Local Likelihood

Master's thesis in mathematical statistics

by

Håkon Otneim

April 23, 2012

Department of Mathematics
University of Bergen
Norway

Preface

Methods for probability density estimation are traditionally classified as either parametric or non-parametric. Fitting a parametric model to observations is generally a good idea when we have sufficient information on the origin of our data; if not, we must turn to non-parametric methods, usually at the cost of poorer performance.

This thesis discusses *local maximum likelihood* estimation of probability density functions, which can be regarded as a compromise between the two mindsets. The idea is to fit a parametric model *locally*, that is, to let the parameters and their estimates depend on the location. If the chosen model is close to the true, unknown density, we keep much of the appealing properties of a full parametric approach. On the other hand, local likelihood density estimates have performance comparable to well known non-parametric methods, even though the locally fitted parametric model differs from the true density in a global sense.

Although traditional methods withstand the test of time as excellent options in many situations, the local maximum likelihood estimator opens up a range of applications. Hjort and Jones [1996], who will serve as the main reference for this thesis, call it semi-parametric density estimation, as it is particularly useful when we have partial knowledge on the shape of the unknown density, but not enough to trust the ordinary, global maximum likelihood estimates. Further, many have built on the idea of locally parametric estimation to applications beyond just density estimation, some of whom have been mentioned and included as references throughout the following chapters.

One-dimensional density estimation will, however, be the primary focus here, with particular emphasis on large sample theory. The main results concern asymptotic bias, which is shown to have a larger order than the bias of traditional kernel estimation as the sample size increases to infinity, and the bandwidth decreases towards zero. Nonetheless, in practical situations with reasonable sample sizes, the local likelihood estimator is shown to perform very well, with an appealing robustness against under- and oversmoothing. Indeed, no experiment performed show signs of deterioration of local likelihood estimates compared to kernel estimation as the sample size grows.

Chapter 1 introduces the notion of likelihood, with basic definitions, examples and properties, as well as some historical remarks. Much of the theory rests upon smoothness conditions that must be imposed on the functions involved. They are stated in some theorems, but most arguments are heuristic of nature.

Chapter 2 motivates the need of modifications to the 'ideal' world of likelihoods in order

to accommodate problems encountered in real life.

The local likelihood function is introduced in Chapter 3 along with some examples of its usefulness in different disciplines of research. Since many results on local likelihood estimation will be compared with equivalents for the non-parametric kernel estimator, a section describing the two mechanisms is included.

Chapters 4, 5 and 6 contain the main results of the thesis. First, some results on asymptotic variance are reviewed, but asymptotic bias receives most of the attention. Section 4.4 is perhaps of particular interest since it is shown that both sources for bias, as we will see arise, have convergence rates depending on the number of parameters in the parametric family, contrary to only one as Hjort and Jones [1996] claim. Simulations then follow to see how the estimator behaves in some constructed situations, especially for bimodal data. A section on bandwidth selection is included as well.

Estimation of densities with bounded support is discussed in Chapter 6. In short, it is demonstrated here that local likelihood estimates perform very much like the kernel estimator or local polynomial estimation near boundaries, depending on which parametric family we choose.

The treatment is concluded with a short review of the popular Cox regression model in light of partial and local likelihood.

The present work would not see the light of day without the pedagogical skills of my two supervisors, Hans A. Karlsen and Dag Tjøstheim, and the steadfast support they have given me the last couple of years. I have truly appreciated our sessions together, and I applaude their ability to let me take the time I need to get my head around the various concepts encountered during the process.

I would also like to thank my fellow students at the Departement of Mathematics for creating the social and positive environment that has made me look forward to get out of bed (almost) every morning. I want to mention my two roomies at the sixth floor, Torbjørn and Silje, in particular. Our endless discussions on topics ranging from existential to trivial, have been most enlightening.

Last, but certainly not least, I thank my absolutely wonderful wife, Karina, and our little boy Kristian, for sticking around, keeping me sane and reminding me each and every day what life is all about.

Bergen, April 2012
Håkon Otneim

Contents

Notation

I have tried to apply standard notational conventions throughout the thesis. These include:

$\mathbb{R}^k$	The k-dimensional Euclidean space
$\boldsymbol{X}$, $\boldsymbol{y}$, $\boldsymbol{\theta}$, ...	Vectors and matrices
$\mathbf{0}$	The zero vector
$\boldsymbol{X}^T$	The transpose of a matrix or vector
$\text{diag}(\cdot)$	Diagonal matrix
$X, Y, Z, \ldots$	Stochastic variables
$f(x) = O(g(x))$ as $x \to a$	$\limsup_{x \to a} \lvert f(x)/g(x) \rvert < \infty$
$f(x) = o(g(x))$ as $x \to a$	$\limsup_{x \to a} \lvert f(x)/g(x) \rvert = 0$
$\xrightarrow{P}$	Convergence in probability
$\xrightarrow{a.s.}$	Convergence almost surely
∇	Gradient vector
$\nabla\nabla$	Matrix of second partial derivatives
I_A	The indicator function for the set A

Chapter 1

Likelihood

A common task for statisticians is parameter estimation. The list of situations in which we use observed data to say something about the underlying model from which the observations originate, is seemingly endless; including regression- and time series analysis, probability density estimation, as well as various applied problems within natural sciences, medical and social research. Often, we assume that observations stem from a certain class of models, determined up to a set of *parameters*. We will hereafter designate this class a *parametric family*. The task is then to construct a functional from the observations to the set of possible parameter values, the *parameter space*, in such a way that the functional value, the *estimate*, which may be a vector, is as close to the true parameters as possible. Through the last couple of centuries, a number of different paths to good parameter estimation have been pursued, and many of these are now used on a day to day basis. The term *likelihood* covers a range of popular methods of estimation, and this chapter will serve as an introduction to the topic before we delve into the more central parts of the thesis.

1.1 Definition

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a realization of the stochastic variable $\boldsymbol{X}$ with probability density function $f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^k$ is a parameter about which we intend to do inference. Based on the observed data, it is in many cases possible to intuitively estimate the unknown parameter by considering different values for it, and determine if one value is more reasonable, or likely, than some other value. This is a subjective exercise, and does not guarantee that our estimate is in fact the most likely. What we need is a likelihood *function*, a function that, given the observed data $\mathbf{x}$, yields larger values for more likely values for $\boldsymbol{\theta}$. If such a function should exist, finding the most likely value for $\boldsymbol{\theta}$ will be reduced to the problem of locating a possible global maximum for the likelihood function.

The likelihood is defined as the joint density function of the observed data, considered as a function of the unknown parameter, $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}|\mathbf{x}) = f_{\boldsymbol{X}}(\mathbf{x}|\boldsymbol{\theta}). \tag{1.1}$$

Given some number or vector $\boldsymbol{\theta}$, $L(\boldsymbol{\theta}|\mathbf{x})$ is the probability of observing $\mathbf{x}$ if $\boldsymbol{\theta}$ is the true parameter value (the discrete case), or in the continuous case, the probability density of the stochastic variable $\mathbf{X}$ at the point $\mathbf{x}$, given $\boldsymbol{\theta}$. The likelihood function can thus be interpreted as how well different values for the parameter explain the observed data. Note that the likelihood function is not a probability function, as it does not necessarily integrate to one with respect to $\boldsymbol{\theta}$.

One might think that there are other functions that measure the reasonability of parameter values, but it will be shown later on that the likelihood function as defined above, possesses many neat properties, and in some sense is the optimal way of creating such a function.

The numerical value of the likelihood function is not interesting in itself. We want to know if the functional value in one point differs from the value in some other point, and if so, by how much they differ. It is common to consider the logarithm of the likelihood function as it will often be easier to analyse:

$$l(\boldsymbol{\theta}|\mathbf{x}) = \log\, L(\boldsymbol{\theta}|\mathbf{x}).$$

We can illustrate the above discussion with a very simple application of the likelihood function. Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ consists of n independent Bernoulli(p)-trials, each with probability mass function (pmf) $P(X = x) = p^x(1-p)^{1-x}$. We want to use the likelihood function to estimate the unknown parameter p based on $\mathbf{x}$. By independence, their joint probability function, and hence the likelihood function, is given as the product of each of the pmf's:

$$\begin{aligned} L(p|\mathbf{x}) &= \prod_{i=1}^{n} P(X = x_i) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^s(1-p)^{n-s}, \end{aligned}$$

where $s = \sum_{i=1}^{n} x_i$. Upon taking logarithm, we get

$$l(p|\mathbf{x}) = s\log p + (n-s)\log(1-p).$$

Differentiating the log-likelihood and equating it to zero will yield a local maximum or minimum.

$$\frac{\partial}{\partial p} l(p|\mathbf{x}) = \frac{s}{p} - \frac{n-s}{1-p} = 0,$$

with solution

$$\widehat{p} = \frac{s}{n} = \overline{x}.$$

By differentiating once more, it is easy to verify that $\widehat{p}$ actually maximizes l. The estimator $\widehat{p}$ for p is called the maximum likelihood estimator (MLE), and the procedure above is quite

standard for obtaining MLEs in simple cases. The derivative of the log-likelihood function with respect to the parameter θ is called the *score* function $u(\theta)$ (which is a vector $\boldsymbol{u}(\boldsymbol{\theta})$ of partial derivatives if there is more than one parameter), and the MLE is the solution of the score equation, $u(\theta) = 0$ (or $\boldsymbol{u}(\boldsymbol{\theta}) = \mathbf{0}$ in the multi-parameter case). In practical situations, it is usually not possible to obtain an analytical solution of the score equation, so numerical methods must be applied.

Finding a global maximum is one thing, but the likelihood function contains more information for us to take advantage of. For instance, the curvature of the likelihood function evaluated at the MLE is called the observed Fisher information. Large Fisher information means that the log-likelihood function has a clear spike, or in other words, the MLE is much more likely to be true than other possible values nearby. Small Fisher information could mean that there are other values that are almost equally likely. In fact, we will show later in this chapter that the asymptotic variance for the maximum likelihood estimator under some regularity conditions is the inverse of the Fisher information, so we can use the second derivative of the log-likelihood function to approximate confidence intervals.

The idea of likelihood can be expanded and generalized in many directions. One can use likelihood to estimate coefficients in regression problems, or to do factor analysis in multivariate statistics. One can also modify the likelihood function to *partial* likelihood (chapter 2) and *local* likelihood (chapter 3 and onwards) to mention a few examples. It is therefore common to speak of likelihood methods as a large collection of procedures and algorithms in statistical analysis, which are based on the idea of a likelihood function as defined by (1.1).

Maximum likelihood theory has not always been as polished as it appears today. Several conjectures and methods have been proven wrong and unusable. Stigler [2007] summarizes the evolution from the first intelligent attempts on finding the the most probable solution to a parametric problem to the theory as it stands in modern statistical analysis. Pearson and Filon [1898] published a method for parameter estimation in a general multivariate and multi-parameter setting, but the approach was soon rejected by Pearson himself, as it did not apply to many problems.

Fisher [1920] discovered sufficiency. Fisher first came to the conclusion that maximizing the likelihood *always* led to a sufficient statistic, but quickly formulated the weaker statement that any sufficient statistic maximizes the likelihood. He later realized that a sufficient statistic of the same dimension as the parameter did not always exist, so he introduced the term *efficiency*: the asymptotic variance of an estimator should be as small as possible (i.e. it reaches the Cramér-Rao lower bound). He proved that if there exists an efficient estimator, then the maximum likelihood estimator is asymptotically efficient under certain regularity conditions.

The theory of Fisher evolved further with correspondence between him and Hotelling and critisism from Neyman. Joe Hodges presented in lectures in 1951 what is called the *'Nasty Ugly Little Fact'*, a so-called super-efficient estimator, with smaller asymptotic variance then the maximum likelihood estimator. This is dealt with in the regularity conditions of Fisher. Hodges' estimator, as well as other examples of estimators that are

better than the MLE, created significant murmur at the time of their discoveries, but they are now considered more as technical details than of practical importance. The theory of maximum likelihood is both powerful and useful, even in situations where no general theorem can be applied, but let us end this section using the words of Stigler [2007] : *'Maximum likelihood remains a truly beautiful theory, even though tragedy may lurk around a corner'.*

1.2 The likelihood function is optimal

Let us now follow the argument of Severini [2000, p.76] to show that the likelihood function, as defined by equation (1.1), is indeed the best likelihood function in a certain sense.

Assume $\mathbf{x}$ is a vector of observations. Our task is to choose between two possible values for the unknown parameter $\boldsymbol{\theta}$: $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$. This is a very simple, but illustrative case. Based on the likelihood function, there are two sets in the sample space, X_1 and X_2, such that we will choose $\boldsymbol{\theta}_1$ to be the true value for $\boldsymbol{\theta}$ if $\mathbf{x} \in X_1$, and likewise let $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ if $\mathbf{x} \in X_2$.

In this setting, there are two types of errors that might occur: We observe $\mathbf{x} \in X_1$ when, in fact, $\boldsymbol{\theta} = \boldsymbol{\theta}_2$, or we observe $\mathbf{x} \notin X_1$ even though $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. It is now reasonable to choose X_1 to be the set that minimizes the probability of the sum of the two possible errors. Let $Q(X_1)$ denote this probability:

$$Q(X_1) = P(\mathbf{X} \in X_1|\boldsymbol{\theta}_2) + 1 - P(\mathbf{X} \in X_1|\boldsymbol{\theta}_1).$$

Minimizing $Q(X_1)$ is equivalent to minimizing

$$\int_{X_1} f(x|\boldsymbol{\theta}_2)\,\mathrm{d}x - \int_{X_1} f(x|\boldsymbol{\theta}_1)\,\mathrm{d}x = \int_{X_1} \left(f(x|\boldsymbol{\theta}_2) - f(x|\boldsymbol{\theta}_1)\right)\,\mathrm{d}x$$

with respect to the set X_1. It is easily seen that the set which gives the largest negative contribution to the integral above is

$$X_1 = \{\mathbf{x} \in X : (f(x|\boldsymbol{\theta}_2) < (f(x|\boldsymbol{\theta}_1)\},$$

or in other words, we choose $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ if $L(\boldsymbol{\theta}_1) > L(\boldsymbol{\theta}_2)$. This justifies our use of the word *likelihood*, because choosing the most likely value for the parameter θ (i.e. maximizing the likelihood function) minimizes the probability of making erroneous conclusions.

Note that the argument above is just a version of the Neyman-Pearson Lemma, see for example Casella and Berger [2002], p. 388.

1.3 The likelihood principle

The business of likelihood is not without controversies. Before reaching conclusions about the unknown parameter, we want to make sure that we have considered all available information. It is pleasing to know that the likelihood function itself is a minimal sufficient

statistic [Pawitan, 2001, p. 56], that is, it captures all the information about the parameter available through an experiment, but nothing more; anything less will lead to loss of information. Does this mean that we only need data and the likelihood function to do inference, and additional information about the experiment is redundant? The likelihood principle says yes:

> *Suppose we have an observation x from a statistical model $\{f(x|\theta); \theta \in \Theta\}$ and an observation y from a statistical model $\{g(y|\theta); \theta \in \Theta\}$ where the parameter θ has the same meaning in both models. Let $L(\theta)$ denote the likelihood function for the model f, and let $\tilde{L}(\theta)$ denote the likelihood function for the model g. If, for given x and y, $L(\theta) = \tilde{L}(\theta)$ for all θ, then our conclusions regarding θ based on observing x should be the same as our conclusions regarding θ based on observing y.*

(Formulation from Severini [2000], p. 77). At first glance, the likelihood principle is reasonable as it somehow ensures the objectiveness of our inference. It is, however, not universally accepted, because there are many examples where the nature of the experiment should be taken into consideration. A commonly used example follows next (from Pawitan [2001], p. 195):

Suppose our task is to estimate the probability p of getting heads when tossing a coin. First, we plan to toss the coin ten times, and of those ten tosses, we observe 8 heads. Then we do a second experiment, where we plan to stop when the second tail is observed. In the second experiment, we also observe 8 heads. How do we proceed to estimate p? The likelihood function is the same in both experiments:

$$L(p) = \text{constant} \times p^8(1-p)^2.$$

By the likelihood principle, we should therefore reach the same conclusion about p in both experiments.

Suppose now that we wish to test H_0: $p = 0.5$ versus H_1: $p > 0.5$. The p-value from the first experiment is

$$\begin{aligned} p_1 &= P(X \geq 0.8 | p = 0.5) \\ &= \sum_{x=8}^{10} \binom{10}{x} 0.5^{10} \\ &= 0.055, \end{aligned}$$

thus we would not reject the null-hypothesis at a 5% level. The p-value from the second experiment is given by

$$\begin{aligned} p_2 &= P(X \geq 0.8 | p = 0.5) \\ &= \sum_{x=8}^{\infty} (x+1) 0.5^{x+2} \\ &= 0.020, \end{aligned}$$

which means that we would reject H_0 at a 5% level. This experiment is therefore in conflict with the likelihood principle, as we will reach different conclusions about p, even though the likelihood function in both cases are the same.

The example above illustrates why many statisticians do not believe that the likelihood principle is valid. It should, however, not lead to its total rejection. In fact, Birnbaum [1962] showed that the likelihood principle is equivalent to the sufficiency and conditionality principles together, and these two principles are somewhat easier to accept.

Casella and Berger [2002, p. 292] defines an experiment E to be the triple $(\mathbf{X}, \boldsymbol{\theta}, f(\mathbf{x}|\boldsymbol{\theta}))$, where $\mathbf{X}$ is a random vector with probability mass function $f(\mathbf{x}|\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$. We denote by $Ev(E, \mathbf{x})$ the inference we make about θ, and call it the evidence arising from the experiment and the observations $\mathbf{x}$, including knowledge of the experiment. The sufficiency principle states:

> *Suppose we perform an experiment E, and $T(\boldsymbol{X})$ is a sufficient statistic. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are sample data from E such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then*
>
> $$Ev(E, \boldsymbol{x}) = Ev(E, \boldsymbol{y}).$$

Since T is sufficient, we know that it contains all information about θ. Therefore, the sufficiency principle seems very reasonable as we should not gain any more or less evidence from the same amount of information.

Suppose now that there are two experiments, $E_1 = (\mathbf{X}_1, \theta, \{f_1(\mathbf{x}_1|\theta)\})$ and $E_2 = (\mathbf{X}_2, \theta, \{f_2(\mathbf{x}_2|\theta)\})$ where θ is the same in both experiments. Consider a mixture experiment, E^*, where a random index J is chosen to be 1 or 2, each with probability 0.5, and that E_J is performed after each trial. The conditionality principle states:

> *The evidence from a mixture experiment is equal to the evidence from the experiment performed:*
>
> $$Ev(E^*, \boldsymbol{x}^*) = Ev(E_j, \boldsymbol{x}_j),$$
>
> *for $j = 1, 2$.*

Birnbaums theorem states that the sufficiency and conditionality principles together are equivalent to the likelihood principle, and the proof can be found, for example, in Pawitan [2001], p. 198.

1.4 Some properties of the MLE

In this section we state some of the most appealing properties of the maximum likelihood estimator. Recall from the historical discussion in Section 1.1 that, although the theory of Fisher, Neyman and others is both beautiful and powerful, the discoveries of counter-examples made it necessary to impose regularity conditions on f in order to prove theorems on e.g. asymptotic normality and consistency. This is why many results in the theory of likelihood begin with a phrase of the kind: 'Under appropriate smoothness conditions on f

. . . '. Such a formulation may seem a little strange, as any statement can be proved by just imposing the necessary conditions for the statement to be true. However, the smoothness conditions in the following theorems are reasonable, but necessary to make the proofs work.

Note further that when obtaining formal results, we usually define maximum likelihood estimators only as roots of the score equation; we do not require them to be global maxima.

1.4.1 The invariance property

Suppose we do not want to estimate the unknown parameter θ, but rather a function of it, say $\tau(\theta)$. The invariance property of the MLE states that if $\hat{\theta}$ is the MLE of θ and $\tau(\theta)$ is any function, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ in general. See Casella and Berger [2002], p. 319, for a short proof and a discussion on the formalities of this problem.

1.4.2 Consistency

One of the most appealing properties of the maximum likelihood estimator is consistency. By imposing sufficient smoothness conditions on the density function f, we can show that the MLE converges to the true value $\boldsymbol{\theta}_0$. Schervish [1995, p. 415] proves convergence almost surely for maximum likelihood estimators in parameter spaces not necessarily compact through the following theorem:

Theorem 1.1. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of iid (independent and identically distributed) variables with density $f_X(x|\boldsymbol{\theta})$ and let Θ denote the parameter space. Assume that the true value for $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$. Define for each $\Theta' \subseteq \Theta$,*

$$Z(\Theta', x) = \inf_{\boldsymbol{\theta} \in \Theta} \log \left\{ \frac{f_X(x|\boldsymbol{\theta}_0)}{f_X(x|\boldsymbol{\theta}_0)} \right\}.$$

Assume that for each $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$ there is an open set $N_{\boldsymbol{\theta}}$ such that $\boldsymbol{\theta} \in N_{\boldsymbol{\theta}}$ and $E_{\boldsymbol{\theta}_0} Z(N_{\boldsymbol{\theta}}, X_i) > 0$. If Θ is not compact, assume further that there is a compact $C \subseteq \Theta$ such that $\boldsymbol{\theta}_0 \in C$ and $E_{\boldsymbol{\theta}_0} Z(\Theta \setminus C, X_i) > 0$. Then the corresponding sequence of maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_n \overset{a.s.}{\to} \boldsymbol{\theta_0}$.

Instead of proving the formal version above, we can follow the heuristic argument of Rice [1995, pp. 261-263] to establish weak consistency in the basic one-dimensional case for a continuous variable:

Maximizing the log-likelihood $l(\theta)$ is equivalent to maximizing

$$\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta).$$

From the law of large numbers, we get that

$$\frac{1}{n} l(\theta) \overset{P}{\to} E \log f(X_i|\theta) = \int \log f(x|\theta) f(x|\theta_0) \, \mathrm{d}x. \tag{1.2}$$

One can now show that the θ that maximizes the left side of (1.2) (i.e. the MLE), converges to the maximizer of the right side as $n \to \infty$. This maximizer could most certainly be θ_0 because the derivative of $E \log f(X|\theta)$ is (smoothness conditions required)

$$\frac{\partial}{\partial \theta} \int \log f(x|\theta) f(x|\theta_0)\, \mathrm{d}x = \int \frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\, \mathrm{d}x,$$

and inserting $\theta = \theta_0$ gives

$$\int \frac{\partial}{\partial \theta} f(x|\theta_0)\, \mathrm{d}x = \frac{\partial}{\partial \theta} \int f(x|\theta_0)\, \mathrm{d}x = 0,$$

so θ_0 is at least a stationary point.

1.4.3 Asymptotic normality

Under certain regularity conditions on f, the MLE is approximately normally distributed as the sample size increases. As in the previous section, we start out by stating a formal version of the theorem, and then illustrate the concept using a simplified argument. Theorem 7.63 of Schervish [1995, p. 421] states:

Theorem 1.2. *Let the parameter space Θ be a subset of $\mathbb{R}^p$, and let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. given $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, each with density $f_{X_1}(x|\boldsymbol{\theta}_0)$. Let $\hat{\boldsymbol{\theta}}_n$ be the MLE based on the n first observations. Assume that $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta_0}$. Further, assume that $f_{X_1}(x|\boldsymbol{\theta})$ has continuous second partial derivatives with respect to $\boldsymbol{\theta}$ and that differentiation can be passed under the integral sign. Assume that there exists a function $H_r(x, \boldsymbol{\theta})$ such that, for each $\boldsymbol{\theta}$ in the interior of Θ and each k, j,*

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le r} \left| \frac{\partial^2}{\partial \theta_k \partial \theta_j} \log f_{X_1}(x|\boldsymbol{\theta}_0) - \frac{\partial^2}{\partial \theta_k \partial \theta_j} \log f_{X_1}(x|\boldsymbol{\theta}) \right| \le H_r(x, \boldsymbol{\theta}_0),$$

with $\lim_{r \to 0} E_{\boldsymbol{\theta}_0} H_r(X, \boldsymbol{\theta}_0) = 0$. Assume that the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ is finite and non-singular. Then,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)).$$

For a simpler presentation, we follow the lemma and theorem as formulated and proved by Rice [1995, pp. 263-264].

Lemma 1.3. *Define $I(\theta)$ by*

$$I(\theta) = E\left[\frac{\partial}{\partial \theta} \log f(X|\theta) \right]^2.$$

Under appropriate smoothness conditions on f (they can be formulated as in Theorem 1.2, $I(\theta)$ may also be expressed as

$$I(\theta) = -E\left[\frac{\partial^2}{\partial^2 \theta} \log f(X|\theta) \right].$$

We may now proceed to sketch a proof of the asymptotic normality of the MLE:

Theorem 1.4. *Let $X_1, \ldots, X_n, \ldots$ be a sequence of independent and identically distributed (i.i.d.) observations from the probability density or mass function $f(x|\theta)$, let $\hat{\theta}_n$ be the MLE of the univariate θ based on $X_1, \ldots, X_n$, and let θ_0 be the true value of θ. Under the smoothness conditions mentioned above, the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0)$ tends to a standard normal distribution.*

Proof. Let $l(\theta)$ be the log-likelihood function. From a Taylor-series expansion we have

$$\begin{aligned} 0 = l'(\hat{\theta}_n) &= l'(\theta_0) + (\hat{\theta}_n - \theta_0) l''(\theta_0) + R, \\ (\hat{\theta}_n - \theta_0) &= \frac{-l'(\theta_0)}{l''(\theta_0)}, \\ n^{1/2}(\hat{\theta}_n - \theta_0) &= \frac{n^{-1/2} l'(\theta_0)}{-n^{-1} l''(\theta_0)}, \end{aligned}$$

where the remainder R has been set to zero. First we consider the numerator for the last expression, and find its expectation.

$$\begin{aligned} \mathrm{E}\left[l'(\theta_0)\right] &= \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\bigg|_{\theta=\theta_0}\right] \\ &= \sum_{i=1}^{n} \int \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\bigg|_{\theta=\theta_0}\right] f(x|\theta_0)\,\mathrm{d}x \\ &= \sum_{i=1}^{n} \int \frac{\partial}{\partial\theta} f(x|\theta_0)\,\mathrm{d}x \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \int f(x|\theta_0)\,\mathrm{d}x \\ &= 0, \end{aligned}$$

because the last integral equals one. The numerator's variance is

$$\begin{aligned} \mathrm{Var}\left[n^{-1/2} l'(\theta_0)\right] &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial}{\partial\theta} \log f(X_i|\theta)\bigg|_{\theta=\theta_0}\right]^2 \\ &= I(\theta_0). \end{aligned}$$

Next, we consider the denominator:

$$-\frac{1}{n} l''(\theta_0) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta_0).$$

By the law of large numbers and from Lemma 1.3, the latter expression converges in probability to

$$-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(X|\theta_0)\right] = I(\theta_0).$$

We have now that

$$n^{1/2}(\hat{\theta}_n - \theta_0) \approx \frac{n^{-1/2} l'(\theta_0)}{I(\theta_0)},$$

with expectation

$$\mathrm{E}\left[n^{1/2}(\hat{\theta}_n - \theta_0)\right] = 0$$

and asymptotic variance

$$\begin{aligned}\mathrm{Var}\left[n^{1/2}(\hat{\theta}_n - \theta_0)\right] &= \frac{I(\theta_0)}{(I(\theta_0))^2} \\ &= \frac{1}{I(\theta_0)},\end{aligned}$$

thus

$$\mathrm{Var}(\hat{\theta}_n - \theta_0) \approx \frac{1}{nI(\theta_0)}.$$

The central limit theorem may be applied to $l'(\theta_0)$, which is a sum of i.i.d. random variables:

$$l'(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_0} \log f(X_i|\theta),$$

and the result follows from Slutsky's theorem. □

The argument above deserves a few comments. Firstly, the proof is not rigorous. Both breaking off the Taylor expansion, the use of Lemma 1.3, as well as interchanging differentiation and integration require smoothness conditions on f which are stated in Theorem 1.2. Secondly, the function $I(\theta)$ is the expected the Fisher information. Note that, as indicated in section 1.1, small Fisher information results in large asymptotic variance, and large Fisher information results in a smaller asymptotic variance of the MLE. Thirdly, Theorem 1.4 is here proved for a univariate parameter only. In a multi parameter case, the argument is similar, but the Fisher information is replaced by the Fisher information matrix, whose inverse serves as the covariance matrix in the asymptotic normal distribution, and consequently need to be assumed non-singular (see Theorem 1.2).

1.4.4 Stochastic properties of the score function

Before we observe data, the score function $\boldsymbol{U}(\boldsymbol{\theta}|X) = \nabla \log l(\boldsymbol{\theta}|X)$ is a stochastic variable (and is thus denoted using the upper case $\boldsymbol{U}$). Knowing its expectation and variance is vital, both for theoretical arguments (some of which we will encounter in later chapters), but also in the development of numerical procedures to calculate estimates of parameters. Let $\boldsymbol{\theta}$ be the p-dimensional vector of parameters with score function $\boldsymbol{U}(\boldsymbol{\theta}|X) = (U_1, U_2, \ldots, U_p)$, where $U_j = \frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}, X)$. Then we have that

- $\mathrm{E}(\boldsymbol{U}(\boldsymbol{\theta}|X)) = \mathbf{0}$

- $\mathrm{Cov}(U_j, U_k) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta_j\partial\theta_k} l(\boldsymbol{\theta}, X)\right) = -\mathrm{E}\left(\frac{\partial U_j}{\partial\theta_k}\right) = -\mathrm{E}\left(\frac{\partial U_k}{\partial\theta_j}\right),$

the second of which is a generalization of Lemma 1.3. In other words, the covariance matrix of the scores is equal to the expected Fisher information matrix. The proofs for these claims are straightforward algebra, and can be found, for example, in Lehmann and Casella [1998], Lemma 5.3.

1.5 Existence and uniqueness

Can we always find a maximum likelihood estimator? If we have found an MLE, can we be sure that it is unique? These are perhaps the most important questions to ask in a general setting, and the answer to both of them is, generally, "no". The optimality argument of Section 1.2 will remain valid if the inequality signs $<$ and $>$ are replaced by $\leq$ and $\geq$ respectively. If the situation should arise that $L(\theta_1) = L(\theta_2)$, which value should we choose? Both of them maximize the likelihood and are thus MLEs.

If Θ is a subset of $\mathbb{R}^1$ or $\mathbb{R}^2$, it is easy to visualize likelihood functions with no maximal value or likelihood functions that are periodic, such that global maxima exist in abundance, but none are unique. Likelihood functions are, however, special due to their construction. They are probability density functions considered as functions of a parameter, and thus follow certain rules. Periodic likelihood functions are rare, but examples where existence and/or uniqueness are not satisfied exist. Again, we have to identify what conditions that suffice in order to guarantee a MLE to exist and be unique. The following theorem does just that [Mäkeläinen et al., 1981].

Theorem 1.5. *Let $L(\boldsymbol{\theta})$ be a twice continuously differentiable likelihood function with $\boldsymbol{\theta}$ varying in a connected subset $\Theta \subset \mathbb{R}^p$. Let $\partial\Theta$ denote the boundary of Θ. Suppose that*

$$\lim_{\boldsymbol{\theta}\to\partial\Theta} L(\boldsymbol{\theta}) = c, \tag{1.3}$$

and that the Hessian matrix

$$\boldsymbol{H}(\boldsymbol{\theta}) = \left\{\frac{\partial^2 L}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta})\right\},$$

of second partial derivatives is negative definite at every point $\boldsymbol{\theta} \in \Theta$ for which the gradient vector

$$\nabla L = \{\partial L/\partial\theta_i\}$$

is zero. Then there is a unique maximum likelihood estimate $\widehat{\boldsymbol{\theta}} \in \Theta$ based on the sample $\boldsymbol{x}$, and the likelihood function attains

- *no other maxima in Θ,*
- *no minima or other stationary points in Θ,*
- *its infimum value c on the boundary $\partial\Theta$ and nowhere else.*

The constant c is either a real number or $-\infty$. A sequence $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ is said to converge to the boundary $\partial\Theta$ if for every compact set $K \subset \Theta$ there exists an integer $k_0 \geq 1$ such that for all $k > k_0$, $\boldsymbol{\theta}_k \notin K$. Consequently, if $\Theta = \mathbb{R}^k$, condition (1.3) means that the functional value of any divergent sequence $\{\boldsymbol{\theta}_k\}$ approaches c.

The easiest way to interpret the conditions may be to see them in the one-dimensional case. The boundary is constant and the second derivative is negative, so the continuous likelihood function will then have exactly one maximum.

1.6 Robustness and M-estimators

When we choose a parametric model for data fitting, we always run the risk of choosing the wrong model. Estimating the mean and variance of a normal distribution is all well and good if the data at hand is actually normally distributed, but is nothing short of a catastrophe if the underlying model is exponentially distributed. In other words, the quality of an estimator is not necessarily determined by its variance and bias only, but also by its *robustness* against misspecification of the parametric model and/or flawed observations.

Casella and Berger [2002] present a simple illustration of this problem. The sample mean is unbiased and has the smallest possible variance as an estimator for the population mean in a normal distribution. If there are observations that are wrong for some reason and much larger or smaller than the rest, however, the sample mean will be affected and perhaps be driven far off its target. The sample median does not suffer very much from this problem. We can actually let the largest half (roughly) of the observations drift off to infinity without affecting the median at all. Thus, the sample median is more robust against outliers than the sample mean, but, as we would expect, we must pay for increased robustness by poorer performance in terms of more traditional measures; in the normal case, the median/mean asymptotic relative efficiency is 0.64 [Casella and Berger, 2002, p. 484].

For independent observations $x_1, x_2, \ldots, x_n$, the mean minimizes $\sum(x_i - a)^2$, the median minimizes $\sum |x_i - a|$, while the maximum likelihood estimator maximizes $\sum \log f(x_i|\theta)$ (and thus minimizes the negative log-likelihood). We can then introduce a larger class of estimators, defined as the minimum of a certain function

$$\sum_{i=1}^{n} \rho(x_i, \theta), \tag{1.4}$$

with respect to the unknown parameter, and where the specific choice of ρ depends on the problem at hand. These estimators are called *M-estimators* (M for "maximum likelihood type" [Huber and Ronchetti, 1981]), and are often found by solving

$$\sum_{i=1}^{n} \psi(x_i, \theta) = 0,$$

where $\psi(x, \theta) = \partial/\partial\theta\ \rho(x, \theta)$. The *Huber estimator* is an M-estimator designed to estimate the mean of a population. It is defined as the minimum of (1.4), with

$$\rho(x) = \begin{cases} \frac{1}{2}(x-\mu)^2 & \text{if } |x-\mu| \leq k \\ k|x-\mu| - \frac{1}{2}k^2 & \text{if } |x-\mu| \geq k, \end{cases}$$

[Casella and Berger, 2002, p.484]. For small deviations from the estimated mean, the Huber estimator behaves like the sample mean, and for larger deviations, it behaves more like the median. We can vary the parameter k, the *tuning parameter*, to adjust the level of robustness. It is also easily checked that ρ is continuous and differentiable in this case.

Theoretically tractable estimators may not always be useful in practical situations. In the next chapter we will look at some ways to deal with likelihood functions that are hard to obtain, or too complicated to work with.

Chapter 2

Variations of the likelihood

A few common variations of the likelihood function will be discussed next, the most important being the *partial* likelihood, that will be revisited in Chapter 7 in light of the main topic in this thesis, *local* likelihood.

2.1 Definition of partial likelihood

In most real-world applications, obtaining the likelihood function and its maximum is not as easy as in the example of Section 1.1. There are also situations where the standard likelihood function will produce misleading conclusions. These problems call for adjustments to the theory such that it can be applied to a larger class of problems. One adjustment is the *partial* likelihood function, introduced by Cox [1975]. Cox points at several motivations for introducing partial likelihood, among others the study of problems in which the full likelihood is complicated or impossible to obtain, and the reduction of dimensionality in the presence of many nuisance parameters, i.e parameters of little or no interest.

Let $\boldsymbol{Y}$ be a random variable with probability density function $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{y})$, and suppose $\boldsymbol{Y}$ can be transformed to new random variables $(\boldsymbol{X}, \boldsymbol{S})$ by a transformation not depending on the unknown parameter. The joint density, and thus the likelihood function, of $(\boldsymbol{X}, \boldsymbol{S})$ is given by

$$L_{\boldsymbol{X},\boldsymbol{S}}(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{s}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{s}|\boldsymbol{x}),$$

where the two factors in special cases are called the marginal and conditional likelihoods of $\boldsymbol{X}$ and $\boldsymbol{S}$ respectively, see Section 2.4. Both factors can be analyzed by themselves, and this is a useful simplification in many applications. In the examples presented below, the choices of what to include in $\boldsymbol{X}$ and $\boldsymbol{S}$ respectively, are fairly clear. In other applications, this separation is perhaps neither obvious nor unique, and should be chosen carefully in order to maximize the efficiency. Suppose now that we have observations $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n)$ that are transformed into a sequence of $\boldsymbol{X}$s and $\boldsymbol{S}$s:

$$\mathbf{Y} = (\boldsymbol{X}_1, \boldsymbol{S}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{S}_n). \tag{2.1}$$

The full likelihood of (2.1) is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}_1)\prod_{i=2}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{d}_i)\prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{s}_i|\boldsymbol{c}_i), \tag{2.2}$$

where $\boldsymbol{d}_i = (\boldsymbol{s}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{s}_{i-1}, \boldsymbol{x}_{i-1})$ and $\boldsymbol{c}_i = (\boldsymbol{s}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{s}_{i-1}, \boldsymbol{x}_{i-1}, \boldsymbol{x}_i)$, and where $\boldsymbol{c}_1 = \boldsymbol{x}_1$. The last product is called the partial likelihood based on $\boldsymbol{S}$. Further analysis of the nature of the unknown parameter $\boldsymbol{\theta}$ can now be done based on the partial likelihood alone, disregarding the first factor. One major strength of this procedure is that we do not need to specify the first factor in any way if we know that it does not contain any crucial information about $\boldsymbol{\theta}$. On the other hand, the biggest problem of partial likelihood estimation is that the partial likelihood function does not carry all available information about the unknown parameter, so we need to control the loss of information somehow.

2.2 Applications of partial likelihood

In the literature, there seem to be particular focus on applying partial likelihood in the following settings:

- Situations with censored or missing data,
- stochastic processes consisting of several more or less distinct probabilistic models, where some models are difficult to handle and/or not interesting to do inference about, and
- splitting the parameter space into components of interest and nuisance parameters such that the partial likelihood depends mostly (preferably only) on the parameters of interest, and the disregarded factor depends essentially on the nuisance parameters.

The following example is presented by Wong [1986].

Suppose we observe J disconnected segments of a Markov chain, $[z_{n_1}, \ldots, z_{m_1}], [z_{n_2}, \ldots, z_{m_2}], \ldots, [z_{n_J}, \ldots, z_{m_J}]$. The values between z_{m_i} and $z_{n_{i+1}}$ are not observed for some reason, and are missing. Suppose the one-step transition probabilities within the observed segments depend on the parameter θ such that $P(Z_n = z_n|Z_{n-1} = z_{n-1}) = p_\theta(z_{n-1}, z_n)$. The full likelihood function for the observed data is given by

$$L(\theta|\mathbf{z}) = \prod_{j=1}^{J}\left[f(z_{n_j}|z_{m_{j-1}})\prod_{n_j+1}^{m_j} p_\theta(z_{n_j}|z_{n_j}-1)\right]. \tag{2.3}$$

The likelihood function is here, without any transformation, factorized with two factors. Comparing (2.3) with (2.2), we see that in this example the Xs describe the path from one observed segment to the next, while the Ss are random variables within the observed segments. The sequence $\{X_n\}$ does not even need to be a Markov chain, and since we do

not have much information to support a conclusion regarding this process, it will be hard to conduct inference about θ based on the full likelihood. Luckily, according to Cox [1975], we can just consider the partial likelihood, which in this example will be the second factor:

$$L_p(\theta|\mathbf{z}) = \prod_{j=1}^{J} \left[\prod_{n_j+1}^{m_j} p_\theta(z_n|z_{n-1}) \right].$$

In situations like the example above, the missing segments will typically have a different, and perhaps a more complicated, probabilistic structure than the observed ones. The partial likelihood function will then serve as a path around the problem of guessing or trying models with little evidence to be true, that would create great uncertainty in the end result.

The next example, included by Cox [1975] and investigated further by Efron [1977], shows how partial likelihood can be used to deal with censored data. It is often referred to as the Cox regression model.

Consider an experiment with initially n individuals at risk of failure. Suppose we observe the failure times of the individuals $i_1, i_2, \ldots, i_J$ to be $t_1 < t_2 < \ldots < t_J$. The hazard rate for subject i is assumed to be on the form

$$\lambda_i(t|\boldsymbol{z}_i) = \lambda_0(t)\Psi_i(t, \boldsymbol{z}_i), \tag{2.4}$$

where Ψ_i often is on the form $\Psi_i(t, \boldsymbol{z}_i, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}\boldsymbol{z}_i(t))$, $\boldsymbol{z}_i(t)$ is the vector of observed covariates, and $\boldsymbol{\beta}$ is a vector of regression coefficients. Efron [1977] uses the following real-world analogue to the above situation: Suppose we want to study the life-length of residents at a senior-citizen's facility. The residents move in at various ages, and their time of death is recorded. Residents may also move out of the facility for reasons other than death, and not all of them will have died at the end of the study. We observe a collection of covariates for each individual, time-varying or constant, such as age, sex, blood pressure or body weight, and our aim is to estimate $\boldsymbol{\beta}$ based on the observations.

The joint probability density function of the failure time of individual i_j and the order statistic $\{t_{(j)}\}_{j=1}^{J}$ can be factorized into the conditional probability that item i_j failed at time t_j, given that exactly one item failed at that time (S given X), and the marginal distribution of the failure times (X). The last factor can be hard to obtain, so we take the first factor to be the partial likelihood. From the hazard rate (2.4), the conditional probability can be shown to be [Efron, 1977]

$$P(i_j \text{ failed at } t = t_j | \text{ one item failed at } t = t_j) = \frac{\Psi_{i_j}(t_j, \boldsymbol{z}_i, \boldsymbol{\beta})}{\sum_{i \in R(t_j)} \Psi_i(t_j, \boldsymbol{z}_i, \boldsymbol{\beta})}$$

where $R(t_j)$ is the *risk set* at time t_j, i.e. the number of individuals on trial. The risk set will vary in size as time goes by due to failures, but also due to censoring. Proceeding as in (2.3), multiplying the conditional probabilities together give the partial likelihood function:

$$L_p(\boldsymbol{\beta}|t_1, t_2, \ldots, t_J, \boldsymbol{z}_i(t)) = \prod_{j=1}^{J} \frac{\Psi_{i_j}(t_j, \boldsymbol{z}_i, \boldsymbol{\beta})}{\sum_{i \in R(t_j)} \Psi_i(t_j, \boldsymbol{z}_i, \boldsymbol{\beta})}. \tag{2.5}$$

Note that although the factor $\lambda_0(t)$ may be of interest, in this example it will play the role of a nuisance parameter, and conveniently disappears from the partial likelihood function due to the multiplicative structure of (2.4).

The important results of Efron [1977] are the efficiency calculations. It is shown that under some regularity conditions, the asymptotic relative efficiency of the partial maximum likelihood estimator (PMLE) compared to the full MLE is one in the example above, even with censoring. This result is achieved by deriving the full likelihood function, and calculating the ratio of the variances of the PMLE and MLE as the sample size increases to infinity. Assuming Ψ_i to be on the form $\Psi_i(t, \boldsymbol{z_i}, \boldsymbol{\beta}) = \exp(\boldsymbol{\beta z}_i(t))$, (2.5) becomes

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{\exp(\boldsymbol{\beta z}_i(t))}{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta z}_i(t))}. \tag{2.6}$$

Further, it is shown by Efron [1977] that the full likelihood of this experiment, when $\lambda_0(t; \boldsymbol{\gamma})$ is a suitable parametrization of the baseline hazard of (2.4), is

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\left(\prod_{j=1}^{J} \frac{\exp(\boldsymbol{\beta z}_i(t))}{\sum_{i \in R(t_j)} \exp(\boldsymbol{\beta z}_i(t))} \right) \\ &\cdot \left(\exp\left(- \int_0^\infty N(t, \boldsymbol{\beta}) H(t, \boldsymbol{\gamma}) \, \mathrm{d}t \right) \prod_{j=1}^{J} N(t_j, \boldsymbol{\beta}) H(t_j, \boldsymbol{\gamma}) \right) \end{aligned} \tag{2.7}$$

where

$$H(t, \boldsymbol{\gamma}) \stackrel{def}{=} \exp(\boldsymbol{\gamma} t),$$

and

$$N(t, \boldsymbol{\beta}) = n \sum_{R(t)} \exp\{\boldsymbol{\beta z}_i(t)\} / \sum_{i=1}^{n} \exp\{\boldsymbol{\beta z}_i(t)\},$$

which, if $\Psi_i(t)$ does not depend on i, is proportional to the number of items at risk at time t. The first term in (2.7) is the partial likelihood, and the second term will be disregarded. We are now in the situation where we know the form of the disregarded term, so we can proceed to calculate the ratio of the limiting variances in the two cases, as Efron [1977] does. The details are not included here, but it is shown that under some general assumptions, the covariance matrix of the PMLE will tend to the covariance matrix of the MLE, which is the inverse of the Fisher information matrix obtained from (2.7). Also, it is shown by examples that even though some of the assumptions are not completely satisfied, such that too heavy censoring, the Cox regression model is close to efficient. In the examples included by Efron [1977], censoring seems to have little effect on the efficiency of partial likelihood estimation. This example will be revisited in Chapter 7.

2.3 Asymptotic evaluations

Before setting off to investigate the asymptotic properties of partial likelihood estimation, it can be wise to ponder the basic framework. The most general definition of partial

likelihood introduced by Cox, requires the researcher of a statistical problem to make a *choice* on how to factorize the full likelihood function. This freedom is an advantage in many applications, allowing one to subjectively adapt the estimation procedure to give the best fit for the problem at hand, within the frames of well-established theory. On the other hand, the lack of a precise definition of partial likelihood, poses a problem when we want to study the behaviour of the estimators when the sample size increases to infinity. There are two obvious ways to attack this problem. Either developing asymptotic theory for each application, or imposing strict enough conditions on the partial likelihood function so that the usual properties can be proven. Both approaches are explored in the next subsections.

2.3.1 AR(1)-process with missing segments

Consider the first example of section 2.2, and suppose the observed segments consist of observations from an AR(1) model, i.e. $Z_{t+1} = \theta Z_t + \epsilon_t$, where Z_{t+1} and Z_t are in the same segment, $-1 < \theta < 1$ and the ϵ_t's are iid $N(0,1)$. The efficiency of the partial likelihood estimator for the unknown parameter depends on how the model behaves in the missing segments, and Wong [1986] considers a few such special cases.

Suppose first that the missing segments also follow an AR(1) model and have length l, while the observed segments have length k. For simplicity, l and k are assumed to be constant. If l/k is negligible, a lower bound for the ARE of $\hat{\theta}$ is close to one. Similarly, if $l \to \infty$, the ARE approaches 1. The last result makes sense because the observed segments are almost independent.

Secondly, suppose that the process shifts by an unknown amount μ_j for each new unobserved segment, i.e. $Z_{t+1} = \theta(Z_t + \mu_j) + \epsilon_t$. Under some mild regularity conditions on the μ_js (uniform boundedness is sufficient), it can actually be shown that the partial MLE is consistent, but the usual MLE is not.

Lastly, Wong [1986] considers the case where the shifts in the unobserved segments are iid random variables with some unknown density function. After some tedious calculations, it is shown that the efficiency of the partial MLE, $\hat{\theta}$, is close to one if the true value of θ is not too close to one and l is large.

All three results above require involved calculations, and further calculations are of course needed if other special cases should be considered.

2.3.2 The Cox regression model

The Cox regression model is studied extensively by, among many others, Aalen et al. [2008], Efron [1977], Kedem and Fokianos [2002], Tsiatis [1981], and Wong [1986]. In particular, it is shown that the asymptotic properties of the partial maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$, are similar to those enjoyed by the MLE based on the full likelihood function. We will here reproduce the results of Tsiatis [1981] on consistency and asymptotic normality of the PMLE in this model. The theorems are proved along the same lines as the corresponding results for the ordinary MLE.

Theorem 2.1. *Let* $\mathbf{Z}$ *denote the vector of covariates in the Cox regression model, and assume that* $E(\mathbf{Z}\exp(\boldsymbol{\beta}\boldsymbol{Z}))$ *is uniformly bounded in a neighbourhood of* $\boldsymbol{\beta}$*. Assume also that* $P(T \geq T_0) > 0$*, where* T *is the survival time of any individual, and* T_0 *is the time at which the study is terminated. Then there exists a sequence of solutions* $\hat{\boldsymbol{\beta}}_n$ *of the score equation* $\nabla L_p(\boldsymbol{\beta}) = \mathbf{0}$ *such that* $\hat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}$

Theorem 2.2. *Assume that* $E(\mathbf{Z}\exp(\boldsymbol{\beta}\boldsymbol{Z}))$ *is uniformly bounded in a neighbourhood of* $\boldsymbol{\beta}$*. Then the statistic* $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *converges in distribution to a multi-normal random variable with expectation zero and covariance matrix equal to* $(I(\boldsymbol{\beta}))^{-1}$*, where* $I(\boldsymbol{\beta})$ *is the information matrix obtained from the partial likelihood function evaluated at the true value* $\boldsymbol{\beta}$*.*

2.3.3 General theory

Wong [1986] establishes some general theory for the partial likelihood. Regularity conditions must be applied to the partial likelihood function in order to prove consistency and asymptotic normality. Let us follow Wong [1986] and start by stating consistency when the parameter space Θ is compact. Let

$$
\begin{aligned}
r_n(\boldsymbol{\theta}) &= \log(f_{\boldsymbol{\theta}_0}(x_n|c_n)/f_{\boldsymbol{\theta}}(x_n|c_n)), & R_N &= \textstyle\sum_{n=1}^N r_n \\
i_n(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}_0}(r_n(\boldsymbol{\theta})|c_n), & I_N &= \textstyle\sum_{n=1}^N i_n \\
j_n(\boldsymbol{\theta}) &= Var_{\boldsymbol{\theta}_0}(r_n(\boldsymbol{\theta})|c_n), & J_N &= \textstyle\sum_{n=1}^N j_n \\
m_n(\boldsymbol{\theta}) &= r_n(\boldsymbol{\theta}) - i_n(\boldsymbol{\theta}), & M_N &= \textstyle\sum_{n=1}^N m_n(\boldsymbol{\theta}).
\end{aligned}
$$

Here, $\prod f_{\boldsymbol{\theta}}(x_n|c_n)$ is the partial likelihood function as defined in (2.2), and $\boldsymbol{\theta}_0$ denotes the true parameter value.

Theorem 2.3. *Suppose* Θ *is compact, and suppose that for any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$*, there exists an open neighbourhood* $O_{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}$ *whose closure* $\overline{O_{\boldsymbol{\theta}}}$ *does not contain* $\boldsymbol{\theta}_0$*, and that there are constants* $\delta > 0$*,* $\alpha_N \uparrow \infty$ *(which may depend on* $\boldsymbol{\theta}$*) such that*

$$
P\left(\inf_{\boldsymbol{\theta}' \in \overline{O_{\boldsymbol{\theta}}}} I_N(\boldsymbol{\theta}')/\alpha_N > \delta\right) \to 1, \tag{2.8}
$$

$$
J_N(\boldsymbol{\theta}')/\alpha_N^2 \xrightarrow{P} 0 \text{ for all } \boldsymbol{\theta}' \in \overline{O_{\boldsymbol{\theta}}}, \tag{2.9}
$$

The distribution of $\alpha_N^{-1} M_N(\boldsymbol{\theta}')$ *is tight*[1] *in* $C(\overline{O_{\boldsymbol{\theta}}})$*, where* $M_N = R_N - I_N$ *and* $C(\overline{O_{\boldsymbol{\theta}}})$ *is the space of continuous functions on* $\overline{O_{\boldsymbol{\theta}}}$*. Then* $\hat{\boldsymbol{\theta}}_N \xrightarrow{P} \boldsymbol{\theta}_0$*.*

Proof. See Wong [1986]. □

[1] For all $\epsilon > 0$, there exists a K such that $\sup_N P(|\alpha_N^{-1} M_N(\boldsymbol{\theta}')| > K) < \epsilon$.

$I_N(\boldsymbol{\theta})$ is the Kullback-Leibler discrimination information, and condition (2.8) ensures that we in the long run accumulate enough information to discriminate between $\prod f_{\boldsymbol{\theta}}$ and $\prod f_{\boldsymbol{\theta}_0}$, while condition (2.9) ensures that the variance of the (partial) likelihood ratio converges in a proper manner. The tightness condition "forces" R_N to approach its expectation I_N, and together with compactness, the theorem can be proved.

It is desirable to ease the condition of compactness in the previous theorem, consider for example $\Theta = \mathbb{R}^k$. This is achieved by imposing the condition that there exists a fixed, compact subset K of Θ such that $\hat{\boldsymbol{\theta}}$ will eventually be contained in K, much in the same way as for the consistency argument of Chapter 1. If this is the case, only slight modifications of Theorem 2.3 are required to prove consistency of the partial maximum likelihood estimator.

This section will now be concluded by stating sufficient conditions for asymptotic normality for the partial maximum likelihood estimator. The theory is again covered in depth by Wong [1986]. Some notation is required; let

$$l_n(\boldsymbol{\theta}) = \log f_{\boldsymbol{\theta}}(x_n|c_n), \qquad L_N = \sum_{n=1}^{N} l_n. \tag{2.10}$$

The vector, matrix and triple array of the first, second and third derivatives of l_n are assumed to exist almost everywhere, and for simplicity we also assume $\Theta \subset \mathbb{R}^k$. Denote the conditional score for the experiment $x_n|c_n$ by $\boldsymbol{u}_n = l_n'(\boldsymbol{\theta}_0)$. Here, c_n has the same meaning as in (2.2). In the same manner as in Section 1.4.4, we have

$$\mathrm{E}(\boldsymbol{U}_n|c_n) = 0, \qquad v_n = \mathrm{Cov}(\boldsymbol{U}_n|c_n) = \mathrm{E}(-l_n''(\boldsymbol{\theta}_0)|c_n). \tag{2.11}$$

Let $\boldsymbol{U}_N = \sum_{n=1}^{N} \boldsymbol{U}_n$ and $\boldsymbol{V}_N = \sum_{n=1}^{N} \boldsymbol{v}_n$.

Theorem 2.4. *Suppose $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}_0 \in$ interior of $\Theta \subset \mathbb{R}^n$, and for each n, l_n has third order derivatives almost surely and (2.11) holds. Assume also that there are constants $a_N \uparrow \infty$ and a neighbourhood O of $\boldsymbol{\theta}_0$ such that*

$$a_N^{-1}\boldsymbol{V}_N \xrightarrow{P} \text{ some positive definite matrix } \boldsymbol{Q}, \tag{2.12}$$

$$a_N^{-1}(-L_N''(\boldsymbol{\theta}_0)) \xrightarrow{P} \text{ some positive definite matrix } \boldsymbol{Q}_1, \tag{2.13}$$

$$P\left(a_N^{-1}\sup_{\boldsymbol{\theta}\in O}|L_N'''(\boldsymbol{\theta})| < M\right) \to 1 \text{ for some constant } M, \tag{2.14}$$

$$a_n^{-3/2}\sum_{n=1}^{N} E(\|\boldsymbol{u}_n\|^3|c_n) \xrightarrow{P} 0. \tag{2.15}$$

Then

$$a_N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{Q}_1^{-1}\boldsymbol{Q}\boldsymbol{Q}_1^{-1}).$$

2.4 Conditional and marginal likelihood

The definition (2.2) of partial likelihood is somewhat loose and unrestrictive. The only thing we demand is that the likelihood function can be factorized, and that the factors of interest carry enough information to do useful analysis. Conditional and marginal likelihoods are special cases of the partial likelihood with stricter definitions.

Let $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)^T$ be a sample with likelihood function $L(\boldsymbol{\theta}, \boldsymbol{X})$, where $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ is the vector of parameters, separated into the parameters of interest, $\boldsymbol{\psi}$, and nuisance parameters $\boldsymbol{\lambda}$. A statistic $T(\boldsymbol{X})$ is *sufficient* for $\boldsymbol{\lambda}$ if the conditional distribution of $\boldsymbol{X}$ given T does not depend on $\boldsymbol{\lambda}$ [Casella and Berger, 2002, p. 272]. If T is sufficient for $\boldsymbol{\lambda}$, but not for $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$ together, then the likelihood can be written

$$L(\boldsymbol{\theta}, \boldsymbol{X}) = f_{\boldsymbol{X}|T}(\boldsymbol{X}|t; \boldsymbol{\psi}) f_T(t; \boldsymbol{\psi}, \boldsymbol{\lambda}),$$

and the *conditional likelihood* is defined to be the first factor;

$$L_{\text{cond}}(\boldsymbol{\psi}|\boldsymbol{X}) = f_{\boldsymbol{X}|T}(\boldsymbol{x}|t; \boldsymbol{\psi}),$$

[Severini, 2000, p. 279]. This is a genuine likelihood function because it is a probability density function regarded as a function of x, but, again, we discard information by analyzing only a part of the full likelihood. For details on this construction, see Kalbfleisch and Sprott [1973].

Suppose now, on the other hand, that there exists a statistic S whose distribution depends only on $\boldsymbol{\psi}$, so the the full likelihood may be written

$$L(\boldsymbol{\psi}, \boldsymbol{\lambda}|\boldsymbol{X}) = f_S(s; \boldsymbol{\psi}) f_{\boldsymbol{X}|S}(\boldsymbol{X}|S; \boldsymbol{\psi}, \boldsymbol{\lambda}).$$

We now extract the marginal distribution of S for further analysis, and denote it the *marginal likelihood based on* S [Severini, 2000, p. 298]. For a simple example of the marginal likelihood, suppose $\boldsymbol{X}$ consists of independent observations from a normal distribution with mean μ and variance σ^2, and that we wish to estimate the latter. We know that $\sum(X_i - \overline{X})^2/\sigma^2$ is chi-squared distributed with $n-1$ degrees of freedom, so that the marginal log-likelihood for σ^2 becomes

$$l_{\text{marg}}(\sigma^2|\boldsymbol{X}) = -\left(\frac{n-3}{2}\right)\log\sigma^2 - \frac{\sum(X_i - \overline{X})^2}{2\sigma^2},$$

with maximum $\widehat{\sigma}^2 = \sum(X_i - \overline{X})^2/(n-3)$. Thus, the marginal likelihood introduces a little more bias than the ordinary maximum likelihood estimator we get when maximizing with respect to both parameters.

There are several other strategies to dealing with nuisance parameters, some of which are discussed by Severini [2000]. These include the *integrated likelihood*, where we integrate out the nuisance parameters with respect to some weight function, and the *profile likelihood* in which the nuisance parameters are replaced by their respective maximum likelihood estimates, calculated while keeping the parameters of interest fixed.

The *local likelihood*, however, modifies the traditional maximum likelihood with a rather different motivation than dealing with nuisance parameters. In the recent years, its applications have been explored in a wide range of diciplines, some of which will be mentioned in later chapters. Our main focus from now on will be probability density estimation using local likelihood, both from a theoretical and practical point of view.

Chapter 3

Introduction to local likelihood

3.1 Motivation and definition of the local likelihood function

Up to this point, we have focused on the parametric aspect of estimation. There are, however, many applications in which a full parametric approach may not be the best path to a good result. For example, suppose you were to estimate an unknown probability density function, but had no prior knowledge of the true density whatsoever. Choosing a parametric family would be plain guesswork and could result in both good, but most certainly also bad conclusions. An alternative approach would then be the traditional non-parametric kernel estimator, $\hat{f}(x) = n^{-1}\sum_{i=1}^{n} K_h(x - x_i)$, where $K_h(z) = h^{-1}K(h^{-1}z)$, $K(z)$ is a unimodal, symmetric density, and n is the total number of observations, with its well known properties, advantages and drawbacks. The purpose of this chapter is to introduce a local likelihood function, which, by varying the bandwidth h, can be interpreted on a continuous scale from the non-parametric kernel estimator, to the fully parametric likelihood function.

As we will see, density estimates calculated using local likelihood are good in practice, but there is usually not much to gain in terms of theoretical performance. The variance is the same as the kernel estimator, and the bias is sometimes better (and sometimes worse). It is nonetheless important to consider, both for the sake of completeness, but also because the local likelihood function is the entry point to more interesting applications, real-world as well as more theoretically motivated.

Let $f(x)$ be our unknown density, and let $\phi(x, \boldsymbol{\theta})$ be a family of densities with a p-dimensional vector of parameters $\boldsymbol{\theta}$. The idea of local likelihood is to approximate f by ϕ *locally*, that is,

$$\hat{f}(x) = \phi(x, \widehat{\boldsymbol{\theta}}(x)).$$

The non-parametric aspect of this estimation is obvious; the choice of parametric family $\phi(x, \boldsymbol{\theta})$ and especially the kernel $K(x)$, should not have too much influence on the esti-

mate $\hat{f}$. Estimating the functions $\theta_1(x), \theta_2(x), \ldots, \theta_p(x)$, however, resemble the parametric approach studied so far. Hjort and Jones [1996] introduce the *local likelihood function*:

$$l_n(x, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \log \phi(X_i, \boldsymbol{\theta}) - \int K_h(x - y) \phi(y, \boldsymbol{\theta}) \, \mathrm{d}y, \tag{3.1}$$

where K_h is some unimodal density symmetric about the origin. Observe that for large h, the local likelihood function is close to $K(0)h^{-1}$ times the ordinary, normalized log-likelihood function, $n^{-1} \sum_{i=1}^{n} \log \phi(X_i, \theta) - 1$, and for small h, the global properties of ϕ will play a very little role in the final result.

One might be tempted to let the local log-likelihood function just be the first term in the above definition. It would certainly be a good candidate at first glance, but simple maximization in the normal case gives the MLEs $\widehat{\mu}(x) = x$ and $\widehat{\sigma}(x) = \infty$, which does not make much sense. To see that (3.1) is a reasonable function to maximize, observe that [Tjøstheim and Hufthammer, 2012]

$$\begin{aligned} \frac{\partial l_n}{\partial \theta_j} &= \frac{1}{n} \sum K_h(x - X_i) u_j(X_i, \boldsymbol{\theta}) - \int K_h(x - y) u_j(y, \boldsymbol{\theta}) \phi(y, \boldsymbol{\theta}) \, \mathrm{d}y \\ &\to \int K_h(x - y) u_j(y, \boldsymbol{\theta}) \{ f(y) - \phi(y, \boldsymbol{\theta}) \} \, \mathrm{d}y \end{aligned}$$

as $n \to \infty$ by the law of large numbers. Here u_j denotes the j'th score function, $\partial / \partial \theta_j \log \phi(x, \boldsymbol{\theta})$. We see that the parameter function $\widehat{\boldsymbol{\theta}}(x)$ satisfying the local score function, $\partial l_n / \partial \theta_j = 0$, requires $\phi(x, \widehat{\boldsymbol{\theta}}(x))$ to be close to the true density $f(x)$.

3.2 Examples of local likelihood estimation

The idea sketched in the previous subsection was first introduced by Tibshirani and Hastie [1987] in regression models by fitting a line locally instead of globally. It was then applied to density estimation by Loader [1996], and also generalized to fit the unknown density $f(x)$ by a low-degree polynomial in a neighbourhood of x. The most general form of local likelihood estimation was introduced by Hjort and Jones [1996], where the locally fitted parametric family is allowed to take any smooth form, and where the Gaussian family has proved especially attractive. This opens up for many interesting applications. Let us briefly mention three of them.

Ordinary linear regression is perhaps one of the most used methods of statistics. In fact, where the subject *statistics* is supposed to be illustrated by a simple figure, perhaps as the icon of a computer program, or at the cover of a book, it is not unusual to see a scatter plot of observations in two variables, and the least squares line estimating the linear relationship between them. Say we observe the pairs (x_i, y_i), $i = 1, \ldots, n$ and that the ys depend on the xs through the equation

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

where α and β are constants, and the ϵ_i's are normally distributed with zero mean and variance equal to σ^2. Simple arguments yield estimators for α and β in the least squares- or maximum likelihood sense. In many situations we do not even have to assume this relationship to exist in order to do interesting statistical analysis. Often, our main objective is to discover if there is any significant growth or decay as x increases, and this can be achieved by testing whether the parameter β is different from zero.

In other cases, however, the shape of the trend can be of significance. Is it exponential? Logarithmic? Or perhaps it is linear after all. In any case, simple linear regression will not answer these questions, as the estimated relationship will be a straight line no matter how the data looks. We can then apply the idea from Tibshirani and Hastie [1987], where the line is fitted locally using only the observations within a certain window, instead of using them all to fit just one line.

Let w be our desired window size, and for simplicity, assume that w is an odd number. Let $y_i = s(x_i) + \epsilon_i$ be our new model, not necessarily linear. For each observation (x_i, y_i), we fit a line using this observation and the $(w-1)/2$ nearest observations on each side measured by the x-values. Call this local line $y_i = m_i(x)$ and let

$$s(x_i) = m_i(x_i).$$

Near the endpoints, we truncate the window corresponding to the points missing. Tibshirani and Hastie [1987] show that the local regression does not suffer from the sometimes severe end effects that arise when applying a simple moving average filter. We will demonstrate this property in a more general setting in Section 6.3.2.

In Figure 3.1 we see local regression in action. We have 51 synthetic, equidistant observations from the function $y = x^3$ plus some Gaussian noise with zero expectation and standard deviation equal to 3. The least squares line captures the growth; $\widehat{\beta}$ is obviously greater than zero, but we also see that the growth is not linear. The local regression model with window size equal to 21 seem to capture the trend in an excellent way.

Bearing in mind that the least squares estimates for α and β equal the corresponding maximum likelihood estimates in a Gaussian model, calling this procedure a local likelihood method is certainly reasonable.

Consider next the problem of two dimensional density estimation, and suppose we want to estimate the density $f(x, y)$ by local likelihood estimation, using the bivariate normal density as our parametric family, that is

$$\begin{aligned}\phi(x, y|\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \\ \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\Bigg\{ -\frac{1}{2(1-\rho^2)}\Bigg[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 \\ - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \Bigg]\Bigg\}.\end{aligned}$$

The multivariate normal distribution is attractive in that the correlation coefficient ρ completely characterizes the dependence between X and Y. The function $\widehat{\rho}(x, y)$ resulting

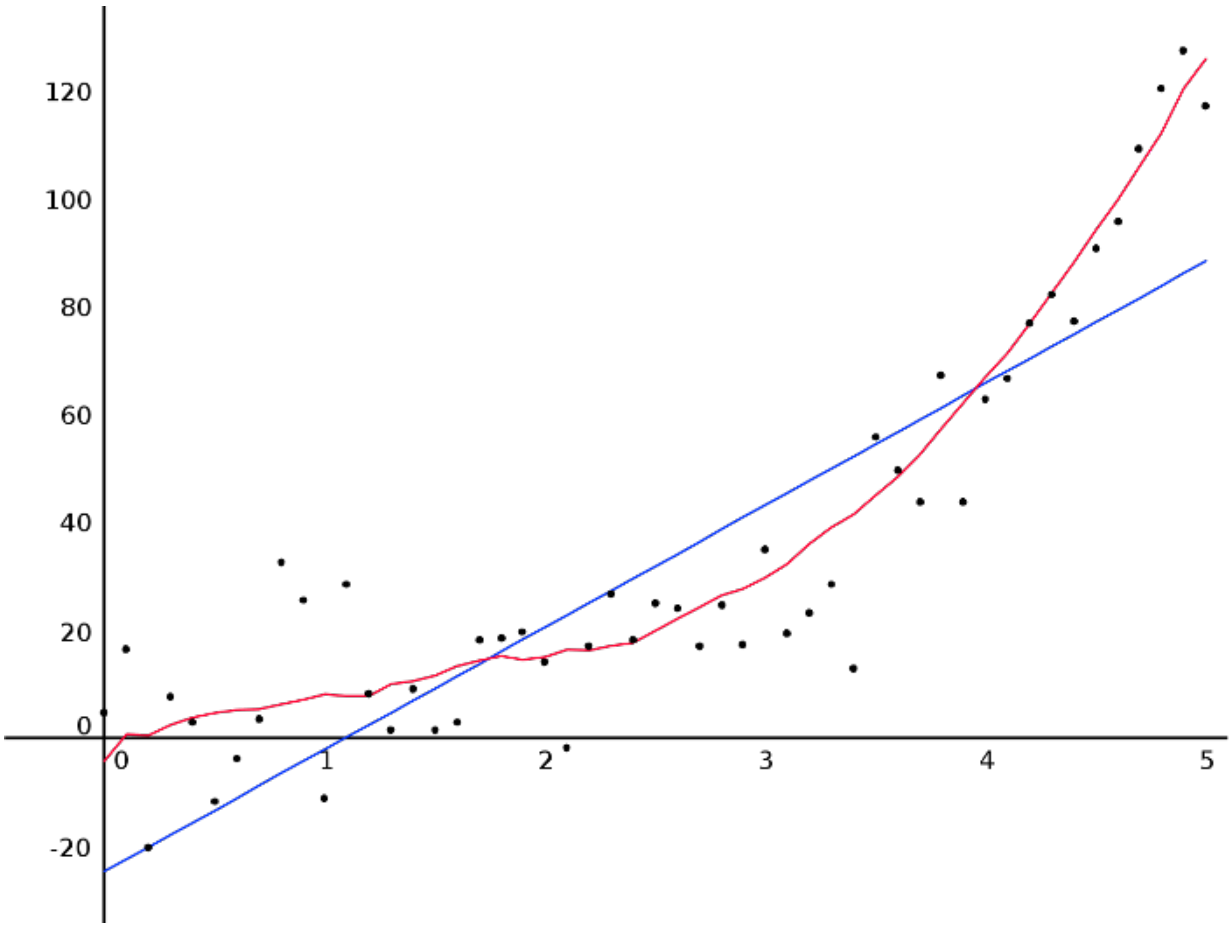


Figure 3.1: Local regression

from a local likelihood estimation, will then be an appealing measure of local correlation between two stochastic variables, see Tjøstheim and Hufthammer [2012].

For a final example, consider the problem of estimating the size of a population of objects in a certain area that can be identified as points on a plane, for example animals or plants. A commonly used method for estimating the intensity D of the population (i.e. the number of individuals per unit area), is line transect sampling. This method consists of two parts. First, straight lines of total length L are randomly drawn across the area of interest, before researchers travel along these lines and record the perpendicular distance from the line to each observed object.

Barabesi [2000] introduces an estimator for D using local likelihood estimation. First, introduce the *detection function* $g(x)$ which is the conditional probability of detecting an object, given that its distance to the line is x. We assume that g is monotonically decreasing and that $g'(0) = 0$. These conditions also apply to the function f, which is the g-function normalized so that it integrates to one. It can be shown that [Burnham and Anderson, 1976]

$$\widehat{D} = \frac{n\widehat{f(0)}}{2L} \tag{3.2}$$

is an unbiased estimator for D if $\widehat{f(0)}$ is an unbiased estimator for $f(0)$. Under some assumptions (e.g. setting the score function $\partial/\partial\theta \log \phi(x, \theta)$ equal to one, justified by Hjort and Jones [1996]), it turns out that the local score equation in this case, evaluated at $x = 0$, is

$$\tilde{f}_{h,K}(0) - \int_0^\infty K_h(t)\phi(t;\theta)(0)\,\mathrm{d}t = 0, \tag{3.3}$$

where $\tilde{f}_{h,K}(0)$ is the ordinary kernel estimator for f evaluated at zero. Zero is a so-called boundary point in this case, which generally leads to consistency problems (see Chapter 6), but by some clever choices for the kernel K and parametric family ϕ, we avoid this problem and show further that we are able to find an explicit estimator for $\theta(0)$, which is the parameter estimate at zero.

Denote by $\psi(x)$ the standard normal distribution and let $K_h(x) = \frac{1}{h}\psi(\frac{x}{h})$. Further, let the parametric family be the so called half-normal density;

$$\phi(x;\theta) = \frac{2}{\theta}\psi\left(\frac{x}{\theta}\right)\mathrm{I}_{[0,\infty)}(x),$$

with one unknown scale parameter θ. In Chapter 6 we will see that since the parametric family ϕ and the true density f share the same boundary $x = 0$, the density estimate will be consistent. The integral in (3.3) is now easy to evaluate using the fact that densities integrate to one. We arrive at the equation

$$\tilde{f}_{h,K}(0) - \frac{1}{\sqrt{2\pi(h^2 + \theta^2(0))}} = 0,$$

which, by solving with respect to $\theta(0)$, yields

$$\widehat{\theta}(0) = \sqrt{\frac{1}{2\pi \tilde{f}_{h,K}(0)^2} - h^2}.$$

Using this result, we get the estimate $\widehat{f(0)} = \phi(0, \widehat{\theta}(0))$, which, in turn, give an estimate of the unknown quantity D through equation (3.2).

3.3 Local likelihood versus the kernel estimator

In the following chapters, we will compare the local likelihood approach to density estimation with the more traditional non-parametric kernel estimator. The concepts are rather different, as we will see below.

The kernel estimator is perhaps the most popular and best understood tool for non-parametric estimation. Suppose we have a vector of one-dimensional observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Then each observation contributes to the estimate with density determined by a symmetric probability density function, K, chosen beforehand:

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h \left(x - x_i \right),$$

where h is a scale parameter in the K-distribution, usually called the bandwidth or smoothing parameter. See Figure 3.2 for a graphical presentation of the kernel estimator. The Gaussian distribution is perhaps the most common choice for the kernel K, but Silverman [1986] shows that minor performance improvements can be made by choosing differently.

The problem of finding the ideal bandwidth h is not trivial, however. Silverman [1986, Section 3.4] discusses this problem in detail. See also section 5.3 on bandwidth selection in connection with the local likelihood case.

Figure 3.3 shows the concept of local maximum likelihood density estimation. For any point x_0 on the x-axis at which we wish to estimate the unknown density, we estimate the parameters of the parametric family locally by maximizing the local likelihood function (3.1), with contribution from each observation determined by the kernel. The resulting estimate, $\widehat{f}(x_0) = \phi(x_0; \widehat{\mu}(x_0), \widehat{\sigma}(x_0))$, in this case using the Gaussian parametric family, will then approximate the true density in a neighbourhood around x_0, and is given in figure 3.3.

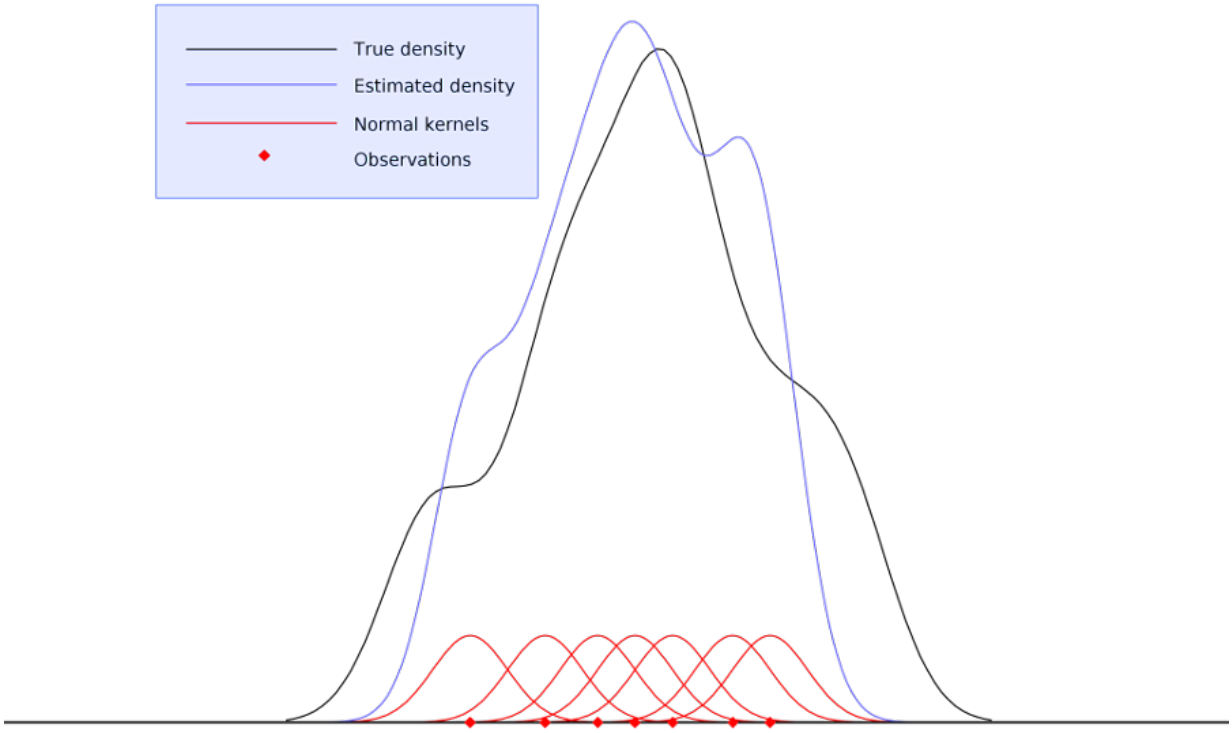


Figure 3.2: Illustration of the Kernel Estimator. The estimate (blue) is defined as the sum of the red kernels, each centred at a datum point. The figure is meant for illustration only. Seven observations will normally not suffice for a good estimate.

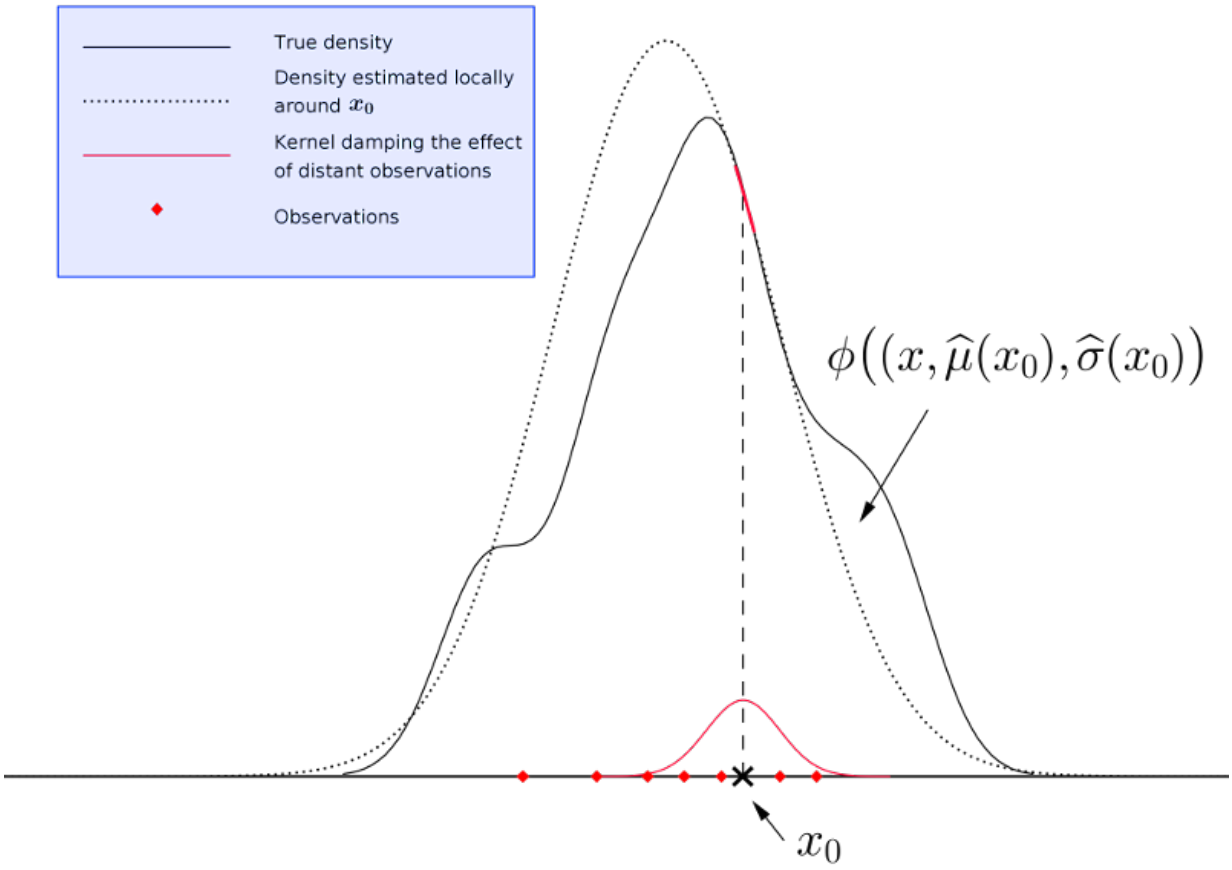


Figure 3.3: Illustration of the Local Maximum Likelihood. We wish to estimate the density at x_0, and maximize the local likelihood function there using the Gaussian distribution as parametric family. The estimate is only valid in a neighbourhood of x_0 (indicated as the red portion of the estimated density). Estimation must be carried out for a selection of points in the interval where we wish to estimate the unknown density. The red kernel indicates that the influence of observations decreases with distance to x_0. Again, the figure is for illustration only. Do not expect seven observations to yield such a good estimate.

Chapter 4

Asymptotic properties

We will now proceed to study local likelihood in more technical detail. Some of the results presented in this section are, for the sake of completeness, merely reproduced here from other sources, while others receive a more thorough treatment than we have found in the literature.

The following results rely on the assumption that there exists a unique solution $\boldsymbol{\theta}_0$ to the system of equations

$$V(x, \boldsymbol{\theta}) = \int K_h(x - y) u_j(y, \boldsymbol{\theta}) \{f(y) - \phi(y, \boldsymbol{\theta})\} \, \mathrm{d}y = 0, \tag{4.1}$$

for $j = 1, \ldots, p$. The number or vector $\boldsymbol{\theta}_0$ will play the role of the 'true' parameter value from now on, and to keep the record straight, we must distinguish between the following quantities:

$f(x)$:	The true, unknown density,
$\widehat{f}(x) = \phi(x, \widehat{\boldsymbol{\theta}}(x))$:	Our density estimate; $\widehat{\boldsymbol{\theta}}$ maximizes (3.1), and
$\phi_0(x) = \phi(x, \boldsymbol{\theta}_0(x))$:	The parametric family with $\boldsymbol{\theta}_0$ as defined above.

The actual existence of a unique $\boldsymbol{\theta}_0$ is yet to be established in general, but this condition essentially requires the true density f to be somewhere within reach of the chosen parametric family. It is nonetheless a reasonable assumption, since the local likelihood function converges in probability to its expectation, namely $V(x, \boldsymbol{\theta})$.

The definition of $\phi_0(x)$ will aid us in developing asymptotic bias expressions, but as a consequence we end up with two sources of bias; one stemming from the approximation of $\mathrm{E}\widehat{f}(x)$ by $\phi_0(x)$, and one from the difference between $f(x)$ and $\phi_0(x)$. The second one is treated by Hjort and Jones [1996] and included in some more detail in Section 4.5. They claim, however, that $\mathrm{E}\widehat{f}(x) = \phi_0(x) + O((nh)^{-1})$ for all dimensions p of the parameter space. It is true for $p = 1$, but we show in section 4.4 that the convergence rate is $O((nh^3)^{-1})$ for $p = 2$. For $p \geq 3$ we no not believe this rate to be any faster, but that is not discussed here.

4.1 Asymptotic normality

Assume (4.1) to hold for some $\boldsymbol{\theta}_0$. Asymptotic normality for the local likelihood estimate is proved much along the same lines as in the ordinary case. The following derivation can be found in Hjort and Jones [1996] and in some more detail in Hufthammer and Tjøstheim [2008b], and the argument is similar to that of the corresponding result in Chapter 1.

First, denote the vector of local score functions by $\boldsymbol{V}_n(\boldsymbol{\theta}) = \{\partial/\partial\theta_j \, l_n(\boldsymbol{\theta}, \boldsymbol{X})\}_{j=1}^p$, where p is the number of parameters in the parametric family. Then, by a first order Taylor expansion, we get

$$\mathbf{0} = \boldsymbol{V}_n(\widehat{\boldsymbol{\theta}}) = \boldsymbol{V}_n(\boldsymbol{\theta}_0) + \nabla \boldsymbol{V}_n(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \boldsymbol{R}.$$

We thus have the approximation

$$(nh)^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\nabla \boldsymbol{V}_n(\boldsymbol{\theta}_0)^{-1}(nh)^{1/2}\boldsymbol{V}_n(\boldsymbol{\theta}_0). \tag{4.2}$$

By performing the p differentiations to obtain an explicit expression for $\boldsymbol{V}_n(\boldsymbol{\theta}_0)$, we obtain

$$\begin{aligned}(nh)^{1/2}\boldsymbol{V}_n(\boldsymbol{\theta}_0) = (nh)^{1/2}\Bigg(& n^{-1}\sum K_h(x - X_i)\boldsymbol{u}(X_i, \boldsymbol{\theta_0}(x)) \\ & - \int K_h(x - y)\boldsymbol{u}(y, \boldsymbol{\theta}_0(x))\phi(y, \boldsymbol{\theta}_0(x)) \, \mathrm{d}y\Bigg),\end{aligned}$$

where $\boldsymbol{u}(x, \boldsymbol{\theta}) = \nabla \log \phi(x, \boldsymbol{\theta})$. Using the Central Limit Theorem, we see that the above expression is normally distributed as $n \to \infty$ when the bandwidth h is held fixed. We know that $\boldsymbol{V}_n(\boldsymbol{\theta}_0)$ converges towards its expectation, and by (4.1), we see that $\mathrm{E}(\boldsymbol{V}_n(\boldsymbol{\theta}_0)) = 0$. For the variance we have

$$\begin{aligned}M_h &\equiv \mathrm{Var}\left((nh)^{1/2}\boldsymbol{V}_n(\boldsymbol{\theta}_0)\right) \\ &= h\Bigg\{\mathrm{E}\Big[K_h(x - X_i)\boldsymbol{u}\left(X_i, \boldsymbol{\theta}_0(x)\right)\Big]\Big[K_h(x - X_i)\boldsymbol{u}^T\left(X_i, \boldsymbol{\theta}_0(x)\right)\Big] \\ &\qquad - \mathrm{E}\Big[K_h(x - X_i)\boldsymbol{u}\left(X_i, \boldsymbol{\theta}_0(x)\right)\Big]\mathrm{E}\Big[K_h(x - X_i)\boldsymbol{u}^T\left(X_i, \boldsymbol{\theta}_0(x)\right)\Big]\Bigg\} \\ &= h\int K_h^2(x - y)\boldsymbol{u}(y, \boldsymbol{\theta}_0(x))\boldsymbol{u}^T(y, \boldsymbol{\theta}_0(x))f(y)\,\mathrm{d}y \\ &\qquad - h\int K_h(x - y)\boldsymbol{u}(y, \boldsymbol{\theta}_0(x))f(y)\,\mathrm{d}y \\ &\qquad\quad \cdot \int K_h(x - y)\boldsymbol{u}^T(y, \boldsymbol{\theta}_0(x))f(y)\,\mathrm{d}y.\end{aligned}$$

We then turn our attention to the leading term of (4.2),

$$\begin{aligned}-\nabla \boldsymbol{V}_n(\boldsymbol{\theta}_0) = -n^{-1}\sum K_h(x - X_i)\nabla \boldsymbol{u}(X_i, \boldsymbol{\theta}_0(x)) \\ + \int K_h(x-y)\Big[\nabla \boldsymbol{u}(y, \boldsymbol{\theta}_0(x))\phi(x, \boldsymbol{\theta}_0(x)) \\ + \boldsymbol{u}(y, \boldsymbol{\theta}_0(x))\nabla\phi(y, \boldsymbol{\theta}_0(x))\Big]\,\mathrm{d}y.\end{aligned}$$

As $n \to \infty$, the above quantity will converge in probability to

$$\begin{aligned}J_h \equiv \int K_h(x-y)\boldsymbol{u}(y, \boldsymbol{\theta}_0(x))\boldsymbol{u}^T(y, \boldsymbol{\theta}_0(x))\phi(y, \boldsymbol{\theta}_0(x))\,\mathrm{d}y \\ - \int K_h(x-y)\nabla \boldsymbol{u}(y, \boldsymbol{\theta}_0(x))\Big[f(y) - \phi(y, \boldsymbol{\theta}_0(x))\Big]\,\mathrm{d}y.\end{aligned}$$

We can finally apply Slutsky's theorem on (4.2) to see that, as $n \to \infty$ with h held fixed,

$$(nh)^{1/2}(\widehat{\boldsymbol{\theta}}(x) - \boldsymbol{\theta_0}(x)) \xrightarrow{d} \mathrm{N}(0, J_h^{-1}M_h(J_h)^{-1})^T). \tag{4.3}$$

This derivation, of course, also goes through in the particular case of one parameter. In that case the quantities J_h and M_h are just numbers instead of matrices and we will avoid a lot of trouble that we will see arise when the parametric model has two or more parameters. In the one parameter case, the delta method gives

$$(nh)^{1/2}(\widehat{f}(x) - \phi(x, \theta_0(x))) \xrightarrow{d} \mathrm{N}(0, \phi(x, \theta_0(x))^2 u(x, \theta_0(x))^2 M_h/J_h^2).$$

4.2 Asymptotic variance for two parameters

For one parameter, we see from the expression above that the asymptotic variance is of order $1/nh$. We now turn to the case of two parameters, which is especially useful because the normal distribution, with its two parameters, seems like a natural choice as parametric family in many cases. The following derivation is merely a reproduction of the argument by Hufthammer and Tjøstheim [2008b], but with a little more attention to the details.

We attack the problem of determining the convergence rate by approximating the matrix $J_h^{-1}M_h(J_h^{-1})^T$ by Taylor expansions for each term, and start out by considering the first term of M_h, denoted by I_M. By making the substitution $s = (x-y)/h$, we have

$$I_M = \int K^2(s)\boldsymbol{u}(x + hs, \boldsymbol{\theta}_0)\boldsymbol{u}^T(x + hs, \boldsymbol{\theta}_0)f(x + hs)\,\mathrm{d}s.$$

We can already now see the reason why this argument is not straightforward. Note that the first order term of the above integral contains the outer product, $\boldsymbol{u}\boldsymbol{u}^T$ as a factor, which is a singular matrix. This expression also appears in J_h, which we must assume invertible. It

is therefore necessary to keep one more term in the expansion of the functions constituting $\boldsymbol{u}\boldsymbol{u}^T$. Keeping only the terms consisting of the scores and their first derivatives, we can write

$$\begin{aligned} I_M &\sim \int K^2(s)\boldsymbol{A}\begin{bmatrix} 1 \\ hs \end{bmatrix}\begin{bmatrix} 1 & hs \end{bmatrix}\boldsymbol{A}^T f(x+hs)\,\mathrm{d}s \\ &= \boldsymbol{A}\begin{bmatrix} 1 & 0 \\ 0 & K_2h^2 \end{bmatrix}\boldsymbol{A}^T f(x) + \boldsymbol{A}\begin{bmatrix} 0 & K_2h^2 \\ K_2h^2 & 0 \end{bmatrix}\boldsymbol{A}^T f'(x) + o(h^2), \end{aligned} \tag{4.4}$$

where the last equality follows from a first order expansion of $f(x+hs)$. We assume that f is differentiable and that $\int |s|^i K^2(s) < \infty$ for $i \leq 3$. The matrix $\boldsymbol{A}$ is here and throughout this chapter defined as

$$\boldsymbol{A} = \begin{bmatrix} u_1(x,\boldsymbol{\theta}_0) & u_1'(x,\boldsymbol{\theta}_0) \\ u_2(x,\boldsymbol{\theta}_0) & u_2'(x,\boldsymbol{\theta}_0) \end{bmatrix}.$$

To arrive at the equality (4.4), we also denote $K_2 \equiv \int s^2K^2(s)\,\mathrm{d}s$ and exploit the fact that $\int s^iK^2(s)\,\mathrm{d}s = 0$ for $i = 1$ and $i = 3$. Further, we note that the second term of M_h is of smaller order because the K-function is not squared, so the whole integral is multiplied by h.

Consider next the first term of J_h, which we denote I_J. Identical calculations as for I_M yield

$$\begin{aligned} I_J &= \int K_h(x-y)\boldsymbol{u}(w,\boldsymbol{\theta}_0)\boldsymbol{u}^T(w,\boldsymbol{\theta})\phi(w,\boldsymbol{\theta}_0)\,\mathrm{d}y \\ &= \boldsymbol{A}\begin{bmatrix} 1 & 0 \\ 0 & K_2h^2 \end{bmatrix}\boldsymbol{A}^T\phi(x,\boldsymbol{\theta}_0) + \boldsymbol{A}\begin{bmatrix} 0 & K_2h^2 \\ K_2h^2 & 0 \end{bmatrix}\boldsymbol{A}^T\phi'(x,\boldsymbol{\theta}_0) + o(h^2) \\ &= \boldsymbol{A}\begin{bmatrix} \phi(x,\boldsymbol{\theta}_0) & \mu_2h^2\phi'(x,\boldsymbol{\theta}_0) \\ \mu_2h^2\phi'(x,\boldsymbol{\theta}_0) & \mu_2h^2\phi(x,\boldsymbol{\theta}_0) \end{bmatrix}\boldsymbol{A}^T, \end{aligned}$$

where $\mu_2 = \int s^2K(s)\,\mathrm{d}s$. For the second term we have, again by a first order Taylor expansion,

$$\begin{aligned} II_J &= \int K_h(x-y)\begin{bmatrix} u_{11}(y,\boldsymbol{\theta}_0) & u_{12}(y,\boldsymbol{\theta}_0) \\ u_{21}(y,\boldsymbol{\theta}_0) & u_{22}(y,\boldsymbol{\theta}_0) \end{bmatrix}\Big(\phi(y,\boldsymbol{\theta}_0) - f(y)\Big)\,\mathrm{d}y \\ &= \boldsymbol{B}\Big(\phi(x,\boldsymbol{\theta}_0) - f(x)\Big) + o(h^2), \end{aligned}$$

where $u_{ij} = \partial u_i/\partial\theta_j$, and where

$$\boldsymbol{B} = \begin{bmatrix} u_{11}(x,\boldsymbol{\theta}_0) & u_{12}(x,\boldsymbol{\theta}_0) \\ u_{21}(x,\boldsymbol{\theta}_0) & u_{22}(x,\boldsymbol{\theta}_0) \end{bmatrix}. \tag{4.5}$$

An important observation here, that will be made clear by Equation (4.15) below when we discuss the bias of the density estimate, is that $\phi(x,\boldsymbol{\theta}_0) - f(x)$ is of order $O(h^2)$ and thus can be expressed as $F(x,\boldsymbol{\theta}_0)h^2$ where $F(x,\boldsymbol{\theta}_0)$ is $O(1)$. Further, and we still follow the

exact argument and notation of Hufthammer and Tjøstheim [2008b], denote by $c_{ij}(x)$ the elements of the matrix $\boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^T$. Then, by direct manipulation, one can verify that

$$J_h \sim \boldsymbol{A}\phi(x,\boldsymbol{\theta}_0)\left[\begin{array}{cc} 1 & 0 \\ 0 & \mu_2 h^2 \end{array}\right]\left\{I_2 + \left[\begin{array}{cc} h^2 a(x) & h^2 b(x) \\ d(x) & e(x) \end{array}\right]\right\}\boldsymbol{A}^T,$$

where I_2 is the two-by-two identity matrix, and

$$a(x) = c_{11}(x)\frac{F(x,\boldsymbol{\theta}_0)}{\phi(x,\boldsymbol{\theta}_0)}, \qquad b(x) = \mu_2\frac{\phi'(x,\boldsymbol{\theta}_0)}{\phi(x,\boldsymbol{\theta}_0)} + c_{12}(x)\frac{F(x,\boldsymbol{\theta}_0)}{\phi(x,\boldsymbol{\theta}_0)},$$

$$d(x) = \frac{\phi'(x,\boldsymbol{\theta}_0)}{\phi(x,\boldsymbol{\theta}_0)} + \mu_2 c_{21}(x)\frac{F(x,\boldsymbol{\theta}_0)}{\phi(x,\boldsymbol{\theta}_0)}, \qquad e(x) = \mu_2 c_{22}(x)\frac{F(x,\boldsymbol{\theta}_0)}{\phi(x,\boldsymbol{\theta}_0)}.$$

By inverting this matrix, it follows that

$$J_h^{-1} \sim \frac{1}{\phi(x,\boldsymbol{\theta}_0)}(\boldsymbol{A}^T)^{-1}\left\{\left[\begin{array}{cc} 1 & 0 \\ 0 & \frac{1}{\mu_2 h^2} \end{array}\right] - \left[\begin{array}{cc} -e(x) & \frac{b(x)}{\mu_2} \\ d(x) & -a(x) \end{array}\right]\right\}\boldsymbol{A}^{-1},$$

which is simplified by the fact that the determinant is $O(1)$ as $h \to 0$. An expression for the covariance matrix, which is correct up to the order of the bandwidth h, is:

$$\begin{aligned} J_h^{-1}M_h(J_h^{-1})^T &\sim \frac{f(x)}{\phi^2(x,\boldsymbol{\theta}_0)}(\boldsymbol{A}^T)^{-1}\{\}_1\boldsymbol{A}^{-1}\boldsymbol{A}\{\}_2\boldsymbol{A}^T(\boldsymbol{A}^T)^{-1}\{\}_3\boldsymbol{A}^{-1} \\ &= \frac{f(x)}{\phi^2(x,\boldsymbol{\theta}_0)}(\boldsymbol{A}^T)^{-1}\{\}_1\{\}_2\{\}_3\boldsymbol{A}^{-1}, \end{aligned} \tag{4.6}$$

where

$$\{\}_1 = \{\}_3^T = \left[\begin{array}{cc} 1 & 0 \\ 0 & \frac{1}{\mu_2 h^2} \end{array}\right] - \left[\begin{array}{cc} -e(x) & \frac{b(x)}{\mu_2} \\ d(x) & -a(x) \end{array}\right]$$

and

$$\{\}_2 = \left[\begin{array}{cc} 1 & 0 \\ 0 & K_2 h^2 \end{array}\right] + \left[\begin{array}{cc} 0 & K_2 h^2 \\ K_2 h^2 & 0 \end{array}\right]\frac{f'(x)}{f(x)}.$$

It is straightforward to calculate the matrix product in (4.6), either by hand, taking only the h-order into account, or by using a symbolic software package such as Maple. In any case, the resulting matrix turns out to be $O(h^{-2})$, and thus the variance of the parameter estimate converges at the somewhat slower rate of $O(1/nh^3)$ in the case of two parameters, compared to the one parameter case, in which the variance converges as $1/nh$. The equivalent convergence rate for the density estimate is $O((nh)^{-1})$, however, as we now proceed to establish. A first order Taylor expansion yields

$$\begin{aligned} \mathrm{Var}\widehat{f}(x) &\sim \mathrm{E}(\phi(x,\widehat{\boldsymbol{\theta}}) - \phi(x,\boldsymbol{\theta}_0))^2 \\ &\sim (\phi'(x,\boldsymbol{\theta}_0))^2\boldsymbol{u}^T(x,\boldsymbol{\theta}_0)\mathrm{E}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\boldsymbol{u}(x,\boldsymbol{\theta}_0) \\ &\sim (\phi'(x,\boldsymbol{\theta}_0))^2\boldsymbol{u}^T J_h^{-1}M_h(J_h^{-1})^T\boldsymbol{u}(x,\boldsymbol{\theta}_0), \end{aligned}$$

but observe that $\boldsymbol{u}(x, \boldsymbol{\theta}_0) = [\ 1 \quad 0\ ]\boldsymbol{A}^T$, and by using this as well as the expression we have already found for $J_h^{-1} M_h (J_h^{-1})^T$, we see that the only dependence on h in this expression is found in the upper left corner of the matrix $\{\}_1\{\}_2\{\}_3$, which is $O(1)$, as easily derived from (4.6). The variance of the density estimate consequently has the same rate as in the one-parameter case.

4.3 Asymptotic bias of $\widehat{f}(x)$ relative to $\phi_0(x)$ for one parameter

In this section, we will see that $\mathrm{E}\widehat{f}(x) = \phi_0(x) + O((nh)^{-1})$ if the parameter has dimension one. This will prove useful later on, when the bias of $\widehat{f}(x)$ relative to the true density $f(x)$ is the subject of investigation. The following argument goes along the same lines as the derivation by Cox and Snell [1968] in the ordinary likelihood case, but with necessary adjustments.

Our starting point will be the local likelihood function (3.1). For large n, we have the following second order approximation of the score equation:

$$l_n'(\widehat{\theta}) = l_n'(\theta_0) + (\widehat{\theta} - \theta_0) l_n''(\theta_0) + \frac{1}{2}(\widehat{\theta} - \theta_0)^2 l_n'''(\theta_0) + R = 0, \tag{4.7}$$

where R is the remainder term from the Taylor expansion. By taking expectation in the above expression, we have from equation (4.1) that the leading term disappear. Since $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$,

$$\begin{aligned} &\mathrm{E}(\widehat{\theta} - \theta_0)\mathrm{E}(l_n''(\theta_0)) + \mathrm{Cov}(\widehat{\theta} - \theta_0, l_n''(\theta_0)) \\ &\quad + \frac{1}{2}\mathrm{E}(\widehat{\theta} - \theta_0)^2 \mathrm{E}(l_n'''(\theta_0)) + \mathrm{Cov}(\frac{1}{2}(\widehat{\theta} - \theta)^2, l_n'''(\theta_0)) + \mathrm{E}(R) = 0, \end{aligned} \tag{4.8}$$

and we proceed by assessing the size of each term.

- The first term contains the bias of $\widehat{\theta}$, in which we are interested. It is multiplied by the expected second derivative of the local likelihood function. By differentiating l_n with respect to the parameter and inserting θ_0, we arrive at

$$\begin{aligned} \mathrm{E}l_n'(\theta_0) &= \mathrm{E}\left(\frac{1}{n} \sum_{i=1}^n K_h(x - X_i) u_0 \right) - \int K_h(x - y) u_0 \phi_0 \, \mathrm{d}y \\ &= \int K_h(x - y) u_0 (f - \phi_0) \, \mathrm{d}y, \end{aligned} \tag{4.9}$$

  By (4.15) below, it follows that $l_n'(\theta_0) = O(h^2)$ as $h \to 0$. For the second derivative,

we have

$$\begin{aligned}\mathrm{E}l_n''(\theta_0) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)u_0'\right) - \int K_h(x-y)\phi_0 u_0' \,\mathrm{d}y - \int K_h(x-y)\phi_0 u_0^2 \,\mathrm{d}y \\ &= \int K_h(x-y)u_0'(f-\phi_0)\,\mathrm{d}y - \int K_h(x-y)\phi_0 u_0^2 \,\mathrm{d}y,\end{aligned}$$

The first integral vanishes as $O(h^2)$ when $h \to 0$ due to the convergence of $f - \phi_0$. The second integral, however, has a finite first order term, $(\phi_0 u_0^2)(x)$, that does not disappear. The second derivative is therefore $O(1)$ asymptotically.

- For the second term, using Schwarz' inequality, we have

$$\mathrm{Cov}(\widehat{\theta} - \theta_0, l_n'''(\theta_0)) \leq \sqrt{\mathrm{Var}(\widehat{\theta})}\sqrt{\mathrm{Var}(l_n''(\theta_0))}, \tag{4.10}$$

  where, from (4.3), the variance of $\widehat{\theta}$ is of order $(nh)^{-1}$. Further, $\mathrm{Var}(l_n''(\theta_0)) = O((nh)^{-1})$ as we will see in Equation (4.12). Together with (4.10), it follows that the second term of (4.8) has a convergence rate not slower than $1/nh$.

- The third term is essentially the variance of $\widehat{\theta}$, which we have already established converges with a rate of $(nh)^{-1}$, multiplied with the expected third derivative, which by similar claculations as that of the second derivative, is $O(1)$.

- The fourth term does not converge any slower than the second, since, using Schwarz' inequality again, it contains $\mathrm{Var}(\widehat{\theta}^2)$, which is seen to have the same order as $\mathrm{Var}(\widehat{\theta})$ when applying the delta method. We neglect the remainder and assume that it is of higher order than the preceeding terms.

The dominating term on the right hand side is therefore $O((nh)^{-1})$ which is the convergence rate for the asymptotic bias of the parameter estimate. Since the delta method here applies without complications, it follows that $\mathrm{E}\widehat{f}(x) = \phi_0(x) + O((nh)^{-1})$.

4.4 Asymptotic bias of $\widehat{f}(x)$ relative to $\phi_0(x)$ for two parameters

Increasing the number of parameters from one to two, created complications in large sample variance calculations due to singular matrices appearing in asymptotic expressions. That is the case for asymptotic bias as well. We start by introducing some notation. Let the three first derivatives of the local likelihood function be given by

$$U_i(\boldsymbol{\theta}) = \frac{\partial l_n}{\partial \theta_i}, \qquad V_{ij}(\boldsymbol{\theta}) = \frac{\partial^2 l_n}{\partial \theta_i \partial \theta_j}, \qquad W_{ijk}(\boldsymbol{\theta}) = \frac{\partial^3 l_n}{\partial \theta_i \partial \theta_j \partial \theta_k}.$$

The matrix of expected second derivatives is denoted by $\boldsymbol{I} = \mathrm{E}\{-V_{ij}\}_{i=1,2,\ j=1,2}$, while the two matrices of expected third derivatives are denoted by $\boldsymbol{J}_i = \mathrm{E}\{W_{ijk}\}_{j=1,2,\ k=1,2}$ for $i = 1, 2$. Let also

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{bmatrix}, \qquad \boldsymbol{\theta}_0 = \begin{bmatrix} \theta_{0,1} \\ \theta_{0,2} \end{bmatrix}.$$

We proceed by expanding the two components of the score function, making the necessary smoothness assumptions on the local likelihood function. For component i, it follows that

$$\begin{aligned} 0 &= U_i(\widehat{\boldsymbol{\theta}}) \\ &= U_i(\boldsymbol{\theta}_0) + \sum_{j=1}^{2}(\theta_{0,j} - \widehat{\theta}_j)V_{ij}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{2}(\theta_{0,j} - \widehat{\theta}_j)(\theta_{0,k} - \widehat{\theta}_k)W_{ijk}(\boldsymbol{\theta}_0) + R. \end{aligned}$$

Upon taking expectations and applying the well known identity that describes the relationship between expectations and covariances, we arrive at the following equality;

$$\begin{aligned} 0 = \sum_{j=1}^{2}\Big[\mathrm{E}(\widehat{\theta}_j - \theta_{0,j})\mathrm{E}(V_{ij}) + \mathrm{Cov}(\widehat{\theta}_j - \theta_{j,0}, V_{ij})\Big] \\ + \frac{1}{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\Big[\mathrm{E}(\widehat{\theta}_j - \theta_{0,j})(\widehat{\theta}_k - \theta_{0,k})\mathrm{E}(W_{ijk}) \\ + \mathrm{Cov}\Big((\widehat{\theta}_j - \theta_{0,j})(\widehat{\theta}_k - \theta_{0,k}), W_{ijk}\Big)\Big] + \mathrm{E}(R), \end{aligned}$$

which we can rewrite using matrix notation when including both components,

$$\begin{aligned} \boldsymbol{I}\mathrm{E}\begin{bmatrix} \widehat{\theta}_1 - \theta_{0,1} \\ \widehat{\theta}_2 - \theta_{0,2} \end{bmatrix} \sim \begin{bmatrix} \mathrm{Cov}\left(\widehat{\theta}_1 - \theta_{0,1}, V_{11}\right) + \mathrm{Cov}\left(\widehat{\theta}_2 - \theta_{02}, V_{12}\right) \\ \mathrm{Cov}\left(\widehat{\theta}_1 - \theta_{0,1}, V_{21}\right) + \mathrm{Cov}\left(\widehat{\theta}_2 - \theta_{02}, V_{22}\right) \end{bmatrix} \\ + \frac{1}{2}\begin{bmatrix} \mathrm{Tr}\left(\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{J}_1\right) \\ \mathrm{Tr}\left(\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{J}_2\right) \end{bmatrix} \\ + \frac{1}{2}\begin{bmatrix} \sum_{jk}^{2}\mathrm{Cov}\left((\widehat{\theta}_j - \theta_{0,j})(\widehat{\theta}_k - \theta_{0,j}), W_{1jk}\right) \\ \sum_{jk}^{2}\mathrm{Cov}\left((\widehat{\theta}_j - \theta_{0,j})(\widehat{\theta}_k - \theta_{0,j}), W_{2jk}\right) \end{bmatrix} + \mathrm{E}(\boldsymbol{R}). \quad (4.11) \end{aligned}$$

Again, we need to discuss the order of each of the factors above.

- $\boldsymbol{I}$ is the matrix of expected second derivatives, whose element (i, j) is given by

$$\begin{aligned} \mathrm{E}\frac{\partial l_n}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta}_0) = \mathrm{E}\left(\sum_{i=1}^{n}K_h(x - x_i)\frac{\partial u_i}{\partial\theta_j}\right) - \int K_h(x - y)\frac{\partial u_i}{\partial\theta_j}\phi_0\,\mathrm{d}y \\ - \int K_h(x - y)u_j u_i\phi_0\,\mathrm{d}y, \end{aligned}$$

and the first two terms cancel each other out as $O(h^2)$ again. The last term is the element of a matrix containing the singular matrix $\boldsymbol{u}\boldsymbol{u}^T$ as a factor, and from the arguments of Section 4.2, such matrices are $O(h^2)$ as well. Thus, the matrix of expected second derivatives is $O(h^2)$ in the two-parameter case.

- The terms on the form $\mathrm{Cov}(\widehat{\theta}_i - \theta_{0,i}, V_{ij})$ for $i, j = 1, 2$ can be considered using Schwarz' inequality:

$$\mathrm{Cov}(\widehat{\theta}_i - \theta_{0,i}, V_{ij}) \leq \sqrt{\mathrm{Var}(\widehat{\theta}_i - \theta_{0,i})}\sqrt{\mathrm{Var}(V_{ij})}$$

The variance of the parameter is known to be $O((nh^3)-1)$. The variance of the second derivatives can be calculated as follows;

$$\begin{aligned}
\mathrm{Var}(V_{ij}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)\frac{\partial u_i}{\partial \theta_j}(x)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathrm{E}\left(\frac{1}{h}K\left(\frac{x - x_i}{h}\right)\frac{\partial u_i}{\partial \theta_j}(x)\right)^2 - \left(\mathrm{E}\left(\frac{1}{h}K\left(\frac{x - x_i}{h}\right)\frac{\partial u_i}{\partial \theta_j}(x)\right)\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[\frac{1}{h}\int K^2(s)\left(\frac{\partial u_i}{\partial \theta_j}(x + sh)\right)^2 f(x + sh)\,\mathrm{d}s\right. \\
&\qquad \left. - \left(\int K(s)\frac{\partial u_i}{\partial \theta_j}(x + sh) f(x + sh)\right)\mathrm{d}s\right] \\
&\sim \frac{1}{nh} - \frac{1}{n} \sim \frac{1}{nh},
\end{aligned} \tag{4.12}$$

and this is really valid for all order of derivatives, and all number of parameters. Thus, the first term on the right side of (4.11) is asymptotically bounded above by the inequality

$$\mathrm{Cov}(\widehat{\theta}_i - \theta_{0,i}, V_{ij}) \leq \sqrt{1/(nh^3)}\sqrt{1/(nh)} = 1/(nh^2).$$

- The covariance matrix of $\widehat{\boldsymbol{\theta}}$ is of order $O((nh^3)^{-1}))$. The two matrices of expected third derivatives, $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ are given by

$$\begin{aligned}
\mathrm{E}\frac{\partial^3 l_n}{\partial\theta_i\partial\theta_j\partial\theta_k}(\boldsymbol{\theta_0}) = \mathrm{E}&\left(\sum_{i=1}^{n} K_h(x - x_i)\frac{\partial^2 u_i}{\partial u_j \partial u_k}\right) - \int K_h(x - y)\frac{\partial^2 u_i}{\partial\theta_j\partial\theta_k}\phi_0\,\mathrm{d}y \\
&- \int K_h(x - y)\frac{\partial u_i}{\partial\theta_j}u_k\phi_0\,\mathrm{d}y - \int K_h(x - y)u_i u_j u_k\phi_0\,\mathrm{d}y \\
&- \int K_h(x - y)\frac{\partial u_i}{\partial\theta_k}u_j\phi_0\,\mathrm{d}y - \int K_h(x - y)u_i\frac{\partial u_j}{\partial\theta_k}\phi_0\,\mathrm{d}y,
\end{aligned}$$

where i denotes the matrix $\boldsymbol{J}_i$, and the indices j and k denotes the elements of that matrix. The two leading terms are $O(h^2)$. The next three terms are all elements of

matrices on the form $\int K_h(x-y)\boldsymbol{v}\boldsymbol{w}^T\phi_0\,\mathrm{d}y$, where $\boldsymbol{v}$ and $\boldsymbol{w}$ are vectors containing various combinations of differentiations of u. Thus they are all $O(h^2)$, following similar arguments as that of Section 4.2. The last term, however, can not be treated this way automatically, as the matrix $\{\partial u_j/\partial\theta_k\}_{j,k=1,2}$ is not singular and can not be written as an outer product. Comparing its elements with the preceeding matrices, however, reveals that

$$\boldsymbol{J}_1 \sim \begin{pmatrix} O(1) & O(h^2) \\ O(1) & O(h^2) \end{pmatrix}, \qquad \boldsymbol{J}_2 \sim \begin{pmatrix} O(h^2) & O(h^2) \\ O(h^2) & O(h^2) \end{pmatrix}. \tag{4.13}$$

- The last term can be shown to be at least as fast as the first by using the same argumentation as in the one-parameter case. We assume that the remainder is dominated by $1/nh^3$ asymptotically.

Upon left-multiplying (4.11) with $\boldsymbol{I}^{-1}$, the asymptotic bias turns out to be $O((nh^3)^{-1})$, that is, the same order as the asymptotic variance.

Recall that, due to some 'lucky' cancellations, the asymptotic variance of *density estimates*, was not affected by the slower convergence rate established for *parameter* estimates. This does not happen for the asymptotic bias, however. We have the following Taylor expansion;

$$\widehat{f}(x) = \phi(x,\widehat{\boldsymbol{\theta}}) = \phi(x,\boldsymbol{\theta}_0) + (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\nabla\phi(x,\boldsymbol{\theta}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\nabla\nabla\phi(x,\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + R,$$

where R is of higher order. Taking expectation, we arrive at the following expression for the bias:

$$\begin{aligned}\mathrm{E}\left[\widehat{f}(x) - \phi(x,\boldsymbol{\theta}_0)\right] \sim\ & \phi(x,\boldsymbol{\theta}_0)\boldsymbol{u}^T(x,\boldsymbol{\theta}_0)\mathrm{E}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \\ & + \frac{1}{2}\mathrm{E}\left[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T(\nabla\phi_0\boldsymbol{u}^T + \phi_0\boldsymbol{B})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right],\end{aligned}$$

where B is the matrix of second derivatives of ϕ, defined by Equation (4.5). Everything involving the parametric family is non-stochastic and independent of h, so the dominating term turns out to be $\mathrm{E}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$, that is, the density estimates share the same slower convergence rates of order $O((nh^3)^{-1})$ as the parameter estimates.

4.5 Asymptotic bias of $\phi_0(x)$ relative to $f(x)$

In this section, we will examine the bias arising due to the difference between $f(x)$ and $\phi_0(x)$.

Hjort and Jones [1996] show that the asymptotic variance in local likelihood density estimation is equal to the variance for the kernel estimator. The bias, although similar, is

shown to differ from its kernel estimator equivalent. The derivation of asymptotic bias is sketched by Hjort and Jones, and the details are included below.

Generally, the unknown parameter $\boldsymbol{\theta}$ is p-dimensional. Denote the score function by

$$\mathbf{u}(x,\boldsymbol{\theta}) = \begin{bmatrix} u_1(x,\boldsymbol{\theta}) \\ u_2(x,\boldsymbol{\theta}) \\ \vdots \\ u_p(x,\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\phi(x,\boldsymbol{\theta}) \\ \frac{\partial}{\partial\theta_2}\phi(x,\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial\theta_p}\phi(x,\boldsymbol{\theta}) \end{bmatrix}.$$

By a standard Taylor series argument it follows that, as $h \to 0$,

$$\int K_h(t-x)g(t)\,\mathrm{d}t = g(x) + \frac{1}{2}K_2h^2g''(x) + O(h^4) \tag{4.14}$$

for any smooth function $g(\cdot)$ and kernel K, and where $K_2 = \int s^2K(s)\,\mathrm{d}s$ as before. To see this, let

$$F(h) = \int K_h(t-x)g(t)\,\mathrm{d}t = \int \frac{1}{h}K\left(\frac{t-x}{h}\right)g(t)\,\mathrm{d}t$$

and make the substitution $z = (t-x)/h$. We then have $\mathrm{d}t = h\mathrm{d}z$, so

$$F(h) = \int K(z)g(x+zh)\,\mathrm{d}t.$$

The quantity $F(h)$ is next expanded about zero;

$$F(h) = F(0) + F'(0)h + \frac{1}{2}F''(0)h^2 + \frac{1}{6}F'''(0)h^3 + O(h^4).$$

$F'(0) = F'''(0) = 0$ because K is even and the functions z and z^3 are odd. It remains that

$$\begin{aligned} F(0) &= \int K(z)g(x)\,\mathrm{d}z = g(x)\int K(z)\,\mathrm{d}z = g(x) \\ F''(0) &= \int z^2K(z)g''(x)\,\mathrm{d}z = K_2g''(x), \end{aligned}$$

from which (4.14) follows. Letting $g(x) = u(x,\theta)(f(x) - \phi_0(x))$, we easily see from (4.14) and (4.1) that, in general, $\phi(x,\theta_0(x)) - f(x) = O(h^2)$:

$$\begin{aligned} 0 &= \int K_h(t-x)u(t,\theta)(f(t) - \phi(t,\theta_0(x))\,\mathrm{d}t \\ &= \int K(z)u(x+zh,\theta)(f(x+zh) - \phi(x+zh,\theta_0(x))\,\mathrm{d}z \\ &= u(x,\theta)(f(x) - \phi(x,\theta_0(x))) + O(h^2). \end{aligned} \tag{4.15}$$

By including another term in the Taylor expansion above, we have under smoothness assumptions on f and the weight functions u_j, and by writing $\phi_0(x) = \phi(x, \theta_0(x))$ and $u_{j,0} = u_j(x, \theta_0)$, that

$$u_{j,0}(x)(\phi_0(x) - f(x)) = \frac{1}{2}K_2 h^2 \{u_{j,0}(f - \phi_0)\}''(x) + O(h^4). \tag{4.16}$$

Dividing through with $u_{j,0}$ and applying the bias results from preceding sections, we have

$$E\widehat{f}(x) = f(x) + \frac{1}{2}K_2 h^2 b_j(x) + O(h^4 + (nh^c)^{-1}),$$

where

$$b_j(x) = \{u_{j,0}(f - \phi_0)\}''(x)/u_{j,0}(x) \tag{4.17}$$

and where c depends on the number of parameters in the parametric family.

4.5.1 The one-parameter case

To proceed further, we need to investigate the bias term by first specifying the number of parameters in the parametric family. Do note that the bias shows a striking resemblance to the bias of ordinary kernel estimation where $b_j(x) = f''(x)$. Finding, and discussing the nature of $b_j(x)$ compared to $f''(x)$ for various cases (i.e. different dimensions of the parameter θ), is imperative in this context, as it seems to be our only possibility of better performance. In the one-parameter case,

$$b_j(x) = f''(x) - \phi_0''(x) + 2\frac{u_0'(x)}{u_0(x)}(f'x) - \phi_0'(x)). \tag{4.18}$$

This follows by performing the differentiation in (4.17). The term containing $(f - \phi_0)$ seemingly disappears, but as we have shown earlier it is of order $O(h^2)$, and thus can be included in the $O(h^4)$-term. As we see, closeness between f and ϕ_0 may reduce the bias compared with the kernel estimator. Indeed, if the unknown density is a member of our parametric family, $b_j(x)$ will vanish altogether, and we will be left with the $O((nh^3)^{-1})$-term.

4.5.2 The two-parameter case

Let the parametric family be given by $\phi_{\boldsymbol{\theta}}(x) = \phi(x, \theta_1, \theta_2)$. From (4.16) we see that

$$\phi_0(x) - f(x) = \frac{1}{2}\sigma_K^2 h^2 \left(f''(x) - \phi_0''(x) + 2\frac{u_{j,0}'(x)}{u_{j,0}(x)}(f'(x) - \phi_0'(x)) \right) \tag{4.19}$$

for $j = 1, 2$. Note that the left side of (4.19) is independent of j, but the right side is not. Thus, equality can only hold if $f'(x) - \phi'(x) = o(h)$ as $h \to 0$. Taking the results of Section 4.4 into account, we can therefore write

$$E\widehat{f}(x) = f(x) + \frac{1}{2}\sigma_K^2 h^2 (f''(x) - \phi_0''(x)) + O(h^3 + (nh^3)^{-1}) \tag{4.20}$$

in the case of two parameters.

4.5.3 Three or more parameters

Hjort and Jones [1996] proceed by investigating the asymptotic bias when we introduce even more local parameters. This sub-section fills in some of the details that Hjort and Jones did not include in their paper. This discussion is really only valid for the difference between f and ϕ_0, though, as neither we, nor Hjort and Jones, consider the difference between $\mathrm{E}\widehat{f}$ and ϕ_0 for three or more parameters. In light of Section 4.4, however, we believe that this difference is not smaller than $O((nh^3)^{-1})$, that was established for two parameters, contrary to $O((nh)^{-1})$ as claimed by Hjort and Jones.

For $p \geq 3$, where p denotes the number of parameters, we can make further improvements by deriving that $(f - \phi_0)''$ is $o(1)$. This is done by differentiating (4.16) four times and writing $g_r \equiv (f - \phi_0)^{(r)} \approx \sum_{i=0}^{4-r} a(r,i)h^i$. In order to obtain enough equations to solve for $a(i,j)$, we approximate $V(x,\theta)$ in (4.1) by a sixth order Taylor expansion:

$$\begin{aligned} &g_0 + \frac{1}{2}k_2h^2(u_{j,0}g_0)''/u_{j,0} \\ &\quad + \frac{1}{24}k_4h^4(u_{j,0}g_0)^{(4)}/u_{j,0} + \frac{1}{720}k_6h^6(u_{j,0}g_0)^{(6)}/u_{j,0} = 0, \end{aligned} \tag{4.21}$$

where $k_i = \int z^i K(z)\,dz$. We proceed by performing the differentiations in the above equations, collecting terms in equal powers of h, and then equating each coefficient of h^i with zero. After some calculations, we see that these coefficients are:

$$\begin{aligned}
1 &: a(0,0) \\
h &: a(0,1) \\
h^2 &: \frac{1}{2}k_2a(2,0) + a(0,2) + k_2a(1,0)\frac{u'_{j,0}}{u_{j,0}} + \frac{1}{2}k_2a(0,0)\frac{u''_{j,0}}{u_{j,0}} \\
h^3 &: \frac{1}{2}k_2a(2,1) + a(0,3) + k_2a(1,1)\frac{u'_{j,0}}{u_{j,0}} + \frac{1}{2}k_2a(0,1)\frac{u''_{j,0}}{u_{j,0}} \\
h^4 &: \frac{1}{2}k_2a(2,2) + \frac{1}{24}k_4a(4,0) + a(0,4) + \left(k_2a(1,2) + \frac{1}{6}k_4a(3,0)\right)\frac{u'_{j,0}}{u_{j,0}} \\
&\quad + \left(\frac{1}{2}k_2a(0,2) + \frac{1}{4}k_4a(2,0)\right)\frac{u''_{j,0}}{u_{j,0}} \\
h^5 &: + \left(\frac{1}{2}k_2a(1,3) + \frac{1}{6}k_4a(3,1)\right)\frac{u'_{j,0}}{u_{j,0}} + \left(\frac{1}{2}k_2a(0,3) + \frac{1}{4}k_4a(2,1)\right)\frac{u''_{j,0}}{u_{j,0}} \\
&\quad + \frac{1}{6}k_4a(1,1)\frac{u^{(3)}_{j,0}}{u_{j,0}} + \frac{1}{24}k_4a(0,1)\frac{u^{(4)}_{j,0}}{u_{j,0}}
\end{aligned}$$

$$h^6 : g_6 + g_5\frac{u'_{j,0}}{u_{j,0}} + \left(\frac{1}{2}k_2a(0,4) + \frac{1}{4}k_4a(2,2) + \frac{1}{48}k_6a(4,0)\right)\frac{u''_{j,0}}{u_{j,0}}$$
$$+ \left(\frac{1}{6}k_4a(1,2) + \frac{1}{36}k_6a(3,0)\right)\frac{u^{(3)}_{j,0}}{u_{j,0}} + \left(\frac{1}{2}k_4a(0,2) + \frac{1}{48}k_6a(2,0)\right)\frac{u^{(4)}_{j,0}}{u_{j,0}}$$
$$+ \frac{1}{120}k_6a(1,0)\frac{u^{(5)}_{j,0}}{u_{j,0}} + \frac{1}{720}k_6a(0,0)\frac{u^{(6)}_{j,0}}{u_{j,0}}$$

First of all, we observe that $a(0,i) = a(j,k) = 0$ for $i = 0,1,2,3$, $j = 1,2$, $k = 0,1$, so we can immediately conclude that $(f - \phi_0)''$ is $o(1)$. Secondly, we pay extra attention to two of the expressions in the h^4-coefficient, both of which must equal zero:

$$\frac{1}{2}k_2a(2,2) + \frac{1}{24}k_4a(4,0) + a(0,4) = 0, \tag{4.22}$$
$$k_2a(1,2) + \frac{1}{6}k_4a(3,0) = 0. \tag{4.23}$$

Consider next the h^6-coefficient, which last three terms have already been shown to be zero. We then have left three equations in the four unknowns g_5, g_6, t and u;

$$g_6u_{j,0} + g_5u'_{j,0} + tu''_{j,0} + uu^{(3)}_{j,0} = 0, \quad j = 1,2,3 \tag{4.24}$$

where

$$t = \frac{1}{2}k_2a(0,4) + \frac{1}{4}k_4a(2,2) + \frac{1}{48}k_6a(4,0) \tag{4.25}$$
$$u = \frac{1}{6}k_4a(1,2) + \frac{1}{36}k_6a(3,0). \tag{4.26}$$

Writing $t = Au$ for a constant A and solving equations (4.22)-(4.23) and (4.25)-(4.26) with respect to $a(0,4)$, we get

$$a(0,4) = \frac{k_2k_6 - k_4^2}{k_4 - k_2^2}\left(\frac{1}{24}a(4,0) - \frac{1}{18}Aa(3,0)\right),$$

where it is seen from (4.24) that A solves the system of equations $g_6u_{j,0}+g_5u'_{j,0}+Au''_{j,0} = u^{(3)}_{j,0}$ for $j = 1,2,3$. Thus, we arrive at the following expression for asymptotic bias for $p = 3$ parameters;

$$\mathrm{E}\widehat{f}(x) = f(x) - \frac{k_2k_6 - k_4^2}{k_4 - k_2^2}h^4\left[\frac{1}{24}\left\{f^{(4)}(x) - \phi_0^{(4)}(x)\right\}\right.$$
$$\left. - \frac{1}{18}A\left\{f^{(3)}(x) - \phi_0^{(3)}(x)\right\}\right] + o(h^4 + d(n,h)),$$

where $d(n,h)$ denotes the unknown asymptotical convergence rate for the expectation of $\widehat{f}(x)$. For four parameters, the expression above simplify, as we may set $A = 0$. On the

other hand, we must introduce another unknown function, $e(n, h)$, to take the difference $\mathrm{E}\widehat{f}(x) - \phi_0(x)$ into account:

$$\mathrm{E}\widehat{f}(x) = f(x) - \frac{k_2 k_6 - k_4^2}{k_4 - k_2^2} h^4 \frac{1}{24} \left\{ f^{(4)}(x) - \phi_0^{(4)}(x) \right\} + o(h^4 + e(n, h)). \tag{4.27}$$

A pattern emerges for even number of parameters, compare equations (4.20) and (4.27), that resembles the similar behaviour seen for curve estimation using local polynomials, see Section 6.3. Hjort and Jones conjecture that this pattern continues with $o(h^6)$-convergence for five and six parameters and so on, but that is subject to more analysis not pursued here.

Chapter 5

Simulations

In this chapter, the practical implementation of local likelihood for density estimation will be demonstrated. We will also compare the performance of the local maximum likelihood estimator with that of the traditional kernel estimator.

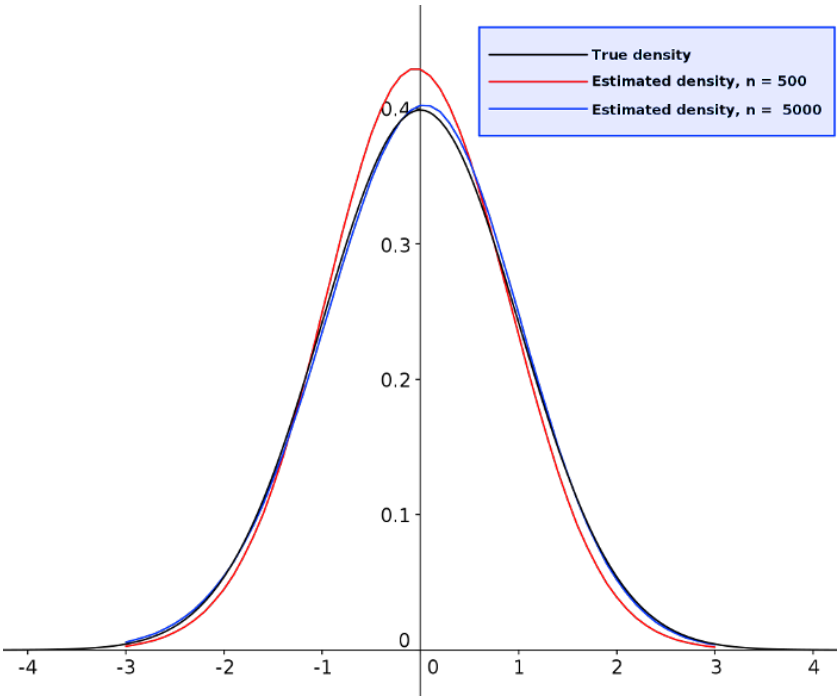
In all of the simulations below, the normal distribution will be used both as kernel $K(\cdot)$ as well as the parametric family $\phi(\cdot;\boldsymbol{\theta})$. Obviously, the subject of estimation is then a two-dimensional parameter; $\boldsymbol{\theta} = (\mu(x), \sigma(x))^T$. Further, all estimations are carried out using the excellent R-implementations developed by K.O. Hufthammer in connection with Hufthammer and Tjøstheim [2008a] and [2008b]. The bandwidths in the following section are chosen so that the estimated mean squared error is minimized (using the unrealistic fact that we know the true density). A general discussion on determination of reasonable bandwidths is included in Section 5.3, and a data driven bandwidth selector is demonstrated in Section 5.4.

5.1 Density estimation

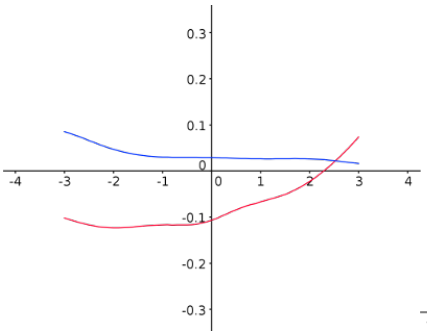
This section treats only local likelihood density estimates, while comparison with the Kernel estimator is treated in Section 5.2.

5.1.1 The normal distribution

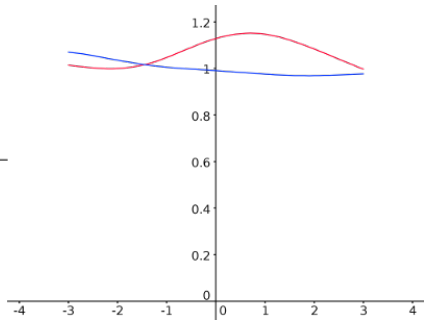
Let us first consider the simple case in which we try to estimate the standard normal density. Of course, the natural choice for the parametric family in this case is the normal distribution, but as we will see in later examples, this works well also for other types of data. In figure 5.1 we see the true, standard normal density $f(x)$, and the estimated density $\phi(x;\widehat{\mu}(x),\widehat{\sigma}(x))$ for $n = 500$ and $n = 5000$.



(a) $\widehat{f}(x) = \phi(x; \widehat{\mu}(x), \widehat{\sigma}(x))$ for two different sample sizes



(b) $\widehat{\mu}(x)$



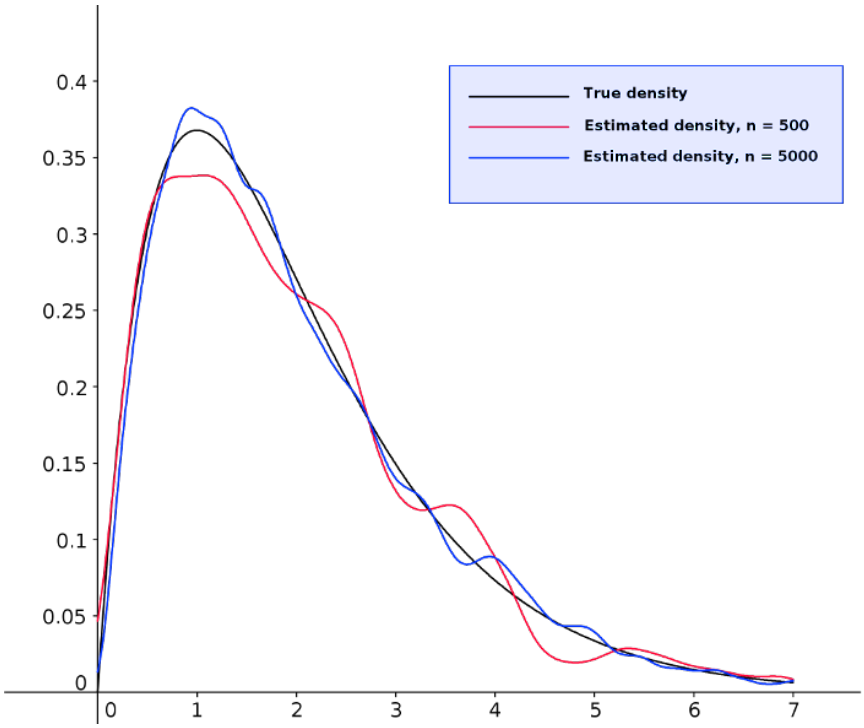
(c) $\widehat{\sigma}(x)$

Figure 5.1: Local maximum likelihood estimate of the standard normal distribution

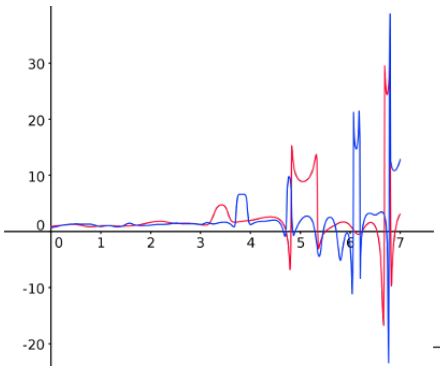
The bandwidth is in all cases chosen to be 1, since choosing a too large bandwidth is hardly a problem when working with the correct parametric family.

We see that in this simple case, the local maximum likelihood estimator performs as expected. With $n = 5000$ observations, the estimate is barely distinguishable from the true density. There is, however, no reason to use this procedure if the form of the underlying distribution is known, and ordinary, global maximum likelihood estimators can be calculated.

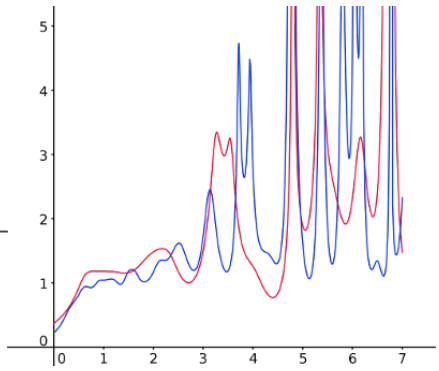
5.1.2 The gamma distribution



(a) $\widehat{f}(x) = \phi(x; \widehat{\mu}(x), \widehat{\sigma}(x))$ for two different sample sizes



(b) $\widehat{\mu}(x)$



(c) $\widehat{\sigma}(x)$

Figure 5.2: Local maximum likelihood estimate of the Gamma($\alpha = 2$, $\beta = 1$)-distribution

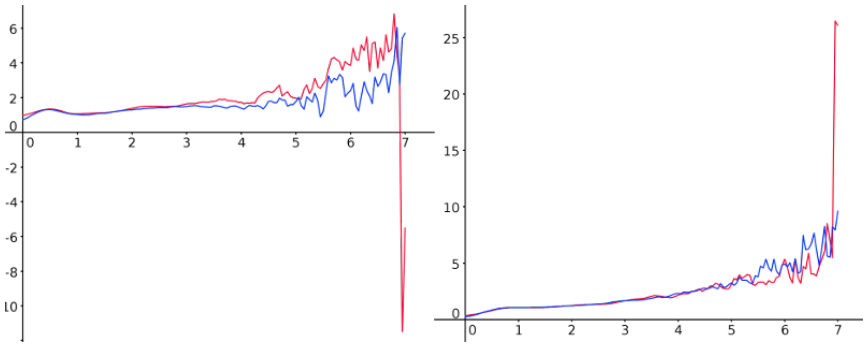
Let us next try to estimate a non-normal density, such as the Gamma(2,1)-distribution, while keeping the Gaussian kernel and parametric family in the local likelihood procedure.

Sums of exponentially distributed variables are gamma distributed, and waiting times are often regarded as such. Several real-world phenomenons can also be modelled using this distribution, such as rainfalls [Aksoy, 2000] and emissions from cars [Zhang et al., 1994]. However, the maximum likelihood estimators for the two parameters, often referred

to as α and β, can not be written out explicitly, so numerical methods must be applied anyhow to estimate the density function. If, in addition, we do not have enough evidence to deem our data gamma distributed, the local maximum likelihood estimator can be a fruitful way of doing inference.

Figure 5.2 summarizes the estimation in the same way as the previous example. The left part of the gamma distribution in question (say for $x < 2.5$) resemble the Gaussian bell-shaped curve, and we recognize this feature in the parameter estimates, which exhibit little variation in this area. The right tail, however, is heavier than its normal counterpart, so we expect $\widehat{\sigma}(x)$ to increase as x increases. The volatility seen in Figure 5.2a and b can be explained intuitively by realizing that estimating a flat portion of a curve locally by a Gaussian curve, can potentially locate $\widehat{\mu}(x)$ at both sides of x, as well as compensating for a distant $\widehat{\mu}$ by a large $\widehat{\sigma}$. Further, we have little data in the tails, which increases both variance and bias.

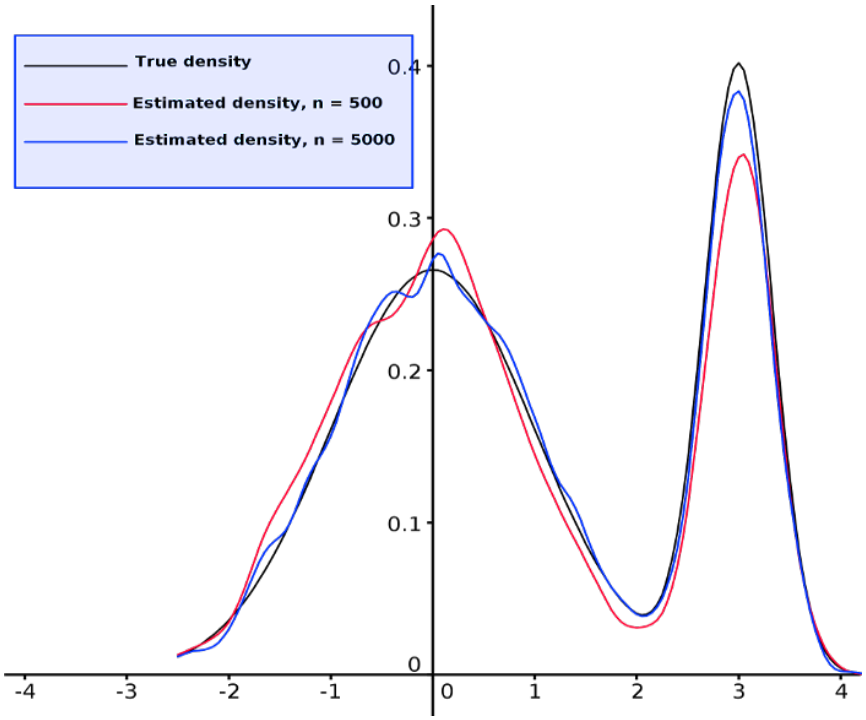
To see that the erratic behaviour of the estimated parameters in figure 5.2 are mostly random fluctuations and not underlying structures in the dataset, we can average the estimates over, say, 100 independent datasets, each with 500 or 5000 observations.



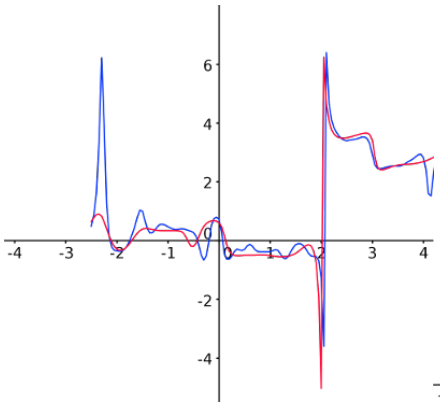
(a) $\overline{\widehat{\mu}(x)}$ (b) $\overline{\widehat{\sigma}(x)}$

Figure 5.3: Averages of parameter estimates over 100 datasets, each with 500 (red) and 5000 (blue) observations.

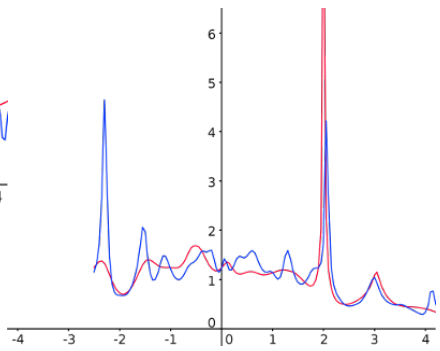
5.1.3 The bimodal normal distribution



(a) $\widehat{f}(x) = \phi(x; \widehat{\mu}(x), \widehat{\sigma}(x))$ for two different sample sizes



(b) $\widehat{\mu}(x)$



(c) $\widehat{\sigma}(x)$

Figure 5.4: Local maximum likelihood estimate of the bimodal normal distribution with parameters $\mu_1 = 0$, $\mu_2 = 3$, $\sigma_1 = 1$, $\sigma_2^2 = 1/3$ and $p = 2/3$

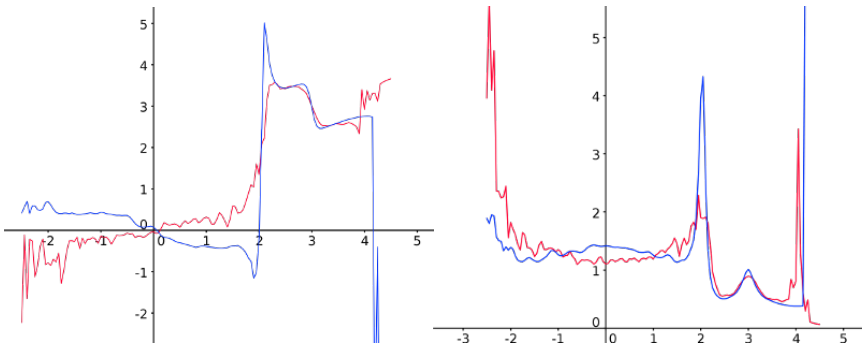
Many natural phenomenons follow bimodal distributions, that is, their probability density functions have two distinct modes. Examples of such are the size of a certain species of ants [Cole and Jones, 1948], and the time between eruptions of the Old Faithful Geyser in Yellowstone National Park, USA [Rinehart, 1969].

A linear combination of two normal distribution will suffice for our purpose of simulations in this section. Let $\phi_{\sigma^2}(x-\mu)$ denote the density function of a normally distributed variable with expectation μ and variance σ^2, then

$$f(x) = p\phi_{\sigma_1^2}(x-\mu_1) + (1-p)\phi_{\sigma_2^2}(x-\mu_2)$$

is the density function of the variable X that is drawn with probability p from a $\mathrm{N}(\mu_1, \sigma_1^2)$-distribution, and with probability $1-p$ from a $\mathrm{N}(\mu_2, \sigma_2^2)$-distribution.

Figure 5.4 shows that the local maximum likelihood estimator performs much as expected in this case. The bi-modality is especially clear in $\widehat{\mu}(x)$, graphed in figure 5.4b, where the two distinct expectations are easily recognized for both sample sizes. This feature is not as clear in $\widehat{\sigma}(x)$, but the density estimates are good nevertheless. The "singularity" seen at approximately $x = 2$ in both $\widehat{\mu}(x)$ and $\widehat{\sigma}(x)$ can be explained using earlier arguments. It coincides with the local minimum located here, and thus with the simplest of classification rules to allocate observations to either of the two normal populations.

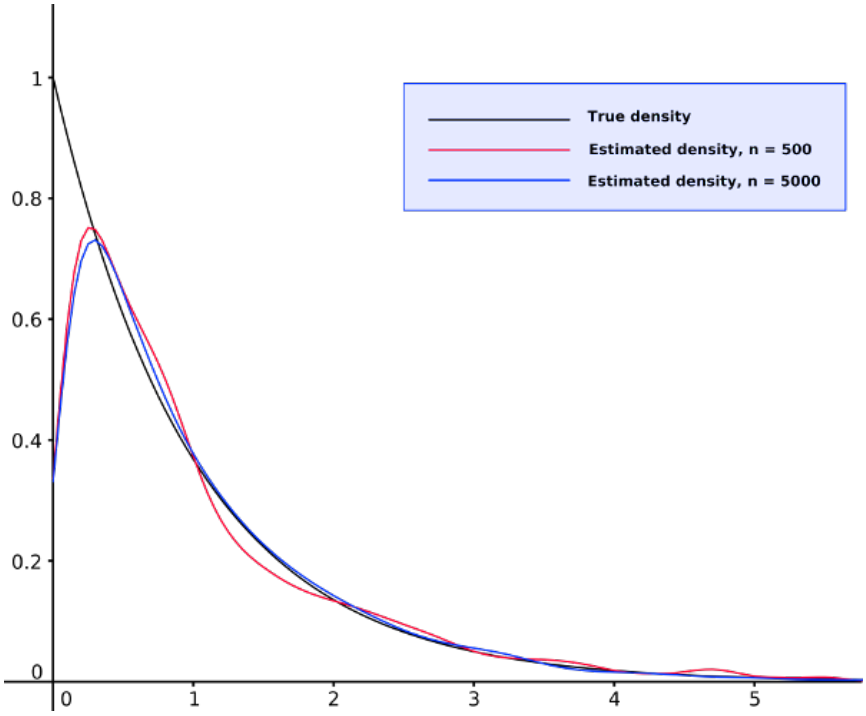


(a) $\overline{\widehat{\mu}(x)}$ (b) $\overline{\widehat{\sigma}(x)}$

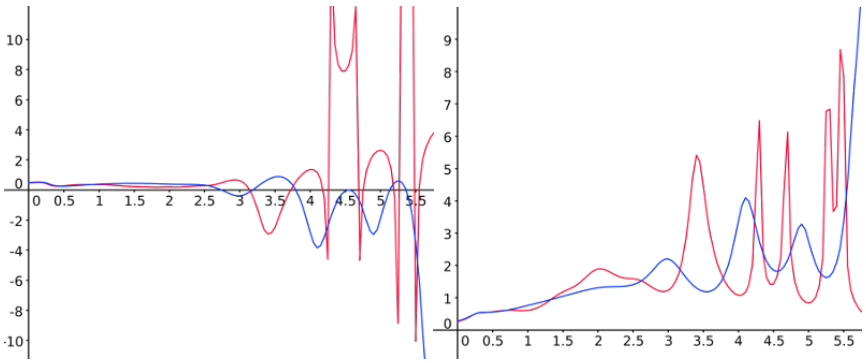
Figure 5.5: Averages of parameter estimates over 100 datasets, each with 500 (red) and 5000 (blue) observations.

Averages over several data sets, seen in Figure 5.5, show the same signs of consistency as for the Gamma(2,1) distribution.

5.1.4 The exponential distribution



(a) $\widehat{f}(x) = \phi(x; \widehat{\mu}(x), \widehat{\sigma}(x))$ for two different sample sizes



(b) $\widehat{\mu}(x)$ (c) $\widehat{\sigma}(x)$

Figure 5.6: Local maximum likelihood estimate of the exponential(1) distribution

The distributions estimated so far in this chapter are well approximated locally by the Gaussian distribution. One possible problem for the Gamma distribution could be the the point $x = 0$, where the density changes from being identically zero to positive values in a non-smooth manner. Even so, since the density approaches zero when $x \to 0$ from the right, this small problem is solved by also letting $\widehat{\sigma}(x) \to 0$. The exponential distribution,

however, with density

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

is not even continuous at zero, so using Normal distributions as building blocks for estimation should be carefully considered in applications.

Figure 5.6 shows that the local maximum likelihood estimator with the Gaussian distribution as parametric family, does not perform very well for small xs when the unknown density is the exponential distribution. This phenomenon persists, even for very large datasets and small bandwidths and is a well known problem also for the kernel estimator. Effective ways to reduce boundary bias have been studied in detail in the literature, and will be discussed in connection with local maximum likelihood in Chapter 6.

5.2 ISE calculations

Let us now investigate how the error of our estimates behave when we vary the bandwidth h, the sample size n, and how they compare with the kernel estimator. We keep the error analysis in this section simple by measuring the distance from the estimated density $\widehat{f}(x)$ to the true density $f(x)$ using the integrated squared error, defined as

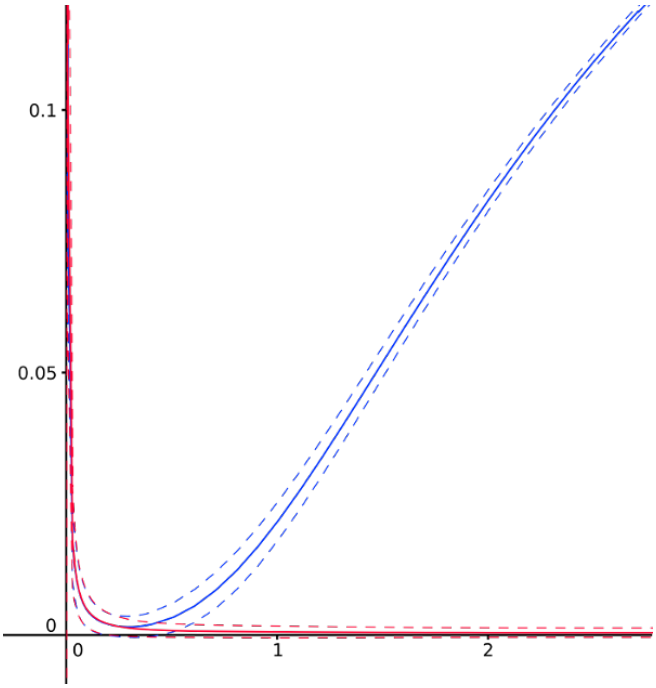
$$\text{ISE}(h) = \int \left(\widehat{f}_h(x) - f(x) \right)^2 \, \mathrm{d}x.$$

We calculate ISE for the distributions considered in the preceding section using the sample sizes $n = 500$ and $n = 5000$ for a set of bandwidths in the interval $[10^{-4}, 3]$, and average over 50 realizations of the data sets. The results are presented in Figures 5.7 - 5.10, where the solid lines represent the calculated ISE as a function of h, and the dashed lines represent approximate 95% confidence bands (ISE $\pm$ two times the empirical standard deviation). Note that, upon averaging over several data sets, the observed ISE will approach its mean, the MISE.

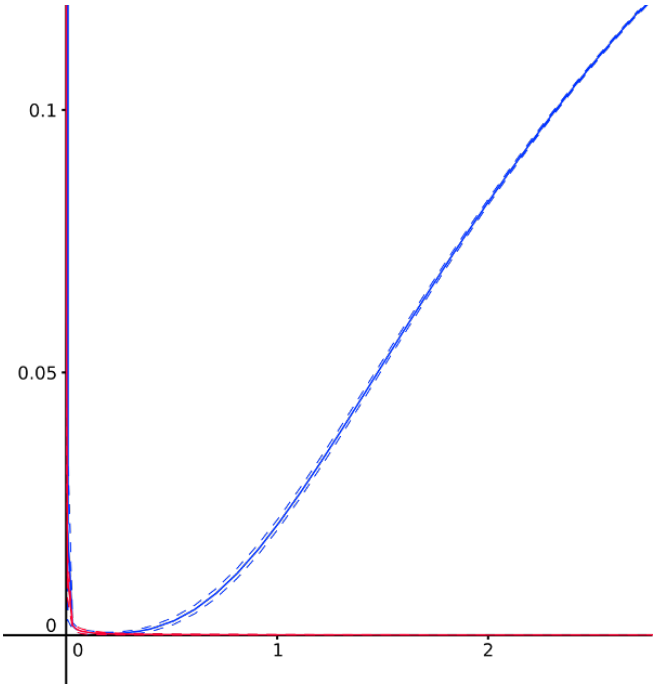
The most important lesson to learn from the following graphs is perhaps that the local maximum likelihood estimator seems to be a *safe choice* compared to the kernel estimator if we know that the unknown distribution is not very far from normal, and we are unsure about which bandwidth to use (in that case choose a moderately large one). Indeed, choosing a too large bandwidth in the local case will be close to a full parametric approach, while doing the same mistake for the kernel estimator will yield a flat, structureless density estimate, and the consequences of this is in terms of ISE are easily seen in the illustrations. In particular, for unimodal data, like the normal- and gamma distribution (Figures 5.7 and 5.8), the error increases much faster for the kernel estimator comared to the local likelihood approach, while for bimodal data (Figure 5.9), a fully parametric Gaussian approach would naturally yield poor results, and thus the local likelihood is more sensitive to large bandwidths.

Local polynomials (discussed in Section 6.3) would probably perform more like local likelihood, as they approach a polynomial as the bandwidth increases.

Of our four examples, the only case in which the kernel estimator is clearly the better choice, is the exponentially distributed data, especially for large n and small h.

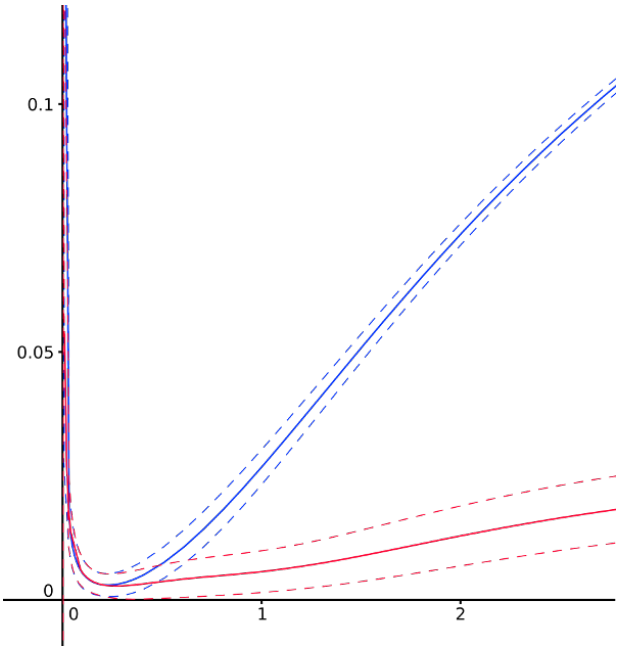


(a) $n = 500$

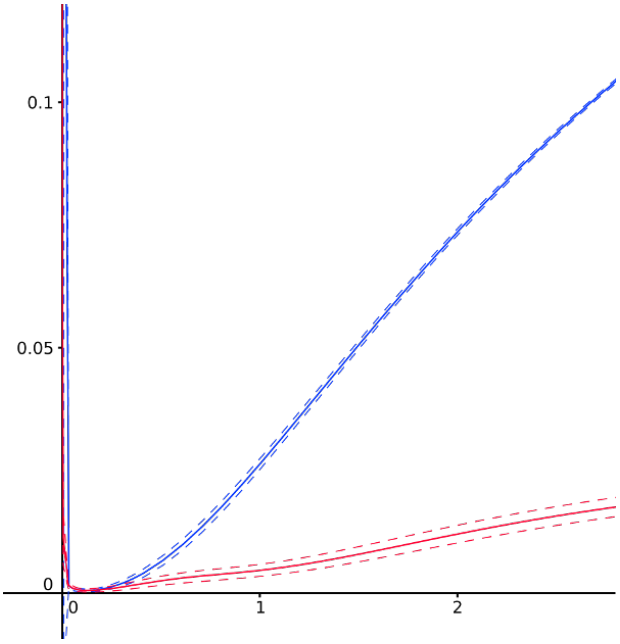


(b) $n = 5000$

Figure 5.7: ISE for estimates of the Standard Normal distribution using the kernel estimator (blue) and local maximum likelihood likelihood estimator (red)

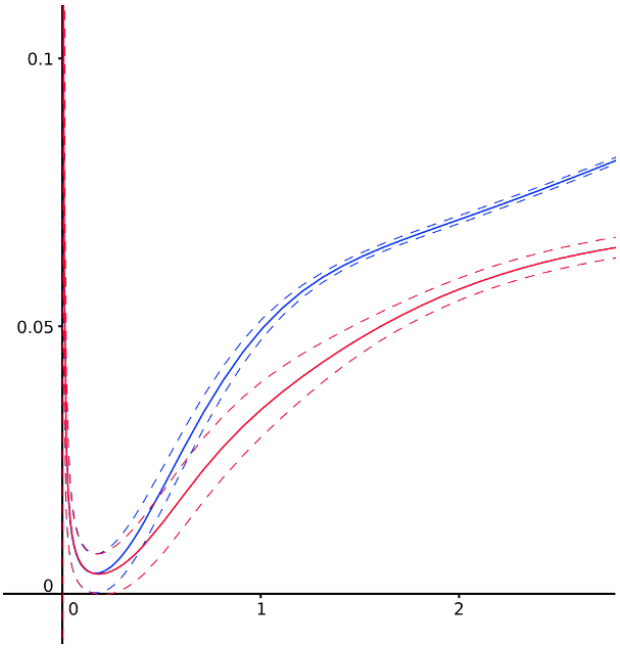


(a) $n = 500$

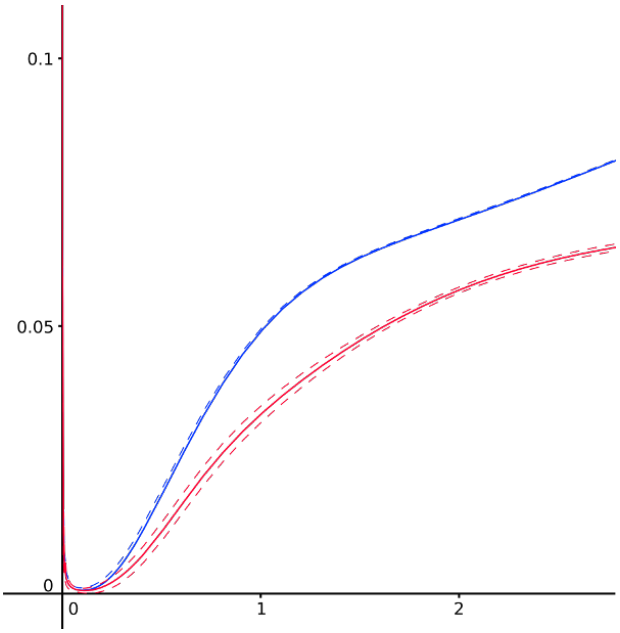


(b) $n = 5000$

Figure 5.8: ISE for estimates of the Gamma(2,1) distribution using the kernel estimator (blue) and local maximum likelihood likelihood estimator (red)

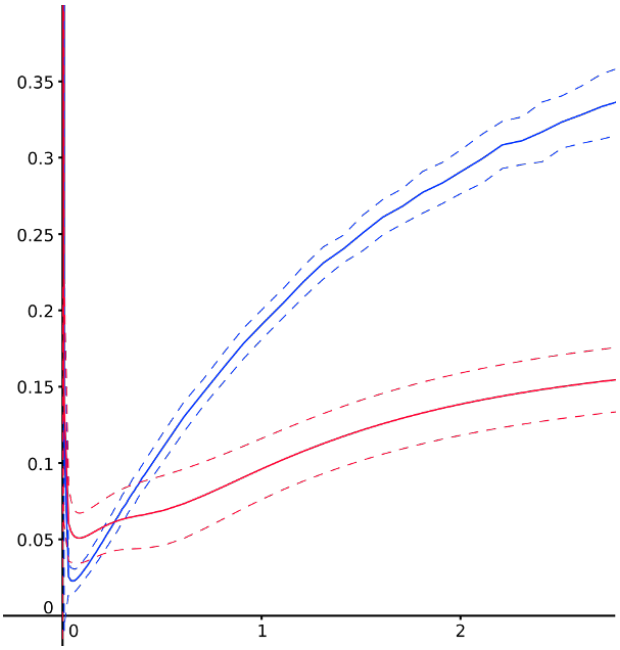


(a) $n = 500$

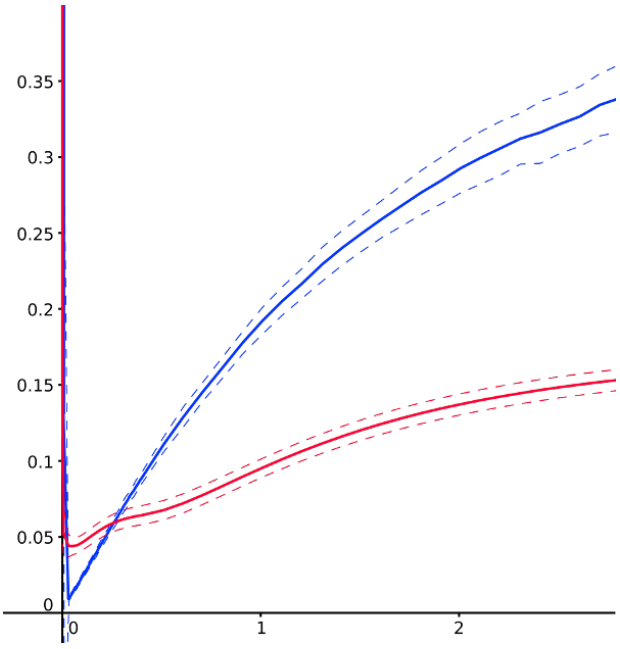


(b) $n = 5000$

Figure 5.9: ISE for estimates of the bimodal Normal Distribution using the kernel estimator (blue) and local maximum likelihood likelihood estimator (red)



(a) $n = 500$



(b) $n = 5000$

Figure 5.10: ISE for estimates of the Exponential(1) distribution using the kernel estimator (blue) and local maximum likelihood likelihood estimator (red)

5.3 Bandwidth selection

The bandwidths applied in Section 5.1 are all chosen to best suit the true density from which the data is sampled. In applications, however, this information is not available, so we need methods for selecting the bandwidth based on the data set at hand. We will therefore review some common methods developed for the traditional kernel estimator, see, e.g., Jones et al. [1996], and then apply one of the to the local likelihood case.

A typical measure of error for the kernel method is the mean integrated squared error (MISE);

$$\mathrm{MISE}(h) = \mathrm{E} \int (\widehat{f}_h - f)^2 \, \mathrm{d}x,$$

which we wish to minimize with respect to h. By letting $n \to \infty$ in the above expression, we obtain the asymptotic mean integrated squared error (AMISE), which can be written as [Wand and Jones, 1995]

$$\mathrm{AMISE}(h) = (nh)^{-1} R(K) + \frac{1}{4} h^4 R\left(f''\right) \left(\int z^2 K(z) \, \mathrm{d}z \right)^2,$$

where $R(g) = \int g^2(x) \, \mathrm{d}x$. A convenient consequence of using AMISE to measure the error, is that it is easily minimized. The optimal bandwidth with respect to AMISE, denoted by h_{AMISE}, is given by

$$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{nR(f'') \left(\int z^2 K \right)^2} \right]^{1/5}. \tag{5.1}$$

The fact that h_{AMISE} is proportional to $n^{-1/5}$ is nice to know in many situations, but we can still not calculate the value exactly since the true density is involved through the expression $R\left(f''\right)$. The latter, however, can be estimated, and this forms the basis for many data driven bandwidth selectors. Jones et al. [1996] discuss several variations of such, as well as other types of algorithms.

One straightforward way to estimate $R\left(f''\right)$ is to replace f by some known distribution such as the normal distribution, but, as Jones et al. [1996] points out, this method tends to oversmooth the data. Another way around the problem of estimating $R\left(f''\right)$ is to replace it by $R\left(\widehat{f}''\right)$, where $\widehat{f}$ is a kernel estimate of f. Again, we arrive at the problem of selecting a suitable bandwidth. Sheather and Jones [1991] suggest then to choose h to be the solution of the fixed-point equation

$$h = \left[\frac{R(K)}{nR\left(\widehat{f}''_{g(h)}\right) \left(\int z^2 K \right)^2} \right]^{1/5}.$$

Note especially that the bandwidth used to estimate the curvature of f is different from the bandwidth used to estimate f itself, a feature that, along with the actual selection of a function g, is discussed by Jones et al. [1996] and Sheather and Jones [1991].

The methods above can in principle be applied to local likelihood estimation directly, but, as [Hjort and Jones, 1996] argues, local likelihood is in many situations *better* than the kernel estimator, and therefore requires smaller bandwidth to obtain minimum error. Specifically, since the variance is the same for both kernel- and local likelihood estimation, and the bias (in the two parameter case) differ only by replacing f'' with $f'' - \phi_0''$, the derivation of the AMISE for local likelihood density estimates goes through in an identical manner, and turns out to be (5.1) but with $R(f'' - \phi_0'')$ in place of $R(f'')$. Further, Hjort and Jones [1996], suggest to use a modified fixed-point method, where h is chosen to be the solution of

$$h = \left[\frac{R(K)}{nR\left(\widehat{f}''_{g(h)} - \phi''_{0_h}\right)\left(\int z^2 K\right)^2} \right]^{1/5}.$$

Most distributions have areas with higher density, and thus more observations, which again call for a smaller bandwidth in order to reveal important structures here. But if that bandwidth is also applied to lower density areas, the estimated density will suffer from undersmoothing as it decreases. On the other hand, if we smooth the tails to a reasonable level, the higher density areas will tend to be oversmoothed. Instead of making a compromise, often resulting in oversmoothed modes and undersmoothed tails, we can let the bandwidth h vary with x. Silverman [1986] discusses an approach for the ordinary kernel estimator, called the nearest neighbour method, where the bandwidth used at X_j is proportional to the distance to the kth nearest observation. Variable bandwidths can greatly improve a density estimate, but the drawback of using such a method is that we, a priori, not only have to chose the proportionality constant, which we by analogy call h, but also k.

5.4 A closer look at the bimodal normal distribution

Figures 5.4 and 5.9 suggest that local maximum likelihood may be suitable to estimate bimodal distributions. In this section we will do some simulations to compare this approach with the kernel estimator in the more realistic situation where we use an automatic bandwidth selector. Silverman [1986] discusses an easily implemented procedure to choose the bandwidth. Start with expression (5.1), but replace $R(f'')$ with $R(\phi'')$, where ϕ is a parametric family, here chosen to be the $N(0, \sigma^2)$-distribution. If we use the Gaussian distribution also for the kernel, some calculations then yield

$$h_{AMISE} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \approx 1.06 \sigma n^{-1/5},$$

where σ must be estimated from the data. It is clear that the integrated curvature $R(f'')$ is greater than $R(\phi'')$ if f is bimodal, so this bandwidth will oversmooth the estimated density if applied unaltered. We make two adjustments to address this problem. First,

choose a more robust estimator for σ than just the empirical standard deviation. The interquartile range (IQR) is less prone to extreme values, so since $IQR = 2\Phi^{-1}(0.75) = 1.34\sigma$ for the $N(0, \sigma)$-distribution, where Φ denotes the cumulative distribution function, we let $\widehat{\sigma}_R = \min(\text{standard deviation}, \text{IQR}/1.34)$ for our purposes. Next, simply reduce the constant 1.06 to reduce smoothing even further. Silverman [1986] suggests to use 0.9 instead for multi-modal distributions, so the bandwidth selector that will be employed in the following experiments becomes

$$h_{AMISE} = 0.9\widehat{\sigma}_R n^{-1/5}.$$

Recall that the bimodal normal density is of the form

$$f(x) = p\phi_{\sigma_1^2}(x - \mu_1) + (1 - p)\phi_{\sigma_2^2}(x - \mu_2),$$

and we will look specifically at six different cases, each with parameters as presented in Table 5.1. The estimates are averaged over 50 datasets consisting of 500 and 5000 observations respectively, and for each estimation, the bandwidth selector discussed in the previous paragraph is used. The bandwidths used for estimation are also averaged over the different realizations, and presented in table 5.2 and 5.3 along with the optimal h that minimizes ISE, using information of the true density.

The automatic bandwidth selector is not very sophisticated, but performs reasonably well. As expected, densities with very large curvatures, like the cases 2, 3 and 6, are oversmoothed because the leading constant is 0.9 regardless of the observations. If we were to know information about the true curvature, for example that there are two sharp peaks far apart, one might consider to reduce this constant even further. Nevertheless, the local maximum likelihood estimator (red curves) is consistently better than the kernel estimator (blue curves) in all cases and for both sample sizes. This is especially true in the difficult areas where f'' is large.

It should be noted here that since the 'unknown' densities here are Gaussian of nature, so using the normal distribution as parametric family will naturally fit very well.

	μ_1	μ_2	σ_1	σ_2	p
Case 1	2	5	1	1	0.5
Case 2	2	7	1	1	0.5
Case 3	2	9	1	1	0.5
Case 4	3.2	5	1	0.5	0.65
Case 5	3.5	5	1	0.5	0.65
Case 6	3.3	5	0.3	0.3	0.65

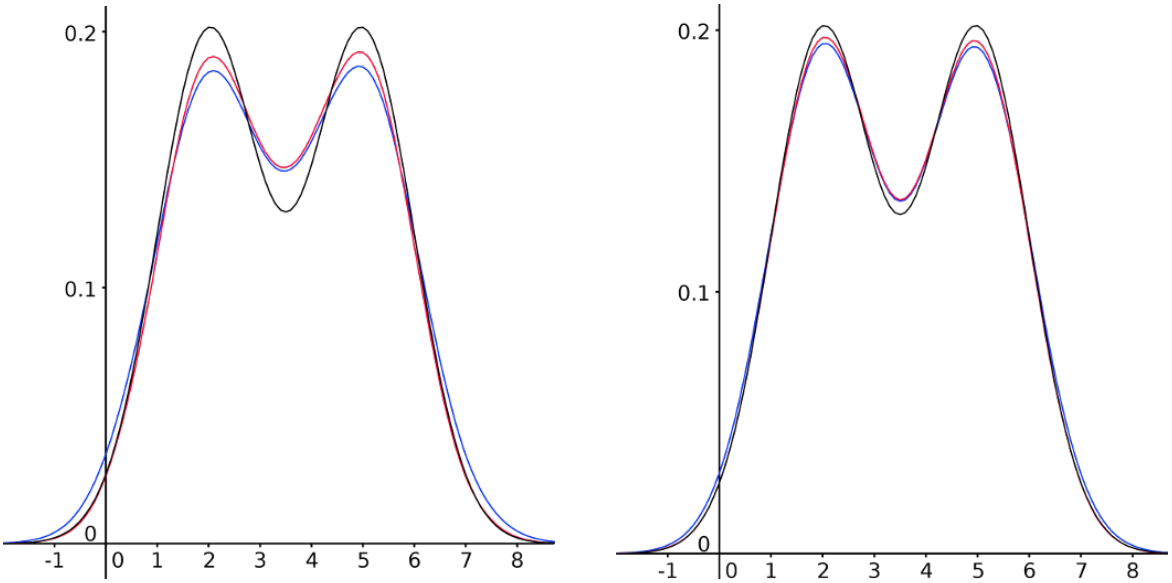
Table 5.1: Six different bimodal normal distributions

	Bandwidth selector	Kernel Estimator			Local Likelihood Estimator		
	h_{AMISE}	h_{ISE}	ISE	$\widehat{\text{ISE}}$	h_{ISE}	ISE	$\widehat{\text{ISE}}$
Case 1	0.47	0.39	0.00152	0.00164	0.41	0.00140	0.00144
Case 2	0.70	0.35	0.00175	0.00475	0.40	0.00167	0.00278
Case 3	0.95	0.37	0.00172	0.00958	0.41	0.00164	0.00380
Case 4	0.31	0.26	0.00231	0.00253	0.29	0.00222	0.00223
Case 5	0.29	0.26	0.00236	0.00243	0.31	0.00222	0.00223
Case 6	0.22	0.11	0.00510	0.01818	0.13	0.00467	0.00730

Table 5.2: Bandwidths calculated for each case, in terms of ISE and AMISE using a sample size of $n = 500$. ISE denotes the observed integrated squares error for each case using the optimal banwidth, h_{ISE} (calculated using the true f), while $\widehat{\text{ISE}}$ denotes the observed integrated squared error using the estimated bandwidth, h_{AMISE}.

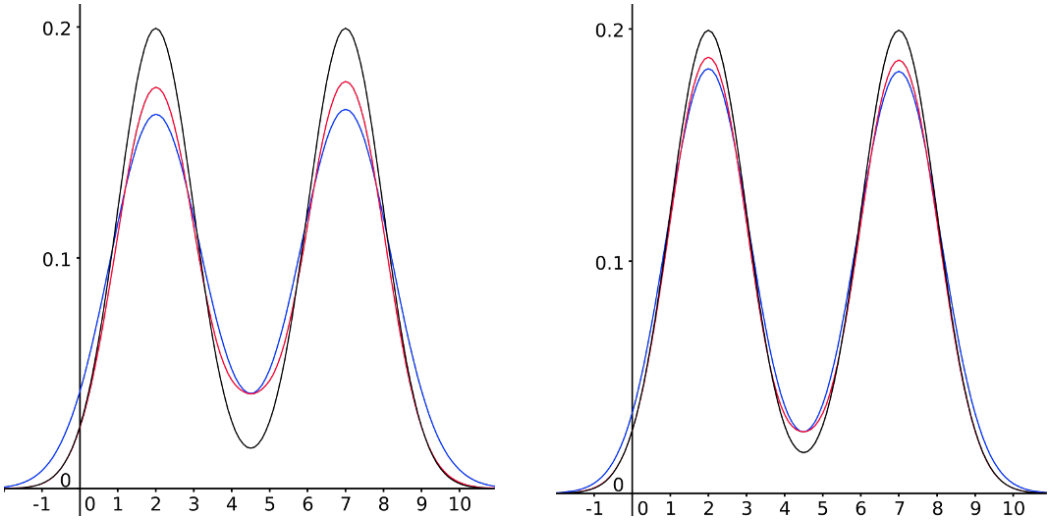
	Bandwidth selector	Kernel Estimator			Local Likelihood Estimator		
	h_{AMISE}	h_{ISE}	ISE	$\widehat{\text{ISE}}$	h_{ISE}	ISE	$\widehat{\text{ISE}}$
Case 1	0.30	0.23	0.00026	0.00030	0.26	0.00025	0.00026
Case 2	0.44	0.22	0.00027	0.00010	0.25	0.00025	0.00054
Case 3	0.60	0.22	0.00027	0.00233	0.26	0.00024	0.00101
Case 4	0.20	0.15	0.00038	0.00046	0.17	0.00037	0.00039
Case 5	0.18	0.15	0.00040	0.00042	0.18	0.00036	0.00036
Case 6	0.14	0.07	0.00097	0.00412	0.08	0.00082	0.00169

Table 5.3: Bandwidths calculated for each case, in terms of ISE and AMISE using a sample size of $n = 5000$. ISE denotes the observed integrated squares error for each case using the optimal bandwidth, h_{ISE} (calculated using the true f), while $\widehat{\text{ISE}}$ denotes the observed integrated squared error using the estimated bandwidth, h_{AMISE}.



(a) $n = 500$ (b) $n = 5000$

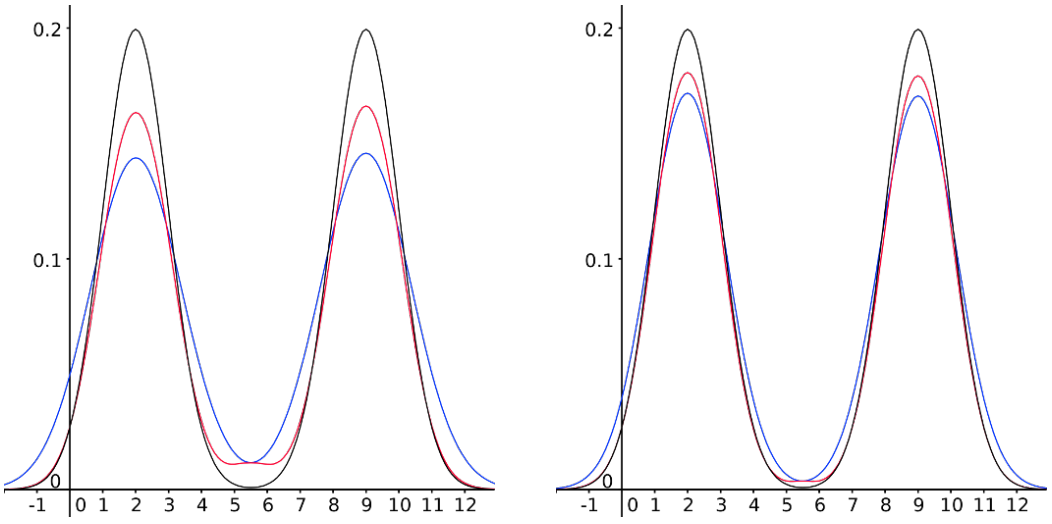
Figure 5.11: Case 1



(a) $n = 500$ (b) $n = 5000$

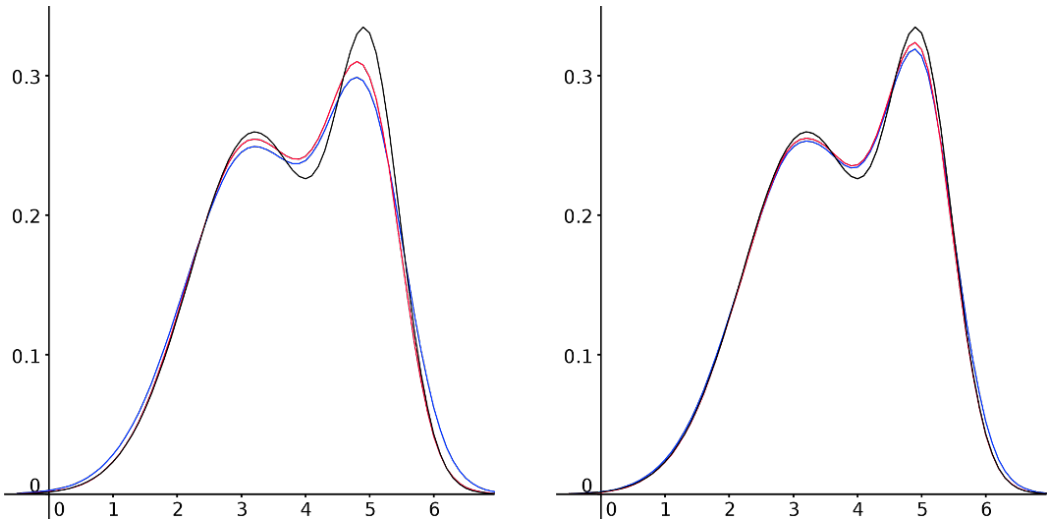
Figure 5.12: Case 2

Red curves: local likelihood, blue curves: the kernel estimator.



(a) $n = 500$ (b) $n = 5000$

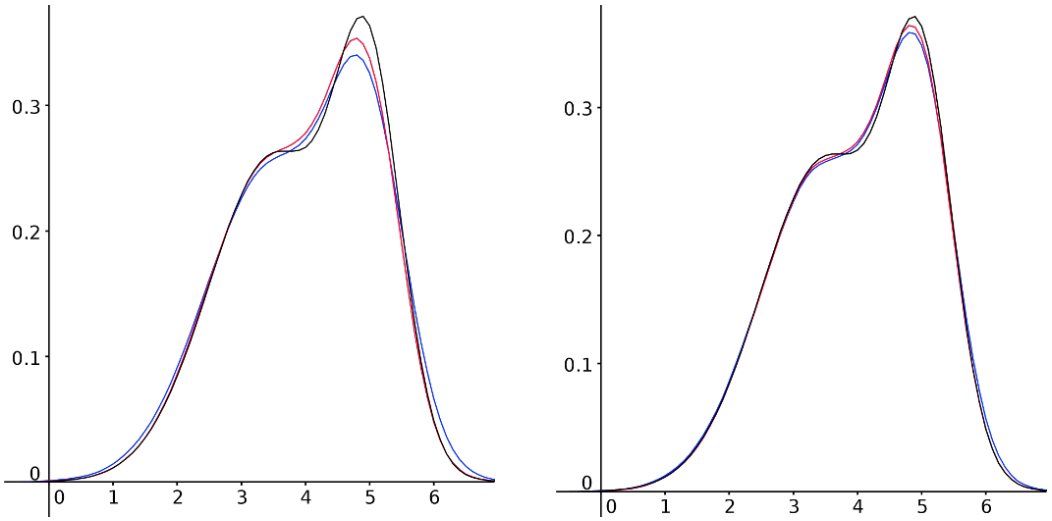
Figure 5.13: Case 3



(a) $n = 500$ (b) $n = 5000$

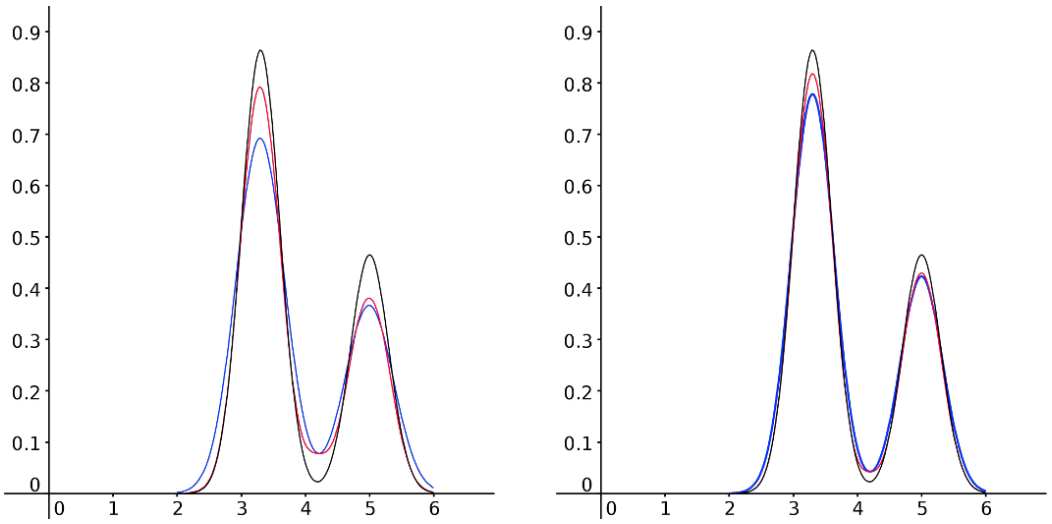
Figure 5.14: Case 4

Red curves: local likelihood, blue curves: the kernel estimator.



(a) $n = 500$ (b) $n = 5000$

Figure 5.15: Case 5



(a) $n = 500$ (b) $n = 5000$

Figure 5.16: Case 6

Red curves: local likelihood, blue curves: the kernel estimator.

Chapter 6

Distributions with bounded support

The local maximum likelihood estimator exhibits the same boundary effects for discontinuous distributions as the kernel estimator, see Figure 5.6. Literature on reduction of these boundary effects exists in abundance in connection with the kernel estimator, and we will review a few of these methods below and demonstrate that they apply also for the local likelihood case.

6.1 The kernel estimator

Suppose first that the unknown density $f(x)$ is positive and smooth on the whole real line. The kernel estimator has then the following well known expression for asymptotic bias:

$$\mathrm{E}\,\widehat{f}(x) = f(x) + \frac{1}{2}h^2\mu_2 f''(x) + O(h^4), \tag{6.1}$$

where $\mu_2 = \int s^2 K(s)\,\mathrm{d}s$ again, and with a similar expressions for the local likelihood [Hjort and Jones, 1996, Section 3]. If the support of f is bounded in either direction, the bias near the boundary is different. Suppose now, for simplicity, and without loss of generality, that the boundary is located at $x = 0$, as is the case for the exponential distribution that we use as example in this chapter. Suppose further that the Kernel K is positive only on $[-1, 1]$, although modifying the arguments below to kernels of infinite support, such as the normal distribution, is possible according to Jones [1993].

For $x < h$, the bias for the kernel estimator is given by [Marron and Ruppert, 1994]:

$$\mathrm{E}\,\widehat{f}(x) \approx a_0(c)f(x) - ha_1(c)f(\prime x) + \frac{1}{2}h^2 a_2(c)f''(x) + o(h^2), \tag{6.2}$$

where $x = ch$, $c \in [0, 1]$ and $a_s(c) = \int_{-1}^{c} u^s K(u)du$. Note that for $c > 1$, this expression reduces to (6.1), so it is really valid for all x. We see that the estimate is not even consistent for $x < h$, but that is easily fixed by dividing our estimate with the inconsistency factor $a_0(c)$. Still, the bias near the boundary is of order $O(h)$, and not $O(h^2)$ as we have in the

interior. Jones [1993] then call for modifications of the kernel function so that $a_0(c) = 1$ and $a_1(c) = 0$. One such modification is given by

$$K_L(x) = \frac{(a_2(c) - a_1(c)x)K(x)}{a_0(c)a_2(c) - a_1^2(c)}, \tag{6.3}$$

for which the desired properties are easy to verify.

Another simple way of reducing bias is by reflecting the data about the boundary and letting the new estimate, $\tilde{f}(x)$, be defined as $\tilde{f}(x) = \widehat{f}(x) + \widehat{f}(-x)$, where $\widehat{f}(\cdot)$ is the traditional kernel estimate based on the reflected data. The resulting estimate is of order $O(h)$, however, as Marron and Ruppert [1994] show:

$$\begin{aligned} \mathrm{E}\tilde{f}(x) = f(x) - 2h\{ca_0(-c) + a_1(-c)\}f'(x) + \frac{h^2}{2}f''(x) \\ + 2h^2\{c^2a_0(-c) + ca_1(-c)\}f''(x) + o(h^2). \end{aligned} \tag{6.4}$$

6.2 Local likelihood

There are two cases worth considering when investigating the boundary bias for the local likelihood estimator; when the parametric family respects the boundaries of the true density, and when that is not the case. The first instance results in consistent estimates, also near the boundary, with convergence rates comparable to those of local polynomial estimation (see Section 6.3). For the second case, the boundary estimates behave like kernel estimates. Details on these properties follow in the following two subsections.

6.2.1 Coinciding support

The choice of parametric family should reflect the knowledge we may have on the true density, also with regard to boundaries. If we know that the true density vanishes for $x < 0$, then the exponential distribution is perhaps better suited for local density estimation than the normal distribution. Assume now that $f(x)$ is zero for $x < 0$, that the parametric family has support $[0, 1]$, and, of course, that the parametric family $\phi(x, \boldsymbol{\theta}(x))$ is actually able to reach the true value $f(x)$ at any given x by varying the parameter. Then the following results, as presented by Hjort and Jones [1996], hold.

Assume again that

$$\int K_h(x - y)u(y, \theta)\{f(y) - \phi(y, \theta)\}\,\mathrm{d}y = 0 \tag{6.5}$$

has a unique solution $\theta(x) = \theta_0(x)$ for all x. In the one-parameter case, we may expand the integral above near zero using the same notation as introduced for the kernel estimator,

resulting in

$$0 = \int_0^\infty K_h(x-y)u_0(y)\{f(y)-\phi_0(y)\}\,\mathrm{d}y$$
$$= a_0(c)u_0(x)\{f(x)-\phi_0(x)\} - ha_1(c)\left[u_0(x)\{f(x)-\phi_0(x)\}\right]'$$
$$+ \frac{1}{2}h^2a_2(c)\left[u_0(x)\{f(x)-\phi_0(c)\}\right]'' + O(h^3), \tag{6.6}$$

where u_0 and ϕ_0 again means the u and ϕ functions with θ_0 inserted as the parameter. Rearranging the terms and replacing $\phi_0(x)$ with $\mathrm{E}\widehat{f}(x)+O((nh)^{-1})$ as before, we arrive at

$$\mathrm{E}\widehat{f}(x) = f(x) - \frac{a_1(c)}{a_0(c)}h(f(x)-\phi_0(x))' + O(h+(nh)^{-1}),$$

where we use the earlier established fact that $f(x)-\phi_0(x) = O(h^2)$ in general. For two locally fitted parameters, however, we regain the appealing $O(h^2)$ bias that we also have in the interior. We now have two equations,

$$0 = \int_0^\infty K_h(x-y)u_{0,j}(y)\{f(y)-\phi_0(y)\}\,\mathrm{d}y,$$

for $j = 1, 2$. We may also write (again from Chapter 4) $(f(x)-\phi_0(x))' = Bh$ and $(f(x)-\phi_0(x))'' = C$ as well as $f(x)-\phi_0 = Ah^2$. Expanding the two integrals aboue to the third power yields a second order term involving $A - \{a_1(c)/a_0(c)\}B + \frac{1}{2}\{a_2(c)/a_0(c)\}C$ and a third order term that contains $-a_1(c)A + a_2(c)B - \frac{1}{6}a_3(c)C$ as a factor. Equating those terms with zero to eliminate B, results in

$$\mathrm{E}\widehat{f}(x) = f(x) + \frac{1}{2}Q(c)h^2\{f(x)-\phi_0(x)\}'' + o(h^2+(nh^3)^{-1}),$$

where

$$Q(c) = \frac{a_2^2(c) - \frac{1}{3}a_1(c)a_3(c)}{a_2(c)a_0(c) - a_1^2(c)}.$$

We will see in Section 6.3 that locally fitted polynomials also have $O(h)$ bias for one parameter, and $O(h^2)$ for two. Hjort and Jones [1996] conjecture that local likelihood fit into this pattern as the number of parameters increases, but that is yet to be proved.

6.2.2 Non-coinciding support

The expressions (6.2) and (6.4) have equivalents in the local likelihood case when the parametric family's support exceeds the limits of the true density. The bias reduction methods proposed by Jones [1993] can therefore be applied. These expressions follow again from expanding integrals like the right hand side of (6.5) in powers of h, but we must

take extra care of the integration limits in order to include all the estimated density. We assume that the kernel has support $[-1, 1]$, and find that:

$$
\begin{aligned}
0 &= \int K_h(x-y)u_0(y)\{f(y)-\phi_0(y)\}\mathrm{d}y \\
&= \frac{1}{h}\int_0^{x+h} K\left(\frac{x-y}{h}\right)u_0(y)\{f(y)-\phi_0(y)\}\,\mathrm{d}y - \frac{1}{h}\int_{x-h}^{0} K\left(\frac{x-y}{h}\right)u_0(y)\phi_0(y)\,\mathrm{d}y \\
&= \frac{1}{h}\int_0^{h(c+1)} K\left(c-\frac{y}{h}\right)u_0(y)\{f(y)-\phi_0(y)\}\,\mathrm{d}y - \frac{1}{h}\int_{h(c-1)}^{0} K\left(c-\frac{y}{h}\right)u_0(y)\phi_0(y)\,\mathrm{d}y \\
&= \int_{-1}^{c} K(u)u_0(x-uh)\{f(x-uh)-\phi_0(x-uh)\}\,\mathrm{d}u - \int_c^1 k(u)u_0(x-uh)\phi_0(x-uh)\,\mathrm{d}u \\
&= a_0(c)u_0(x)\{f(x)-\phi_0(x)\} - a_1(c)h[u_0(x)\{f(x)-\phi_0(x)\}]' \\
&\qquad + \frac{1}{2}a_2(c)h^2[u_0(x)\{f(x)-\phi_0(x)\}]'' - b_0(c)u_0(x)\phi_0(x) + b_1(c)h[u_0(x)\phi_0(x)]' \\
&\qquad\qquad + \frac{1}{2}b_2(c)h^2[u_0(x)\phi_0(x)]'', \qquad (6.7)
\end{aligned}
$$

where $b_s(c) = \int_c^1 u^s K(u)\,\mathrm{d}u$. The argument above is very much like that of Marron and Ruppert [1994], with the necessary modifications. The extra terms resulting from the wider integration limits ruin our hope for consistency and $O(h^2)$-convergence no matter how many parameters we include. Using that $a_0(c) + b_0(c) = 1$ and $a_1(c) + b_1(c) = 0$, the boundary bias then becomes

$$
\begin{aligned}
\mathrm{E}\widehat{f}(x) = a_0(c)f(x) - ha_1(c)&\frac{(f(x)u_0(x))'}{u_0(x)} \\
&\frac{h^2}{2}\left(\frac{a_2(c)(f(x)u_0(x))'' - a_2(1)(u_0(x)\phi_0(x))''}{u_0(x)}\right) + O(h^2 + (nh)^{-1}).
\end{aligned}
$$

We see that the coefficients causing inconsistency and $O(h)$-bias are the same as for the kernel estimator, so we can apply the same alternative kernel (6.3) in order to achieve $O(h^2)$-bias.

As for the reflection method, similar calculations confirm that the boundary bias is of order $O(h)$ also for the local maximum likelihood estimator. Figure 6.1 displays the reflection method in practice when we estimate the Exponential(1)-distribution using the Gaussian parametric family. The estimates are averaged over 50 realizations, each with 500 observations. There is still some $O(h)$ bias near zero, but this simple method at least ensures consistency for all x, which is the case for the Kernel estimator as well.

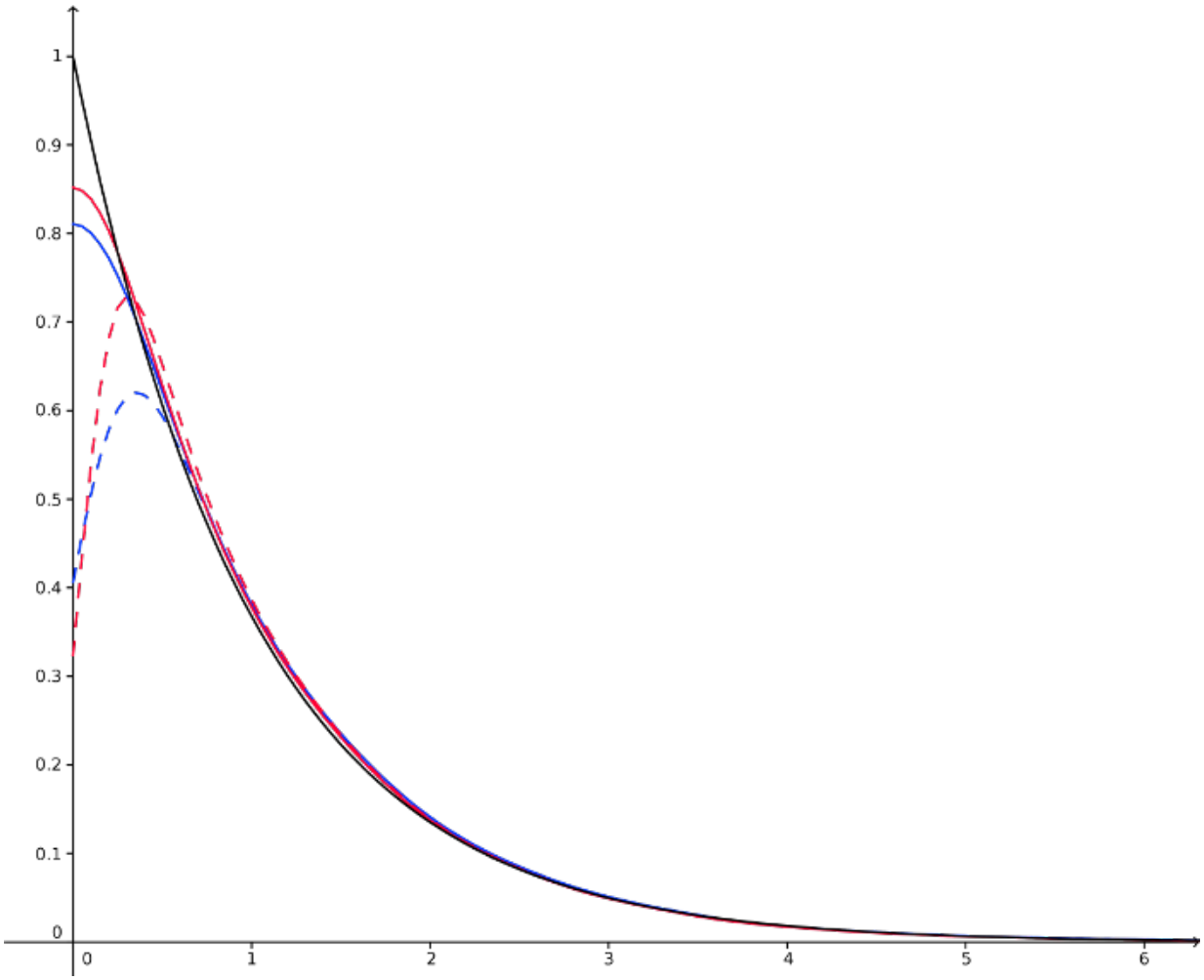


Figure 6.1: The dashed lines show unaltered density estimates for the exponential distribution (solid black line), using local likelihood (red) and the kernel method (blue). The solid red and blue lines show corresponding estimates when using the reflection method. The parametric family is here the normal distribution.

6.3 The local polynomial connection

Recall the example from Chapter 3 where the least squares regression line was estimated locally rather than globally. This kind of non-parametric approach makes us able to estimate any smooth relationship between explanatory- and response variables. We can generalize the method to include polynomials of higher degree, and more efficient kernels than the uniform kernel we applied in Chapter 3. To see how local polynomials compare with the local maximum likelihood method when applied to density estimation, we follow the treatment of local polynomial estimation by and Fan and Gijbels [1996].

Suppose we observe n pairs of observations, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, and that

there exists an unknown relationship between the two variables on the form

$$\mathrm{E}(Y|X = x) = m(X).$$

If the pth derivative of $m(x)$ exists at x_0, then the following approximation is valid in a neighbourhood of x_0:

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{1}{2!}m''(x_0)(x - x_0)^2 + \cdots + \frac{1}{p!}m^{(p)}(x_0)(x - x_0)^p.$$

The unknown coefficients $\beta_k = m^{(k)}(x_0)/k!$ may be estimated by minimizing

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j (X_i - x_0)^j \right)^2 K_h(X_i - x_0), \tag{6.8}$$

where, following convention, K is a unimodal and symmetric kernel damping the influence of observations far from x_0, and $K_h(x) = h^{-1}K(x/h)$. The estimates are valid only in the vicinity of x_0, so estimates must be calculated for each point of a reasonably chosen set of xs. It follows immediately that $\widehat{m}(x_0) = \widehat{\beta}_0(x_0)$, and equivalently for higher order derivatives; $\widehat{m^{(k)}}(x_0) = k!\widehat{\beta}_k(x_0)$. Minimizing (6.8) is a weighted least squares problem; let

$$\boldsymbol{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix}$$

be the design matrix, and collect response variables and estimates in the following vectors:

$$\boldsymbol{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix},$$

and let $\boldsymbol{W} = \mathrm{diag}(K_h(X_i - x_0))$. The estimates are then given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}.$$

6.3.1 Density estimation

Density estimation can be formulated as a regression problem, as Fan and Gijbels [1996] explain. Suppose we have n observations $X_1, X_2, \ldots, X_n$ from an unknown probability density function, $f(x)$, that we wish to estimate on the interval $[a, b]$. Let $\{I_k, k = 1, \ldots, n\}$ be a partition of $[a, b]$ consisting of N intervals of equal size δ, and denote x_k as the center of I_k. Denote further by y_k the proportion of observations that are covered by I_k divided by δ. One can easily show that $\mathrm{E}(y_k) \approx f(x_k)$ as $N \to \infty$ and $n \to \infty$. Therefore, we

may estimate f by local polynomials using the weighted least squares method from the previous section. Expression (6.8) becomes

$$\sum_{k=1}^{n}\left(y_k - \sum_{j=1}^{p}\beta_j(x_k - x)^j\right)^2 K_h(x_k - x) \tag{6.9}$$

for each x, and the resulting density estimate is $\widehat{f}(x) = \widehat{\beta}_0(x)$, with estimators for derivatives given by $\widehat{f^{(k)}}(x) = k!\widehat{\beta}_k(x)$.

Recall from Section 4.5.3 that for local maximum likelihood, the difference between f and f_0 is of order h^2 in the one- and two-parameter cases, h^4 for three or four parameters, and that Hjort and Jones [1996] conjecture that this pattern continues with h^6-convergence for five and six parameters and so on. Hjort and Jones [1996] demonstrate that boundary bias also follow this pattern with a wisely chosen parametric family. This is analogous with local polynomials, as summarized in a general theorem by Fan and Gijbels [1996] (Theorem 3.1). For a local linear fit, we estimate two parameters locally, namely the intercept and the slope. In that case, the bias is of order h^2. Second and third order polynomials mean three and four parameters, and thus h^4-bias. This pattern holds for all orders, and similar results are derived for estimates of derivatives.

Fan and Gijbels [1996] show that polynomials of *odd* orders are preferable to those of even orders, so in practice, the local linear fit seems like a logical choice.

6.3.2 Automatic boundary correction

A striking feature of curve estimation by local polynomials is that it does not suffer from the boundary effects that occur when we work with the kernel estimator or local maximum likelihood, a fact also demonstrated by Fan and Gijbels [1996]. We adjusted for this quite easily by reflecting the data or by kernel adjustments, but we still had to know the location of the boundary point. *This is not required by local polynomials.* In fact, the local polynomial estimator automatically adjusts for boundary effects automatically by its construction. To see this, we introduce some notation.

Let e_ν be the unit vector, with one in the νth position and zero elsewhere. Further, let

$$S_{n,j} = \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^j,$$

so that $\boldsymbol{S}_n \stackrel{def}{=} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ is the $(p+1)\times(p+1)$ matrix $\{S_{n,j+l}\}_{0\le j,l\le p}$. From now on, we focus on the density curve itself, but identical derivations may be carried out for its derivatives up to order p. The estimate can now be written as

$$\begin{aligned}\widehat{\beta}_0 = e_1^T\widehat{\boldsymbol{\beta}} &= e_1^T S_n^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} \\ &= \sum_{i=1}^{n} W_n\left(\frac{X_i - x_0}{h}\right)Y_i,\end{aligned}$$

where $W_n(x) = e_1^T \boldsymbol{S}_n^{-1}\{1, xh, \cdots, (xh)^p\}^T K(t)/h$, and p is the order of the approximating polynomial. We see now that local polynomials very much resemble ordinary kernel estimation, but the kernel W_n in this case changes its shape with location along the x-axis. We will see below that for points near the boundary, this 'kernel' is exactly on the form (6.3) for a local linear fit, that is, boundary bias will be taken care of automatically.

Fan and Gijbels [1996] proceed by showing that

$$S_{n,j} = nh^j f(x_0)\mu_j\{1 + o_P(1)\} \tag{6.10}$$

where $o_P(1)$ is a quantity that tends to zero in probability as $h \to 0$ and $nh \to \infty$, and $\mu_j = \int u^j K(u)\,\mathrm{d}u$ as before. Let now $\boldsymbol{S} = \{\mu_{j+l}\}_{0\le j,l\le p}$, so that we can write

$$\boldsymbol{S}_n = nf(x_0)\boldsymbol{H}\boldsymbol{S}\boldsymbol{H}\{1 + o_P(1)\},$$

where $\boldsymbol{H} = \mathrm{diag}(1, h, \ldots, h^p)$. Substituting this back into the definition of W_n, we get

$$W_n(x) = \frac{1}{nhf(x_0)} e_1^T \boldsymbol{S}^{-1}(1, x, \ldots, x^p)^T K(x)\{1 + o_P(1)\},$$

which in turn yields

$$\widehat{f}(x_0) = \widehat{\beta}_0 = \frac{1}{nhf(x_0)} \sum_{i=1}^{n} K^* \left(\frac{X_i - x_0}{h}\right) Y_i\{1 + o_P(1)\},$$

where

$$K^*(x) = e_1^T \boldsymbol{S}^{-1}(1, x, \ldots, x^p)^T K(t).$$

Fan and Gijbels [1996] refer to $K^*(t)$ as the *equivalent kernel*.

Consider now a boundary point, $x = ch$. The quantity $S_{n,j}$ is again given by (6.10), but with $a_j(c)$ in place of μ_j. The equivalent kernel then turn out to be

$$K_c^*(x) = e_1^T \boldsymbol{S}_c^{-1}(1, x, \ldots, x^p)^T K(x),$$

where $\boldsymbol{S}_c = \{a_{j+l}(c)\}_{0\le j,l\le p}$. When we write out the equivalent boundary kernel in the linear case, we finally see that

$$K_c^*(x) = \frac{(a_2(c) - a_1(c)x)\,K(x)}{a_0(c)a_2(c) - a_1^2(c)},$$

which is exactly the kernel we introduced for the kernel estimator in (6.3) to eliminate boundary bias. Fan and Gijbels [1996] show that similar bias-eliminating equivalent kernels arise for all order of derivatives up to an arbitrary order p.

To sum up, boundary estimates generated by the **Kernel estimator** with a kernel K that is symmetric and has support $[-1, 1]$, are generally not consistent for distributions with bounded support. This is easily fixed, though, by employing alternative kernels as described by Jones [1993].

The **Local Likelihood Estimator** yields consistent boundary estimates, *if* the parametric family respects the boundary a shown in Section 6.2.2. Otherwise, the methods by Jones [1993] will work also here.

Compact supports of the kernels give simple calculations, but according to Jones [1993], this assumption can be relaxed. Especially distributions with exponential decay, such as the Gaussian distribution, should not have too much influence on the results. Note further that both these methods require knowledge on the location of the boundary, and also *active implementation* by the researcher, either to modify the kernel, or to choose a suitable parametric family.

The **Local Polynomial** approach takes care of boundary estimation automatically, as demonstrated above. This is, of course, much easier from a programmer's point of view, and is a good alternative when the estimated curve is our ultimate goal.

Chapter 7

Local partial likelihood in the Cox regression model

The previous chapters have dealt manily with probability density estimation. We will now turn our attention to another application, in which a local likelihood approach can be useful in practice.

Recall the Cox regression model introduced in Section 2.1. We wish to estimate the failure rate of an item (or individual) as a function of time, t, based on a set of observed covariates, $\boldsymbol{x}$. The model in its most general form, is given by [Fan et al., 1997]

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\Psi(\boldsymbol{x}, t), \tag{7.1}$$

where λ is the hazard rate of interest, λ_0 is the *baseline* hazard, and Ψ is the effect on the baseline hazard resulting from covariates. By assuming $\Psi(0) = 1$, the hazard rate is just the baseline hazard when all covariates are zero. Consequently, a common reformulation of (7.1) is

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\left(\psi(\boldsymbol{x}, t)\right).$$

We take censoring into account as follows. Suppose for individual number i that we observe either the failure time t_i or the censoring time c_i, as well as the corresponding set of covariates $\boldsymbol{x}_i$ as a realization from the bivariate variable $\{T_i, \boldsymbol{X}_i\}$ or $\{C_i, \boldsymbol{X}_i\}$. The censoring mechanism is assumed independent from failure times. Denote $Z_i = \min(T_i, C_i)$, and the indicator variable δ_i, being one if the observed event-time is uncensored, and zero otherwise. Thus, we observe the triples

$$\left\{(\boldsymbol{X}_i, Z_i, \delta_i), i = 1, \ldots, n\right\},$$

which are iid samples from the population

$$\left(\boldsymbol{X}, \min(T, C), I_{\{T \leq C\}}\right).$$

We will now consider three different levels of parametrizations of the model (7.1), in which ordinary likelihood, local likelihood and local partial likelihood will be employed correspondingly, all of which are studied in detail by Fan et al. [1997]. For the time being, assume that all covariates are independent of time.

7.1 Parametric baseline, parametric covariate effects: ordinary likelihood

Suppose the baseline hazard is parametrized by $\lambda_0(t) = \lambda_0(t;\boldsymbol{\theta})$ and that ψ is parametrized by $\psi(x) = \psi(x;\boldsymbol{\beta})$, for example a linear function with $\boldsymbol{\beta}$ as vector of coefficients as mentioned in Chapter 2. According to Fan et al. [1997], the likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ conditioned on $\boldsymbol{X}$ is given by

$$L = \prod_u f(Z_i|X_i) \prod_c S(Z_i|X_i),$$

where $f(t|x)$ is the conditional density function of T given X, $S(t|x) = P(T > t|X = x)$ is the conditional survival function, and $\prod_u$ and $\prod_c$ denote multiplication over uncensored and censored individuals respectively. It is further shown that under the model (7.1), the log-likelihood function becomes

$$\log\ L = \sum_{i=1}^{n} \Big[\delta_i \big\{ \log\lambda_0(Z_i;\boldsymbol{\theta}) + \psi(X_i;\beta) \big\} - \Lambda_0(Z_i;\boldsymbol{\theta}) \exp\big\{\psi(X_i;\beta)\big\} \Big], \tag{7.2}$$

where $\Lambda_0 = \int_0^t \lambda_0 \, \mathrm{d}t$ is the cumulative hazard function. Note that (7.2) is the logarithm of (2.7) with censoring taken into account. Maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ follow upon maximization of (7.2).

7.2 Parametric baseline: local likelihood

Next, keep the parametric assumption for the baseline hazard, but assume that $\psi(x)$ is not specified by any parametric model, only that it is smooth enough for a Taylor expansion. In the case of only one covariate, we have

$$\psi(X) \sim \psi(x) + \psi'(x)(X - x) + \cdots + \frac{\psi^{(p)}(x)}{p!}(X - x)^p \tag{7.3}$$

in a neighbourhood of X. We can then construct the vectors

$$\boldsymbol{X} = \{1, X - x, \ldots, (X - x)^p\}^T \quad \text{and} \quad \boldsymbol{X_i} = \{1, X_i - x, \ldots, (X_i - x)^p\}^T,$$

so that in a neighbourhood of x, ψ can be modelled as

$$\psi(X) \sim \boldsymbol{X}^T\boldsymbol{\beta}, \tag{7.4}$$

where $\boldsymbol{\beta} = \{\psi(x), \psi'(x), \ldots, \psi^{(p)}(x)/p!\}$. By introducing a kernel function $K_h = h^{-1}K$ such as the normal distribution, a localized version of (7.2) becomes

$$l_{loc} = \sum_{i=1}^{n} \Big[\delta_i \big\{ \log\lambda_0(Z_i;\boldsymbol{\theta}) + \boldsymbol{X}_i^T\boldsymbol{\beta} \big\} - \Lambda_0(Z_i;\boldsymbol{\theta}) \exp\big\{\boldsymbol{X}_i^T\boldsymbol{\beta}\big\} \Big] K_h(X_i - x). \tag{7.5}$$

Again, maximization yield estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, but they are now dependant on x.

7.3 Non-parametric baseline: local partial likelihood

Lastly, we discard the parametric model for the baseline hazard, and follow Fan et al. [1997] in the derivation of a local partial likelihood function. Let $t_1 < \cdots < t_N$ be the ordered failure times, and let (j) denote the item failing at time t_j. The cumulative baseline hazard function, Λ_0, is non-decreasing by definition, so the weakest assumption we are in the position to do, is that Λ_0 has a jump θ_j at t_j. Thus, $\Lambda_0 = \sum_{j=1}^{N} \theta_j I\{t_j \leq t\}$, so

$$\Lambda_0(Z_i;\boldsymbol{\theta}) = \sum_{i \in R(t_j)} \theta_j,$$

where $R(t_j)$ denotes the risk set at time t_j again. Substituting this into (7.2), the log-likelihood becomes

$$\log\ L = \sum_{j=1}^{J} \Big[\big\{ \log\theta_j + \psi(X_i;\beta) \big\} \Big] - \sum_{i=1}^{n} \Big[\sum_{i \in R(t_j)} \theta_j \exp\{X_i;\beta\} \Big], \tag{7.6}$$

where the δ_i's have become superfluous, because we only sum over the non-censored individuals in the first term. According to Breslow [1972], the maximizer of $\log L$ with respect to θ_j is

$$\widehat{\theta}_j = \Big[\sum_{i \in R(t_j)} \exp\{\psi(X_i;\boldsymbol{\beta})\} \Big]^{-1},$$

so that the likelihood function with respect to $\boldsymbol{\beta}$ turns out to be

$$\max_{\lambda_0} \log\ L = \sum_{j=1}^{N} \Big[\psi(X_{(j)};\boldsymbol{\beta}) - \log \Big\{ \sum_{i \in R(t_j)} \exp\big(\psi(X_i;\boldsymbol{\beta})\big) \Big\} \Big] - N, \tag{7.7}$$

which upon maximization, is equivalent to the partial likelihood function for the Cox regression model used in Chapter 2. By analogy with the preceding sub-section, the *local*

partial likelihood function is then

$$l_{ploc} = \sum_{j=1}^{N} K_h\left(X_{(j)} - x\right)\left[\boldsymbol{X}_{(j)}^T\boldsymbol{\beta} - \log\left\{\sum_{i\in R(t_j)} \exp\left(\boldsymbol{X}_{(j)}^T\boldsymbol{\beta}\right)\right\}K_h\left(X_i - x\right)\right]. \tag{7.8}$$

Apart from just being a localized version of (7.7), according to Fan et al. [1997], it can also be derived more formally from the local likelihood function (7.2).

This is not the place for technical proofs on properties of the local partial likelihood estimators. Suffice it to say, theorems concerning consistency, asymptotic normality, existence and uniqueness under suitable regularity conditions are all covered by Fan et al. [1997]. An important, and perhaps also surprising, point made in this reference, is that the asymptotic bias and variance are the same in the two preceding situations. We do not gain any more information by selecting a parametric model for the baseline hazard (asymptotically, that is), so the local partial likelihood is indeed preferable, as we avoid parametric misspecification without losing performance.

7.3.1 Other variations

Let us briefly consider a few variations of de model discussed above. Obviously, we need to be able to derive a version of (7.5) for situations with more than one covariate. Although demanding more involved notation, the Taylor expansion (7.3) is straightforward to perform with respect to more than one covariate. The vectors $\boldsymbol{X}$ and $\boldsymbol{X}_i$ then becomes matrices, and we have to let go of the compact notation (7.4), especially for more than two covariates.

A second natural extension of the theory, as was mentioned in Chapter 2, is the introduction of time-dependant covariates. Fan et al. [2006] discusses this option, using the following model as a starting point. Let

$$\lambda(t|w,\boldsymbol{z}) = \lambda_0(t)\exp\Big(\boldsymbol{\beta}(W(t))^T\boldsymbol{Z}(t) + g(W(t))\Big),$$

where $\boldsymbol{\beta}(\cdot)$ and $g(\cdot)$ are unknown coefficient functions depending on the value of the exposure variable $W(t)$, which usually represents just time when dealing with individuals, or perhaps mechanical stress when we study the lifetime of certain devices. $\boldsymbol{Z}(t)$ denotes the vector of time-dependant covariates. The partial likelihood is

$$L(\boldsymbol{\beta}(\cdot), g(\cdot)) = \prod_{i=1}^{n}\left[\frac{\exp\left(\boldsymbol{\beta}(W_i)^T\boldsymbol{Z}_i + g(W_i)\right)}{\sum_{j\in R(t_i)}\exp\left(\boldsymbol{\beta}(W_j)^T\boldsymbol{Z}_j + g(W_j)\right)}\right]^{\delta_i}, \tag{7.9}$$

which we localize in time (or at a certain value of the exposure variable if it measures something else) by the following Taylor expansions

$$\begin{aligned}\boldsymbol{\beta}(w) &\approx \boldsymbol{\beta}(w_0) + \boldsymbol{\beta}'(w_0)(w - w_0) \stackrel{def}{=} \boldsymbol{\tau} + \boldsymbol{\eta}(w - w_0)\\ g(w) &\approx g(w_0) + g'(w_0)(w - w_0) \stackrel{def}{=} \alpha + \gamma(w - w_0).\end{aligned}$$

When we insert these expressions into the partial likelihood function, note that α disappears. By analogy with (7.8), the local partial log-likelihood becomes

$$
\begin{aligned}
l(\gamma, \boldsymbol{\tau}, \boldsymbol{\eta}) = n^{-1} \sum_{j=1}^{N} & K_h(W_i - w_0) \\
& \times \Bigg[\boldsymbol{\tau}^T \boldsymbol{Z}_j + \boldsymbol{\eta}^T \boldsymbol{Z}_j (W_j - w_0) + \gamma (W_j - w_0) \\
& \quad - \log \Bigg(\sum_{i \in R(t_j)} \exp\{ \boldsymbol{\tau}^T \boldsymbol{Z}_i + \boldsymbol{\eta}^T \boldsymbol{Z}_i (W_i - w_0) + \gamma (W_j - w_0) \} \times K_h(W_i - w_0) \Bigg) \Bigg].
\end{aligned}
$$

Maximization yield estimators $\widehat{\boldsymbol{\beta}}(w_0) = \widehat{\boldsymbol{\tau}}(w_0)$ and $\widehat{g}(w_0)$ by integration of $\widehat{g}'(w_0) = \widehat{\gamma}(w_0)$.

Chapter 8

Concluding remarks

There is little doubt that local likelihood methods are valuable contributions to a statistician's toolbox. Some of the mentioned references point at useful applications, and many more exist that were not included in the preceding, more theoretically motivated work. Perhaps more important than the local likelihood function itself, is the idea of locally parametric estimation which, if used wisely, can draw on appealing characteristics from well-established methods, both parametric and non-parametric.

Judging by the calculations in Section 5.2, local likelihood estimates seem to be more robust against bad bandwidths than the kernel estimator, which, if true more generally, is a major advantage. Also, we see from the illustrations in Section 5.4 that we get good performance in the difficult areas of large curvature.

The theoretical foundation, however, is in need of more thorough investigations. In light of Chapter 4, it appears that the asymptotic behaviour of local likelihood estimates is highly dependant on the number of parameters in the parametric model. A general theory on asymptotic variance and bias like that of local polynomials is yet to be established. Also, it would be interesting to perform a more systematic study on performance compared with the kernel estimator based on smaller sample sizes than the simple simulations of Chapter 5.

Further, a review on which situations we may assume existence and uniqueness for the 'true' parameter $\boldsymbol{\theta}_0$, as defined for example by equation (4.1), would be useful.

Bibliography

O.O. Aalen, Ø. Borgan, and H.K. Gjessing. *Survival and event history analysis: a process point of view*. Springer Verlag, 2008. ISBN 0387202870.

H. Aksoy. Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, 24(6):419–428, 2000.

L. Barabesi. Local likelihood density estimation in line transect sampling. *Environmetrics*, 11(4):413–422, 2000.

A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962. ISSN 0162-1459.

N.E. Breslow. Comment on "Regression and life tables" by D.R. Cox. *Journal of the Royal Statistical Society, Series B*, 34:216–217, 1972.

K.P. Burnham and D.R. Anderson. Mathematical models for nonparametric inferences from line transect data. *Biometrics*, pages 325–336, 1976.

G. Casella and R.L. Berger. *Statistical inference*. Duxbury, 2002. ISBN 0534243126.

Jr. Cole, A. C. and Jr. Jones, J. W. A study of the weaver ant, oecophylla smaragdina (fab.)1. *American Midland Naturalist*, 39(3):pp. 641–651, 1948.

D.R. Cox. Partial likelihood. *Biometrika*, 62(2):269, 1975. ISSN 0006-3444.

D.R. Cox and E.J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275, 1968.

B. Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977. ISSN 0162-1459.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66. Chapman & Hall/CRC, 1996.

J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25(4):1661–1690, 1997.

J. Fan, H. Lin, and Y. Zhou. Local partial-likelihood estimation for lifetime data. *The Annals of Statistics*, 34(1):290–325, 2006.

R.A. Fisher. A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, 80:758–770, 1920.

N.L. Hjort and M.C. Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, 24(4):1619–1647, 1996. ISSN 0090-5364.

P.J. Huber and E. Ronchetti. *Robust statistics*, volume 1. Wiley Online Library, 1981.

K.O. Hufthammer and D. Tjøstheim. Describing multivariate dependence by local Gaussian Covariances. 2008a.

K.O. Hufthammer and D. Tjøstheim. Local gaussian likelihood and local gaussian correlation. 2008b.

M.C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

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*, pages 401–407, 1996.

J.D. Kalbfleisch and D.A. Sprott. Marginal and conditional likelihoods. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 311–328, 1973.

B. Kedem and K. Fokianos. *Regression models for time series analysis*. John Wiley and Sons, 2002. ISBN 0471363553.

E.L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer Verlag, 1998.

C.R. Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996. ISSN 0090-5364.

J.S. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 653–671, 1994.

T. Mäkeläinen, K. Schmidt, and G.P.H. Styan. On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *The Annals of Statistics*, 9(4):758–767, 1981. ISSN 0090-5364.

Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, USA, 2001. ISBN 0198507658.

K. Pearson and L.N.G. Filon. Mathematical contributions to the theory of evolution. iv. on the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 191:229–311, 1898.

J.A. Rice. *Mathematical statistics and data analysis*, volume 2. Duxbury press Belmont, CA, 1995.

J. Rinehart. Old faithful geyser performance 1870 through 1966. *Bulletin of Volcanology*, 33:153–163, 1969.

M.J. Schervish. *Theory of statistics*. Springer, 1995. ISBN 0387945466.

T.A. Severini. *Likelihood methods in statistics*. Oxford University Press, USA, 2000. ISBN 0198506503.

S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690, 1991.

B.W. Silverman. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

S.M. Stigler. The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620, 2007. ISSN 0883-4237.

R. Tibshirani and T. Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987. ISSN 0162-1459.

D. Tjøstheim and K.O. Hufthammer. Local gaussian correlation: A new measure of dependence. 2012.

A.A. Tsiatis. A large sample study of Cox's regression model. *The Annals of Statistics*, 9 (1):93–108, 1981. ISSN 0090-5364.

M.P. Wand and M.C. Jones. *Kernel smoothing*, volume 60. Chapman & Hall/CRC, 1995.

W.H. Wong. Theory of partial likelihood. *The Annals of Statistics*, 14(1):88–123, 1986. ISSN 0090-5364.

Y. Zhang, GA Bishop, and D.H. Stedman. Automobile emissions are statistically gamma distributed. *Environmental science & technology*, 28(7):1370–1374, 1994.