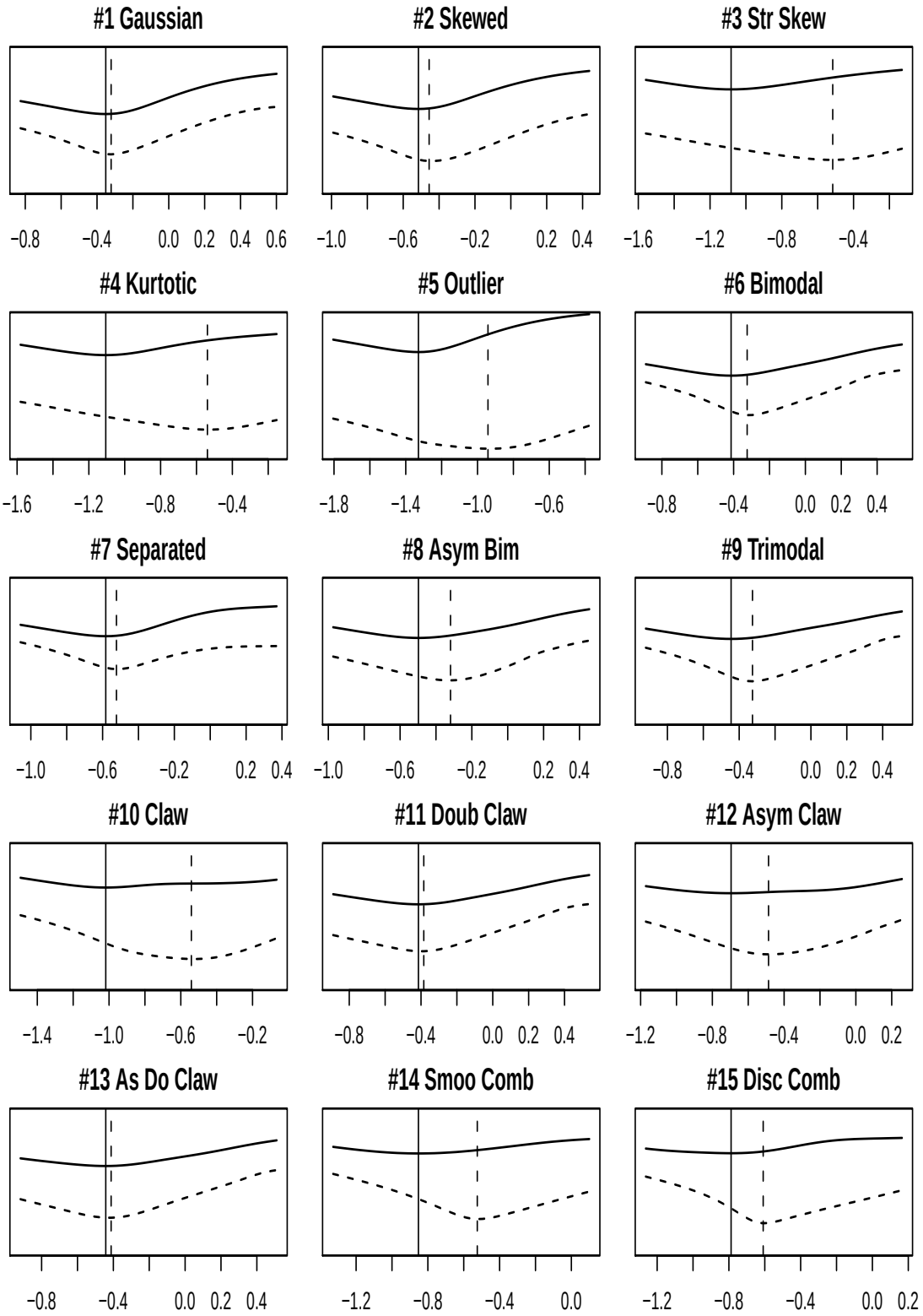

Figure 4.1: $EVE2(h)$ and $MISE(h)$ for the 15 target densities in this report plotted, on a $log10$ scale of h. Dotted lines for $EVE2(h)$ and solid line for $MISE(h)$. The vertical lines are their respective minimizers.

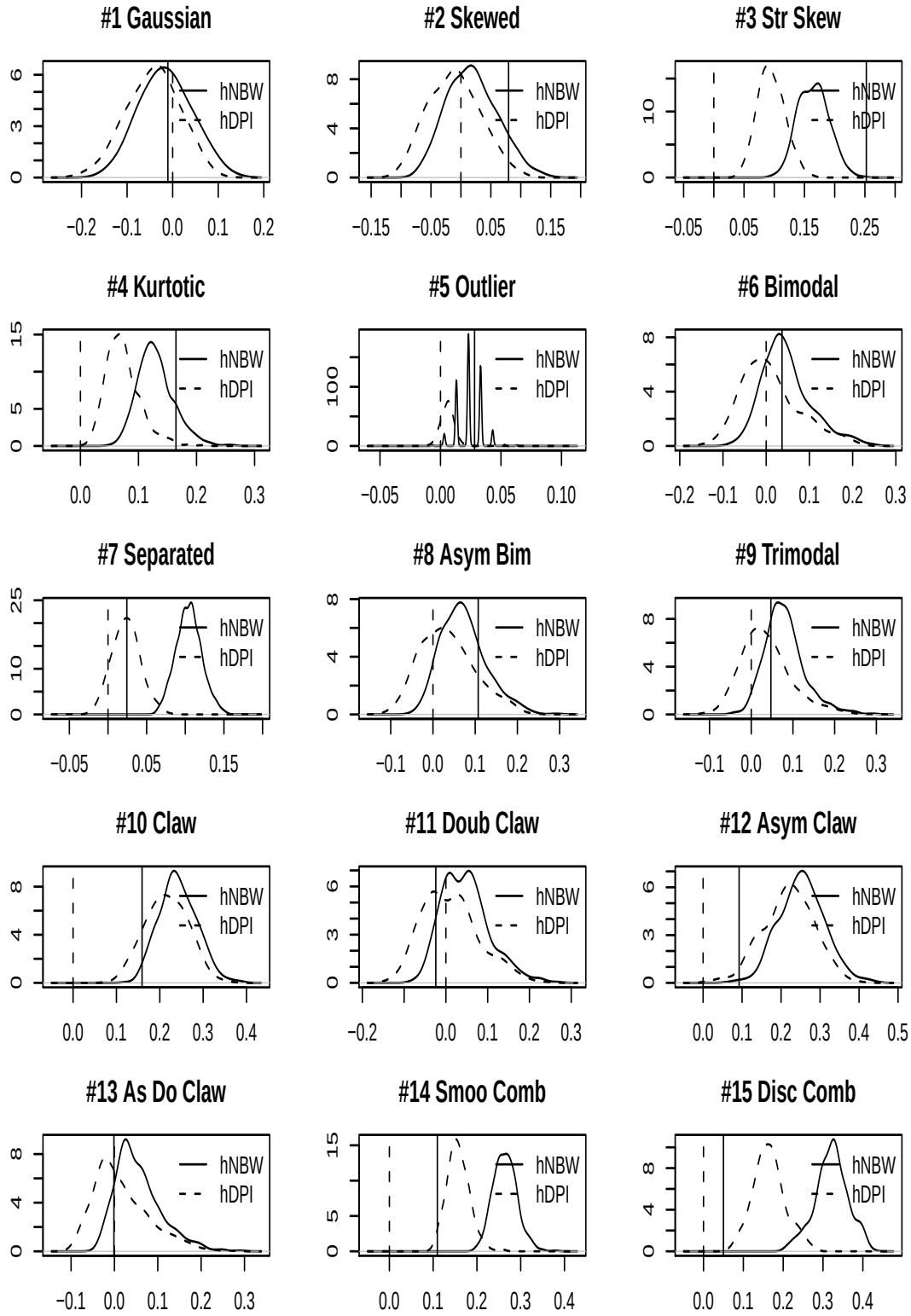

Figure 4.2: The distributions of $h_{NBW} - h_{MISE}$ and $h_{DPI} - h_{MISE}$ for $n = 100$ and 500 simulations for the 15 target densities. The vertical line at 0 is the h_{MISE} optimal bandwidth, while the solid vertical line is the $h_{EVE2} - h_{MISE}$.

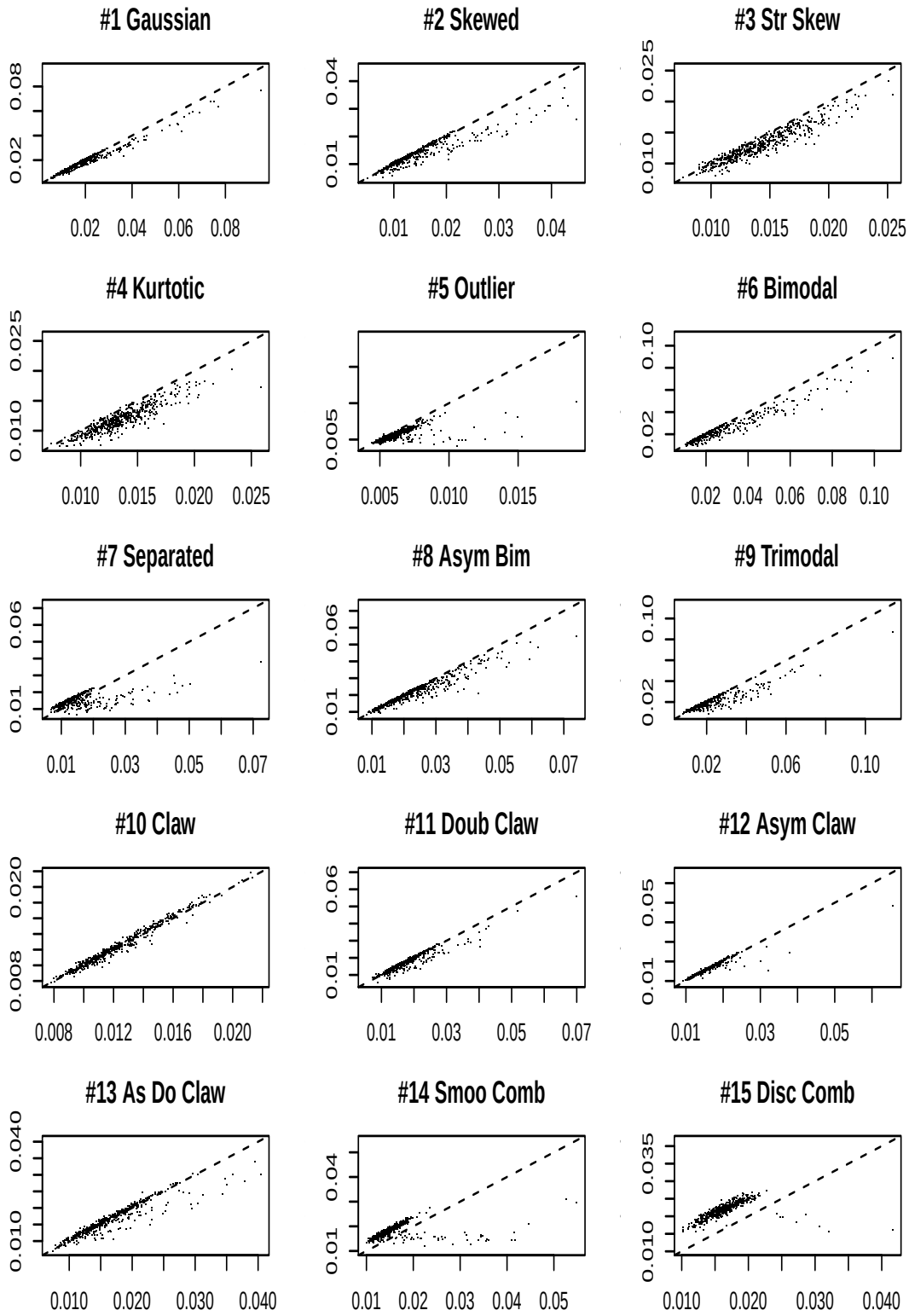

Figure 4.3: Plot of $VE(h_{NBW})$ on the y axis and $VE(h_{DPI})$ on the x axis, for the 15 target densities with $n = 100$ and 500 simulations

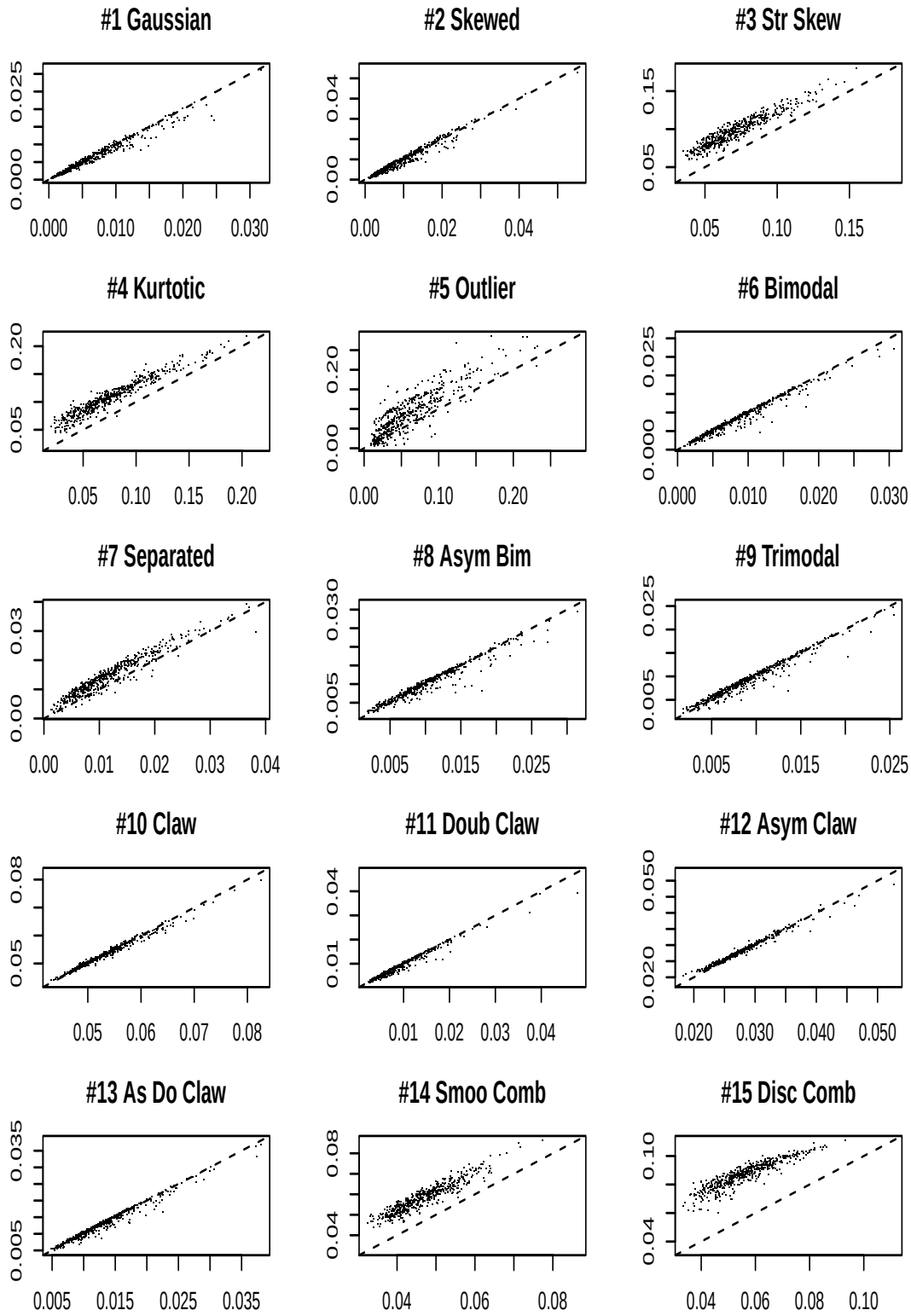

Figure 4.4: Plot of $ISE(h_{NBW})$ on the y axis and $ISE(h_{DPI})$ on the x axis, for the 15 target densities with $n = 100$ and 500 simulations

Bibliography

[1] R. L. Burden and J. D. Faires. *Numerical Analysis*. Thomson Brooks Cole, 2005.

[2] J.E. Cahcon and C. Tenreiro. A data-based methods for choosing the number of pilot stages for plug-in bandwidth selection. *Preprint Number 08-59, Universidade de Coimbra*, 2009.

[3] R. Cao, A. Cuevas, and W.G. Manteiga. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, 1994.

[4] L. Devroye. *A course in density estimation*. Birkhauser, 1987.

[5] L. Devroye. The double kernel method in density estimation. *Annales de l'Institut Henri PoincarŐ. ProbabilitŐs et Statistiques*, 1989.

[6] L. Devroye and L. Gyorfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley New York, 8th edition edition, 1985.

[7] I.K. Glad, N.L. Hjort, and N. Ushakov. Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, 2003.

[8] P. Hall and J.S. Marron. Estimation of integrated squared density derivatives. *Statistics and Probability Letters*, 1988.

[9] P. Hall and M.P. Wand. Minimizing l_1 distance in nonparametric density estimation. *Journal of Multivariate Analysis*, 1988.

[10] P. Hall and M.P. Wand. On the minimization of absolute distance in kernel density estimation. *Statistics and Probability Letters*, 1988.

[11] M.C. Jones, J.S. Marron, and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 1996.

[12] M.C. Jones, J.S. Marron, and S.J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 1996.

[13] M.C. Jones and S.J. Sheater. Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability Letters*, 1991.

[14] A.F. Karr. *Probability*. Springer Verlag, 1993.

[15] H. Kile. Bandwidth selection in univariate kernel density estimation. Specialization Project, Department of Mathematical Sciences, Norwegian University of Science and Technology, 2009.

[16] E. Mammen. On qualitative smoothness of kernel density estimates. *Statistics*, 1993.

[17] J.S. Marron. Assessing bandwidth selectors with visual error criteria. *Computational Statistics*, 1998.

[18] J.S. Marron and A.B. Tsybakov. Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association*, 1995.

[19] J.S. Marron and M.P. Wand. Exact mean integrated squared error. *Ann. Stat.*, 1992.

[20] E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 1962.

[21] D.W. Scott and M.P. Wand. Feasibility of multivariate density estimates. *Biometrika Trust*, 1991.

[22] S.J. Sheater and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, 1991.

[23] B.W. Silverman. *Density Estimation*. Champman and Hall, 1986.

[24] V.G. Ushakov and N.G. Ushakov. On the choice of smoothing parameter in kernel density estimation. *Moscow University Computational Mathematics and Cybernetics*, 2009.

[25] M.P. Wand and L. Devroye. How easy is a given density to estimate? *Computational Statistics and Data Analysis*, 1993.

[26] M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

Appendix A

Numerical methods

In this appendix we will quickly describe the numerical methods used in the previous chapters. We will not pay too much attention to the deduction of the error terms, rather just state their order, $O(\cdot)$. However, these errors are important, and for the interested reader we will in each case refer to a reference where more details can be found.

A.1 Taylor's Theorem

This is just a reminder of the well know Taylor's theorem. Here stated as in theorem 1.14 in Burden and Faires (2005), [1]:

Suppose $f \in C^{(n)}[a,b]$, that $f^{(n+1)}$ exist on $[a,b]$, and $x_0 \in [a,b]$. For every $x \in [a,b]$ there exist a number $\xi(x)$ between x_0 and x with

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}$$

A.2 Lagrange polynomial

In numerical differentiation and integration we need to fit polynomials to data represented as $(x_i, f(x_i))$ for $i = 0, 1, ..., n$. Such a procedure is called interpolation, and Taylor polynomials are not appropriate in this situation. One way of doing such an interpolation is trough Lagrange polynomials.

To understand the idea behind a Lagrange polynomial, we will first take a look at the most fundamental situation, where we fit a first degree polynomial to $(x_0, f(x_0))$ and $(x_1, f(x_1))$. The idea is to construct a polynomial, $P(x)$, which agrees with the values at each observation point; i.e. $P(x_0) = f(x_0)$ and $P(x_1) = f(x_1)$. If we construct two functions:

$$L_0(x) = \frac{x - x_1}{x_0 - x_1}$$
$$L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

it is quite easy to see that

$$P(x) = L_0(x) \cdot f(x_0) + L_1(x) \cdot f(x_1)$$

leads to $P(x_0) = f(x_0)$ and $P(x_1) = f(x_1)$.

In the general situation we want to fit a Lagrange polynomial to $(n + 1)$ distinct points: $(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n))$. We follow the same idea as above, and construct functions $L_{n,k}(x)$ such that

$$L_{n,k}(x_i) = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}$$

This is done in theorem 3.2 in Burden and Faires (2005), [1]. We get:

$$P(x) = \sum_{k=0}^{n} f(x_k) \cdot L_{n,k}(x)$$

$$L_{n,k} = \prod_{i=0, i \neq k}^{n} \frac{(x - x_i)}{(x_k - x_i)}.$$

Such a polynomial will be a nth Lagrange interpolating polynomial. For more details, see chapter 3.1 in Burden and Faires (2005), [1].

A.3 Numerical differentiation

This is a quick summary of the methods described in chapter 4.1 in Burden and Faires (2005), [1]. The first derivative of a function $f(x)$ at x_0 is defined as

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

We can use the Lagrange polynomial above to derive a five point formula for the first derivative.

$$f'(x_0) = \frac{1}{12h}[f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)] + O(h^4)$$

This formula can be found chapter 4.1 in Burden and Faires (2005), [1]. From this formula it is fairly straight forward to derive the formula for the second derivative.

$$f''(x_0) = \frac{1}{h^2}[f(x_0 - h) - 2f(x_0) + f(x_0 + h)] + O(h^2) \tag{A.1}$$

Again we will refer to chapter 4.1 in Burden and Faires (2005), [1] for a more extensive discussion of the methods and the error analysis.

A.4 Numerical Integration

Our numerical integration technique is called numerical quadrature, which means that $\int_a^b f(x)dx$ will be approximated by

$$\sum_{i=1}^{n} a_i f(x_i).$$

Numerical quadrature is based on interpolation polynomials. We start by dividing $[a, b]$ into $n+1$ points, $[x_0, x_1, ...x_n]$. From these points we can construct a Lagrange polynomial:

$$P_n(x) = \sum_{i=0}^{n} f(x_i) \cdot L_{n,i}(x).$$

Then we can integrate the Lagrange polynomial.

$$\begin{aligned}\int_a^b f(x)dx &= \int_a^b \sum_{i=0}^{n} f(x_i) \cdot L_{n,i}(x)dx + ERROR \\ &= \sum_{i=0}^{n} a_i f(x_i) + ERROR\end{aligned}$$

where $a_i = \int_a^b L_{n,i}(x)dx, i = 0, ..., n$. And the $ERROR$ will depend on the degree of our polynomial. See chapter 4.3 in Burden and Faires (2005), [1].

Two famous numerical integration techniques, based upon numerical quadrature, are the Trapezoidal rule and Simpson's rule. We will only consider Simpson's rule in the following, and we will refer to chapter 4.3 in Burden and Faires (2005), [1], for a discussion regarding the Trapezoidal rule.

Below we discuss three numerical integration techniques based upon the Simpson's rule. First we will take a look at Simpson's rule. When we want to integrate over large intervals this method is quite inaccurate, and we will see that we can modify the rule and obtain Composite Simpson's rule. Alternatively, we can change our approach and make use of an adaptive quadrature method.

Simpson's rule

Simpson's rule follows from integrating a second degree Lagrange polynomial over $[a, b]$, i.e. integrate a Lagrange polynomial with nodes $x_0 = a, x_1 = a + h$ and $x_2 = b$ where $h = \frac{b-a}{2}$. This is done on page 189 in Burden and Faires (2005), [1]. But we can obtain a better error term if we derive the formula by doing a Taylor expansion of $f(x)$ around x_1. This Taylor expansion will then be:

$$f(x) = f(x_1) + (x - x_1)f'(x_1) + \frac{(x - x_1)^2}{2}f''(x_1) + \frac{(x - x_1)^3}{6}f'''(x_1) + HOT$$

where HOT is short for higher order terms. Then if follows that

$$\int_{x_0}^{x_2} f(x)dx = \Big[f(x_1)(x-x_0) + \frac{f'(x_1)}{2}(x-x_1)^2 + \frac{f''(x_1)}{6}(x-x_1)^3 + \frac{f'''(x_1)}{24}(x-x_1)^4 + HOT\Big]_{x_0}^{x_2}$$

which can be simplified to

$$\int_{x_0}^{x_2} f(x)dx = 2hf(x_1) + \frac{h^3}{3}f''(x_1) + O(h^5),$$

by noting that

$$(x_2-x_1)^2 - (x_0-x_1)^2 = (x_2-x_1)^4 - (x_0-x_1)^4 = 0$$

$$(x_2-x_1)^3 - (x_0-x_1)^3 = 2h^3.$$

See page 190 in Burden and Faires (2005), [1]. If we now approximate $f''(x_1)$ as in (A.1), we obtain Simpson's rule:

$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] + O(h^5),$$

A detailed discussion regarding the error can be found in chapter 4.3 in Burden and Faires (2005), [1].

Composite Simpson's rule

If we want to integrate over large intervals, we can use Composite Simpson's rule to obtain a better estimate. The idea behind the Composite Simpson's rule is to divide the interval $[a, b]$ into smaller subintervals and apply Simpson's rule on each of the subintervals, i.e. do a piecewise numerical integration.

Divide the interval $[a, b]$ into an even number of subinterval n. Then

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{j=1}^{n/2} \int_{x_{2j-2}}^{x_{2j}} f(x)dx \\ &= \sum_{j=1}^{n/2} \left\{\frac{h}{3}[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})]\right\} + O(h^4)\end{aligned}$$

where $h = \frac{b-a}{n}$ and $x_j = a + jh$ for each $j = 0, 1, ..., n$. We can simplify this expression and obtain the Composite Simpson's Rule.

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=2}^{(n/2)} f(x_{2j+1}) + f(b)\right] + O(h^4)$$

See chapter 4.4 in Burden and Faires (2005), [1], for more details.

Composite Simpson's rule can be implemented on a computer as in Algorithm 4.1 in Burden and Faires (2005), [1]. In the algorithm below we will approximate the integral $I = \int_a^b f(x)dx$

Algorithm 3: Composite Simpson's Rule

Input: endpoints a, b; even positive integer n
Output: approximation XI of I

begin
  $h \leftarrow \frac{(b-a)}{n}$
  $XI0 \leftarrow f(a) + f(b)$
  $XI1 \leftarrow 0$ `(Summation of` $f(x_{2i-1})$ `)`
  $XI2 \leftarrow 0$ `(Summation of` $f(x_{2i})$ `)`
  for $i = 1$ **to** $n - 1$ **do**
    $X \leftarrow a + ih$
    if *i is even* **then**
      $XI2 \rightarrow XI2 + f(x)$
    else
      $XI1 \rightarrow XI1 + f(x)$
    end
  end
  $XI \leftarrow h(XI0 + 2 \cdot XI2 + 4 \cdot XI1)/3$
  RETURN(XI)
end

Adaptive Quadrature

The Composite Simpson's rule divides the integration interval into equal subintervals. When integrating over large intervals the functional variation might vary a great deal from subinterval to subinterval. A method that divides the integration interval into subintervals which reflects the amount of functional variation might be better and more efficient in many situations. In other words, we should use smaller subintervals in regions with large functional variation, and larger subintervals where there is little functional variation.

Such a method can distribute a predetermined maximal error evenly over the whole integration interval, and might be more efficient since it might require fewer function evaluations.

Here we will describe the idea behind an adaptive quadrature method based on Simpson's rule. More details can be found in chapter 4.6 in Burden and Faires (2005), [1].

The idea is to approximate $\int_a^b f(x)dx$ with an error less than ϵ. We start by using Simpson's rule over the whole integration interval,

$$\int_a^b f(x)dx = S(a,b) - \frac{h^5}{90} f^{(4)}(\mu),\ \mu \in (a,b) \tag{A.2}$$

where $h = \frac{b-a}{2}$ and

$$S(a,b) = \frac{h}{3}\left[f(a) + 4f(a+h) + f(b)\right].$$

Now we divide $[a,b]$ into two subinterval and we use Simpson's rule on each subinterval.

$$\int_a^b f(x)dx = S(a, \frac{b-a}{2}) + S(\frac{b-a}{2}, b) - \frac{1}{16}\frac{h^5}{90}f^{(4)}(\tilde{\mu}),\ \tilde{\mu} \in (a,b) \tag{A.3}$$

where h is the same as above and

$$S(a, \frac{b-a}{2}) = \frac{h}{6}[f(a) + 4f(a+\frac{h}{2}) + f(a+h)]$$
$$S(\frac{b-a}{2}, b) = \frac{h}{6}[f(a+h) + 4f(a+\frac{3h}{2}) + f(b)].$$

If we now assume that $f^{(4)}(\mu) \approx f^{(4)}(\tilde{\mu})$, we can combine (A.2) and (A.3)

$$S(a,b) - \frac{h^5}{90}f^{(4)}(\mu) \approx S(a, \frac{b-a}{2}) + S(\frac{b-a}{2}, b) - \frac{1}{16}\frac{h^5}{90}f^{(4)}(\tilde{\mu})$$

and it follows that

$$\frac{h^5}{90}f^{(4)}(\mu) \approx \frac{16}{15}\left[S(a,b) - S(a, \frac{b-a}{2}) - S(\frac{b-a}{2}, b)\right].$$

If we use this estimate in (A.3) we get

$$\left|\int_a^b f(x)dx - S(a, \frac{b-a}{2}) - S(\frac{b-a}{2}, b)\right| \approx \frac{1}{15}\left|S(a,b) - S(a, \frac{b-a}{2}) - S(\frac{b-a}{2}, b)\right|.$$

This means that $S(a, \frac{b-a}{2}) + S(\frac{b-a}{2})$ approximates the integral about 15 times better than $S(a,b)$. So if

$$\left|S(a,b) - S(a, \frac{b-a}{2}) - S(\frac{b-a}{2}, b)\right| \leq 15\epsilon$$

then

$$\left|\int_a^b f(x)dx - S(a, \frac{b-a}{2}) - S(\frac{b-a}{2}, b)\right| \leq \epsilon$$

and we assume

$$S(a, \frac{b-a}{2}) + S(\frac{b-a}{2}, b)$$

to be an accurate enough approximation of $\int_a^b f(x)dx$. If this does not result in an error estimate less than ϵ, we can do the same procedure on each subinterval, and continue until we obtain a small enough error estimate.

One should realize that this procedure depends on our assumption that $f^{(4)}(\mu) \approx f^{(4)}(\tilde{\mu})$. If this is not true, this method might fail.

The adaptive quadrature can be implemented as done i algorithm 4.3 in Burden and Faires (2005), [1]. In this study this algorithm is implemented i R. The algorithm below approximates the integral $I = \int_a^b f(x)dx$

Algorithm 4: Adaptive Quadrature

Input: endpoints a, b, tolerance TOL, limit number of levels N
Output: approximation APP or message that N is exceeded

begin
  $APP \leftarrow 0$
  $i \leftarrow 1$
  $TOL_i \leftarrow 10 \cdot TOl$
  $a_i \leftarrow a$
  $h_i \leftarrow (b-a)/2$
  $a_i \leftarrow a$
  $FA_i \leftarrow f(a)$
  $FC_i \leftarrow f(a+h)$
  $FB_i \leftarrow (b)$
  $S_i \leftarrow h_i(FA_i + 4FC_i + FB_i)/3$ `Approximation from Simpson's method for entire interval`
  $L_i \leftarrow 1$
  while $i > 0$ **do**
    $FD \leftarrow f(a_i + h_i/2)$
    $FE \leftarrow f(a_i + 3h_i/2)$
    `Approximations from Simpson's method for halves subintervals`
    $S1 \leftarrow h_i(FA_i + 4FD + FC_i)/6$
    $S1 \leftarrow h_i(FC_i + 4FE + FB_i)/6$
    $v_1 \leftarrow a_i$ `Save data at this level`
    $v_2 \leftarrow FA_i$
    $v_3 \leftarrow FC_i$
    $v_4 \leftarrow FB_i$
    $v_5 \leftarrow h_i$
    $v_6 \leftarrow TOL_i$
    $v_7 \leftarrow S_i$
    $v_8 \leftarrow L_i$
    $i \leftarrow i - 1$
    if $|S1 + S2 - v_7| < v_6$ **then**
      $APP \leftarrow APP + (S1 + S2)$
    else
      `Run the algorithm "Adaptive Quadrature Two"`
      RUN(Adaptive Quadrature Two)
    end
  end
  RETURN(APP)
end

Algorithm 5: Adaptive Quadrature Two

begin
  if $v_8 \geq N$ **then**
    OUTPUT('Level Exceeded. Method Fails')
    STOP
  else
    $i \leftarrow i + 1$ `Data for right half subinterval`
1 $a_i \leftarrow v_1 + v_5$
2 $FA_i \leftarrow v_3$
3 $FC_i \leftarrow FE$
4 $FB_i \leftarrow v_4$
5 $h_i \leftarrow v_5/2$
6 $TOL_i \leftarrow v_6/2$
7 $S_i \leftarrow S2$
8 $L_i \leftarrow v_8 + 1$
9 $i \leftarrow i + 1$ `Data for left half subinterval`
10 $a_i \leftarrow v_1$
11 $FA_i \leftarrow v_2$
12 $FC_i \leftarrow FD$
13 $FB_i \leftarrow v_3$
14 $h_i \leftarrow h_{i-1}$
15 $TOL_i \leftarrow TOL_{i-1}$
16 $S_i \leftarrow S1$
17 $L_i \leftarrow L_{i-1}$
  end
end

Appendix B

Additional Tables and Figures

In this appendix we will just report some of our results. They are not part of our main interest, but are interesting and might be a basis for further research.

B.1 From chapter 3

Here we present some results from chapter 3.

Table B.1: Optimal MILPE bandwidths, $n = 100$

Density	$p = 0.5$	$p = 1$	$p = 1.5$	$p = 2.0$
1	0.41	0.42	0.44	0.45
2	0.30	0.30	0.30	0.31
3	0.16	0.12	0.09	0.08
4	0.15	0.10	0.09	0.08
5	0.05	0.05	0.05	0.05
6	0.34	0.36	0.37	0.39
7	0.24	0.25	0.25	0.26
8	0.31	0.31	0.32	0.32
9	0.31	0.33	0.34	0.36
10	0.10	0.10	0.10	0.10
11	0.34	0.36	0.37	0.39
12	0.20	0.21	0.17	0.20
13	0.32	0.34	0.35	0.36
14	0.32	0.34	0.35	0.14
15	0.13	0.15	0.16	0.16

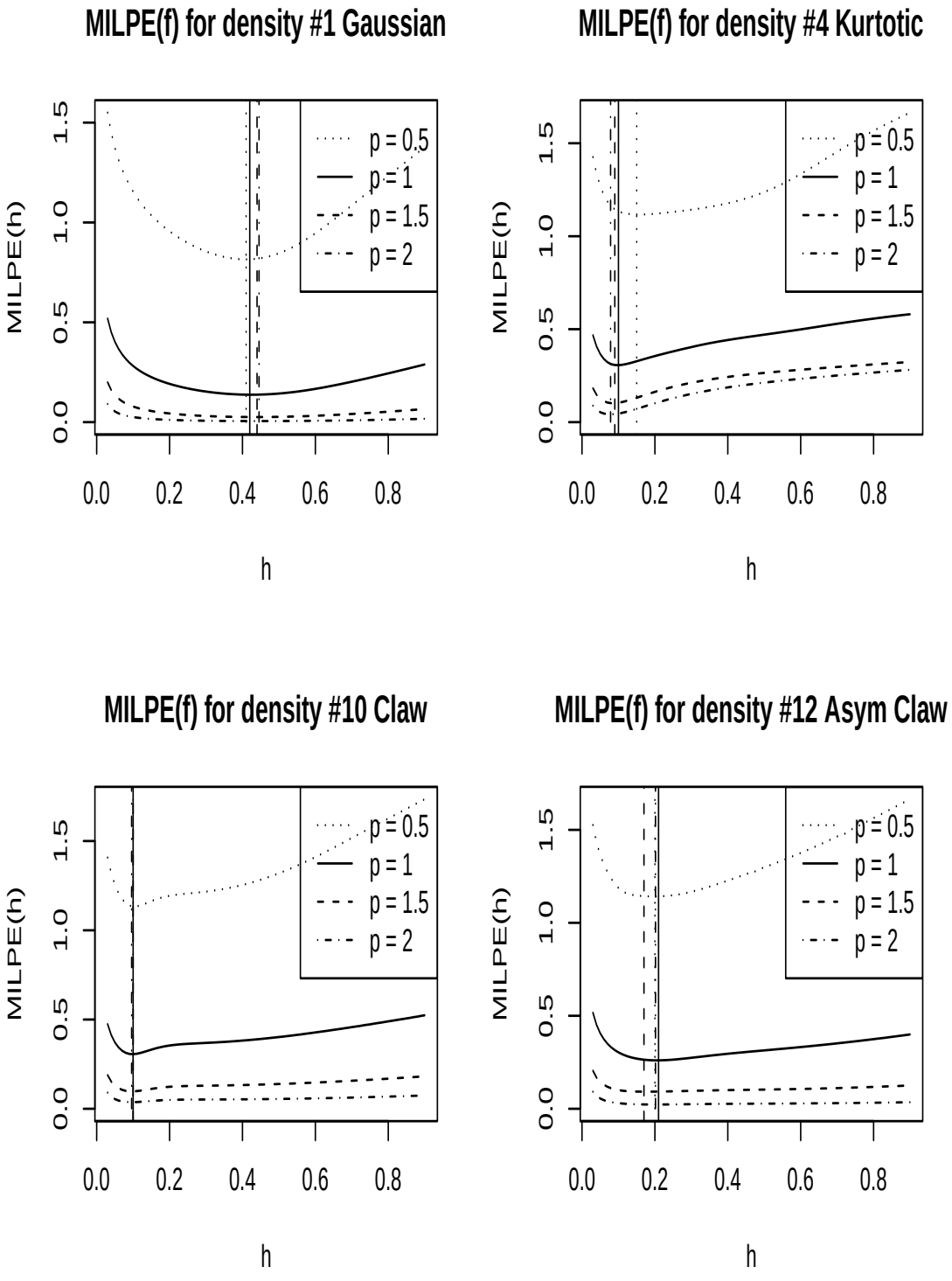

Figure B.1: Plot of $MILPE(h)$ for $p = \{0.5, 1, 1.5, 2\}$ for some target densities.

B.2 From chapter 4

Here we present some results from chapter 4.

Table B.2: Summary statistics for h_{NBW} and h_{DPI}, from the simulation in algorithm 2

Density	$E(h_{NBW})$	$\hat{\sigma}(h_{NBW})$	$E(h_{DPI})$	$\hat{\sigma}(h_{DPI})$
1	0.43	0.058	0.40	0.058
2	0.32	0.042	0.29	0.043
3	0.24	0.025	0.17	0.022
4	0.20	0.031	0.15	0.030
5	0.07	0.009	0.05	0.005
6	0.43	0.058	0.39	0.068
7	0.37	0.016	0.29	0.018
8	0.39	0.054	0.35	0.064
9	0.44	0.048	0.39	0.060
10	0.34	0.043	0.31	0.048
11	0.43	0.057	0.38	0.068
12	0.45	0.058	0.42	0.065
13	0.42	0.053	0.38	0.064
14	0.41	0.027	0.30	0.026
15	0.48	0.041	0.33	0.040

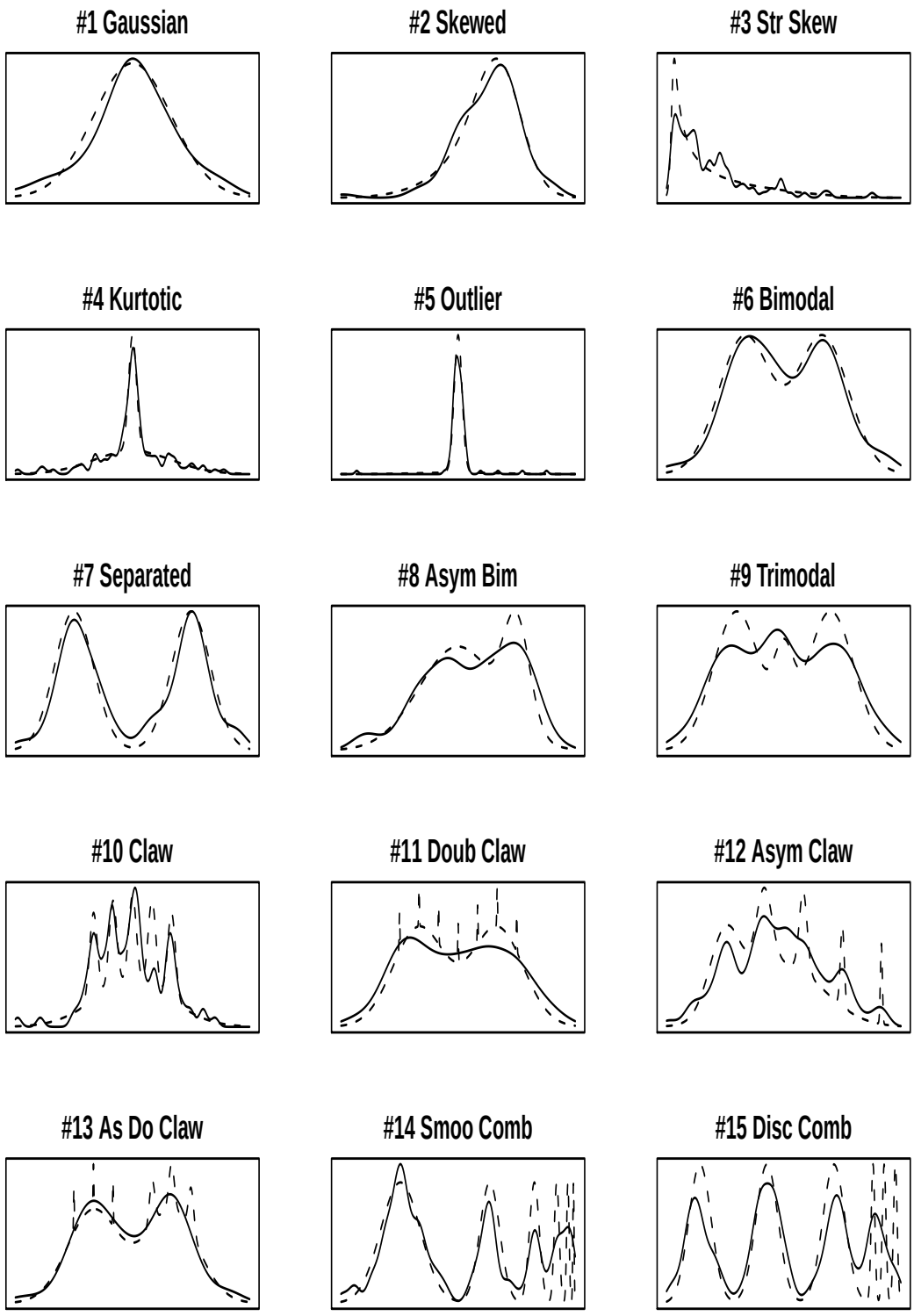

Figure B.2: Plot of kernel density estimated with h_{MISE} as bandwidth and sample size 100. Note that many of these estimates are undersmooth.

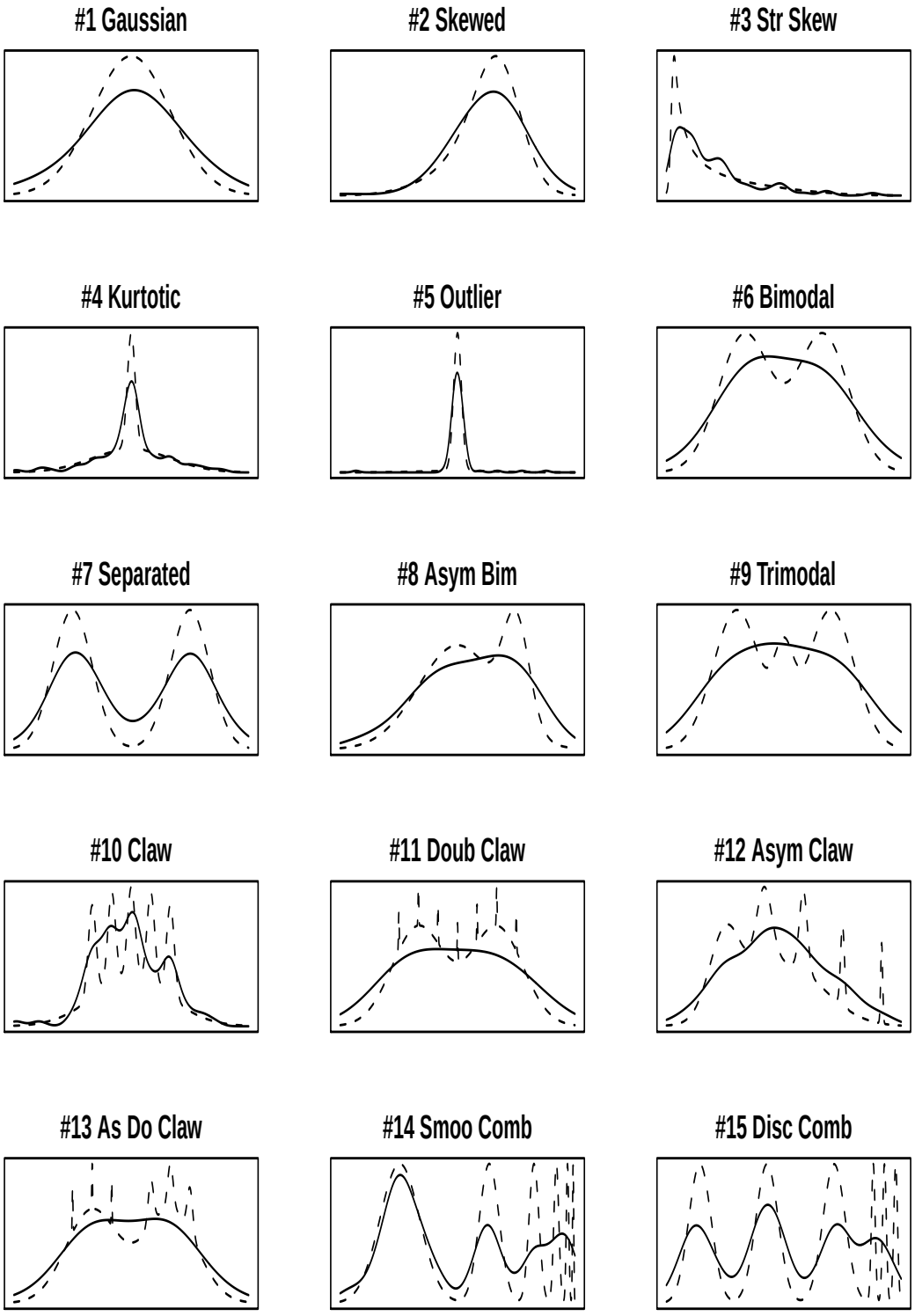

Figure B.3: Plot of kernel density estimates with $2 * h_{MISE}$ as bandwidth and sample size 100. Note that these estimates are smoother, and many of the estimates seems like better estimates from a visual perspective, compared to the density estimates in figure B.2.

Appendix C

Notation

C.1 Definition of error criteria

Mean absolute error: $MAE = E[|\hat{f}_n(x;h) - f(x)|]$
Mean integrated absolute error: $MIAE = E[\int |\hat{f}_n(x;h) - f(x)|dx]$
Mean squared error: $MSE = E[(\hat{f}_n(x;h) - f(x))^2]$
Mean integrated squared error: $MISE = E[\int (\hat{f}_n(x;h) - f(x))^2 dx]$
Visual distance: $d((x, \hat{f}(x;h)), G_{f_X}) = \inf_{(x',y') \in G_{f_X}} ||(x, \hat{f}(x;h)) - (x', y')||_2$
Visual error: $VE_2(\hat{f} \to f_X) = \left[\int_a^b d((x, \hat{f}(x;h)), G_{f_X})^2 dx\right]^{1/2}$

C.2 Notation concerning univariate functions

Let f be a real valued function

$$f_h(x) = f(x/h)/hh > 0$$
$$\int f(x)dx = \int_{-\infty}^{\infty} f(x)dx$$
$$R(f) = \int f^2(x)dx$$
$$\mu_k(f) = \int x^k f(x)dx$$
$$(f * f_1)(x) = \int f(x-y)f_1(y)dy$$
$$\varphi_f(t) = \int e^{itx} f(x)dx$$