# Local Smoothing Methods for the Analysis of Multivariate Complex Data Structures

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

Jochen Einbeck
am 12. August 2003

# Local Smoothing Methods for the Analysis of Multivariate Complex Data Structures

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

Jochen Einbeck
am 12. August 2003

| | |
|---|---|
| 1. Gutachter: | Prof. Dr. G. Tutz |
| 2. Gutachter: | Prof. Dr. L. Fahrmeir |
| 3. Gutachter: | Prof. J. O. Ramsay, Ph.D. |

| | |
|---|---|
| Rigorosum: | 27.11.2003 |

# Vorwort

Bedanken möchte ich mich zuallererst bei meinem Doktorvater Gerhard Tutz, der mir die entscheidenden Impulse zu dieser Arbeit gab, mir dabei aber auch viel Freiheit in der Auswahl der Forschungsschwerpunkte ließ. Nun ja, so viel Auswahl gab es vielleicht gar nicht, nachdem mich jedes wie auch immer geartete Forschungsprojekt zu guter letzt zwangsläufig zu lokalen Glättungsmethoden hinzutreiben scheint, so dass mein Chef kürzlich, beinahe resignierend, feststellte: *"Sie sind halt einfach ein Lokaler..."*. Den Grundstein für meine innige Beziehung zu Kernfunktionen, und allem, was damit zusammenhängt, legte zu Zeiten meiner Staatsexamensarbeit Helmut Pruscha, dem ich dafür ebenfalls danke. Mein besonderer Dank gilt auch den Koautoren, die neben Gerhard Tutz an einzelnen Teilen dieser Arbeit mitgewirkt haben, allen voran Göran Kauermann, der mir zu Beginn der Promotion maßgeblich geholfen hat, in die wissenschaftliche Welt hineinzufinden, weiterhin Julio da Motta Singer und Carmen Diva Saldiva de André, mit denen die Zusammenarbeit in São Paulo einen ungeheuren Spaß gemacht hat, und schließlich Ludger Evers, dessen Diplomarbeit mitzubetreuen eine wahre Freude war. Bedanken möchte ich mich ferner bei Ludwig Fahrmeir und Jim Ramsay, die sich freundlicherweise zur Begutachtung dieser Arbeit bereit erklärt haben, sowie bei meinen Kollegen am Institut, die für ein familiäres und freundschaftliches, aber auch motivierendes, Arbeitsklima gesorgt haben.

Ferner danke ich meinen Eltern, die mein Interesse an Zahlen und Daten schon sehr früh weckten, Paul Hix, der mir hilfreich zur Seite stand, wenn ich mal wieder an die Grenzen meiner Englischkenntnisse stieß, Martin Pohl, der mir unermüdlich klarzumachen versuchte, dass die Welt nicht nur aus Statistik besteht, sowie den "Eberwurzern" und den am Institut für ihre fünften Plätze berüchtigten "Zufallstreffern", die dafür sorgten, dass die nötige Ablenkung mittels eines ledernen Pentakisdodekaeders erfolgte.

Danken möchte ich schließlich meiner Tochter Nathalie, die mit mir viele Stunden zu Hause mehr oder weniger geduldig über irgendwelchen Statistik-Papern verbrachte (während die Mutti beim Deutschkurs war), und die sich mittlerweile in nichtparametrischer Statistik auch schon recht gut auskennt... Além disso queria agradecer minha esposa Flávia, que mudou para Alemanha para eu poder acabar o meu doutorado aqui, e que me apoiou com paciência e amor. Obrigado, meu anjo!

## Zusammenfassung

Ein weitverbreiteter Ansatz zur Lösung nichtparametrischer Regressionsprobleme besteht darin, eine Gerade oder ein Polynom lokal - d.h. nur in einer bestimmten Umgebung des jeweils interessierenden Punktes - an die Daten anzupassen. Diese Arbeit widmet sich zunächst der Frage, wie die bestehenden Konzepte verallgemeinert, und damit, wenn möglich, auch verbessert werden können. Der Schwerpunkt liegt hierbei auf Generalisierungen in zwei Richtungen: Einerseits werden die Polynome durch geeignete glatte allgemeine Basisfunktionen ersetzt, und andererseits wird der Schätzmechanismus, der auf der Kleinste-Quadrate-Methode beruht, geeignet modifiziert. Wie sich zeigt, ist ersteres Konzept vornehmlich zur Bias- Reduktion interessant, wohingegen letzteres vor allem unter dem Gesichtspunkt der Ausreißer-Robustheit sinnvoll einsetzbar ist. Es ergeben sich hierbei interessante Parallelverbindungen zu anderen Gebieten der Mathematik und Statistik, insbesondere zu den bekannten Sätzen von Taylor und Horvitz-Thompson.

Im weiteren Verlauf der Arbeit werden lokale Ansätze für bestimmte Problemstellungen herangezogen, für die sie bisher wenig beachtet wurden. Es stellt sich heraus, dass lokale Glättungsmethoden, geeignet kombiniert, sinnvoll für das Online-Monitoring anwendbar sind, d.h. zur Echtzeitüberwachung von Zeitreihen hinsichtlich plötzlicher Strukturbrüche. Zum Ende wird der bisherige Rahmen des funktionalen Zusammenhangs zwischen abhängigen und unabhängigen Variablen verlassen und ein lokaler Ansatz zur Berechnung glatter Kurven durch die “Mitte” einer mehrdimensionalen, möglicherweise verzweigten, Datenwolke entwickelt.

## Summary

Local smoothing methods are a widely used tool in the context of nonparametric regression. The essential idea is to perform a linear or polynomial regression locally in a neighborhood of the target point. This method is generalized in two ways. Firstly, the polynomials are substituted by arbitrary smooth basis functions, and secondly, the estimating methodology, which is based on the least squares method, is modified in a suitable way. It appears that the first concept is useful for bias reduction, while the second one is interesting for robustifcation against outliers in the predictors. As by-products some interesting relations to other mathematical and statistical topics are unveiled, concerning in particular the theorems from Taylor and Horvitz-Thompson.

In the further course of the thesis the interest turns to some particular problems which have not been a domain of local methods so far. It turns out that local smoothing methods, suitably combined, are useful for the online monitoring of time series in order to detect sudden breaks or jumps. Finally, the restriction of modelling only functional data is abandoned and a new approach to calculate principal curves, i.e. smooth curves which pass through the "middle" of a multidimensional, possibly multiply branched, data cloud, is developed.

# Contents

**1 Introduction** **1**

1.1 Thoughts about local smoothing . . . . . . . . . . . . . . . . . . . . . 1

1.2 Guideline through the thesis . . . . . . . . . . . . . . . . . . . . . . . 6

**2 Local Fitting with General Basis Functions** **9**

2.1 Univariate predictors . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.1.2 The power basis . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.1.3 Asymptotics . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

2.1.4 A simulated example . . . . . . . . . . . . . . . . . . . . . . . 16

2.1.5 Bias reduction with data-adaptive basis functions . . . . . . 18

2.1.6 Notes about bandwidth selection . . . . . . . . . . . . . . . . 21

2.1.7 A real data example . . . . . . . . . . . . . . . . . . . . . . . 23

2.1.8 Outlook . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

2.1.9 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

2.2 Multivariate predictors . . . . . . . . . . . . . . . . . . . . . . . . . . 31

2.2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

2.2.2 Multivariate locally weighted regression using a general basis 32

2.2.3 Asymptotics for basis functions without interactions . . . . . 34

2.2.4 Example . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

2.2.5 Finding a data-driven basis function . . . . . . . . . . . . . . 38

2.2.6 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

2.2.7 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

2.3 Remarks on the generalized Taylor expansion . . . . . . . . . . . . 47

2.3.1 Univariate generalized Tayor expansion . . . . . . . . . . . . 47

2.3.2 Multivariate generalized Taylor expansion . . . . . . . . . . 54

**3 Local Smoothing with Robustness against Outlying Predictors 59**

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

3.2 An overview of related concepts . . . . . . . . . . . . . . . . . . . 61

3.2.1 Ridging . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

3.2.2 Variable bandwidth . . . . . . . . . . . . . . . . . . . . . . 61

3.3 Robustness against outlying predictors . . . . . . . . . . . . . . . 62

3.3.1 Soft robustification . . . . . . . . . . . . . . . . . . . . . . 62

3.3.2 Hard robustification . . . . . . . . . . . . . . . . . . . . . 63

3.3.3 Example . . . . . . . . . . . . . . . . . . . . . . . . . . . 64

3.3.4 Simultaneous robustness for predictor and response variables 68

3.3.5 Some notes about bandwidth selection . . . . . . . . . . . . 69

3.4 Relative risk curves for respiratory deaths . . . . . . . . . . . . . . 72

3.5 Discussion and Outlook . . . . . . . . . . . . . . . . . . . . . . . 74

3.6 Additional asymptotics . . . . . . . . . . . . . . . . . . . . . . . 74

3.7 Relation to the Horvitz-Thompson estimator . . . . . . . . . . . . . 79

**4 Online Monitoring with Local Smoothing Methods and Adaptive Ridging 81**

4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 81

4.2 Local linear smoothing and breakpoint detection . . . . . . . . . . . 83

4.3 Practical adjustments . . . . . . . . . . . . . . . . . . . . . . . . 87

4.3.1 Ridging . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

4.3.2 Choice of $h$, $h_1$ and $h_2$. . . . . . . . . . . . . . . . . . . . . 90

4.3.3 Missing values and outliers . . . . . . . . . . . . . . . . . . . 92

4.3.4 Variance calculation . . . . . . . . . . . . . . . . . . . . . . 93
4.4 Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93
4.4.1 Cardio beats . . . . . . . . . . . . . . . . . . . . . . . . . 93
4.4.2 ECG measurements . . . . . . . . . . . . . . . . . . . . . 95
4.5 Comparison to other methods . . . . . . . . . . . . . . . . . . . . 97
4.5.1 Autoregressive models . . . . . . . . . . . . . . . . . . . . 97
4.5.2 Phase space models . . . . . . . . . . . . . . . . . . . . . 98
4.5.3 State space models . . . . . . . . . . . . . . . . . . . . . 98
4.6 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100
4.7 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100

**5 Local Principal Curves** **103**

5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . 103
5.2 The algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . 106
5.3 Technical details . . . . . . . . . . . . . . . . . . . . . . . . . 108
5.3.1 Maintainig the direction . . . . . . . . . . . . . . . . . . 108
5.3.2 Running backwards from $x_{(0)}$ . . . . . . . . . . . . . . . . . 110
5.3.3 Angle penalization . . . . . . . . . . . . . . . . . . . . . 110
5.3.4 Multiple initializations . . . . . . . . . . . . . . . . . . . 111
5.4 Examples . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
5.4.1 2-dimensional data . . . . . . . . . . . . . . . . . . . . . 111
5.4.2 3-dimensional data . . . . . . . . . . . . . . . . . . . . . 113
5.5 Theoretical justification . . . . . . . . . . . . . . . . . . . . . 117
5.6 Coverage . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
5.7 Selection of parameters . . . . . . . . . . . . . . . . . . . . . . 121
5.8 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

**6 Perspectives** **127**

**References** **130**

# Chapter 1

# Introduction

## 1.1 Thoughts about local smoothing

The roots of local polynomial modelling reach far back into the nineteenth century. The Italian meteorologist Schiaparelli (1866) and the American mathematician De Forest (1873) were amongst the first to work on local polynomials. In these early approaches the predictors were confined to be fixed integers, as e.g. in Spencer (1904), who fitted mortality rates against age. The methods were based on graduation, i.e. a suitably weighted variant of moving averages. Henderson (1916) further elaborated the concept, but then the topic was largely forgotten in statistics. The ideas however survived in economic time series, starting with the book by Macaulay (1931) and producing the program X-11 in 1954, developed by the U.S. Bureau of Census, which provided the first computer implementation of smoothing methods. Details about X-11 are found in Shiskin, Young & Musgrave (1967).

At that time, statisticians rediscovered the issue and began to develop local modelling as it is understood today, weighting the (possibly random) independent variables locally by means of kernel functions. The first ones to treat this problem were Nadaraya (1964) and Watson (1964), who confined to the case of local constant smoothing. Stone (1977), Cleveland (1979), Katkovnik (1979) and Tsybakov (1986) did groundbreaking work on extending this concept to local polynomials. We will not provide a complete repetition of the theory and application of local polynomial modelling in this introduction, since this will be done implicitely in the introductions of the following chapters. For now, we will briefly provide as much backgound as necessary for this introduction. For bivariate data $(X_i, Y_i), i = 1, \ldots, n$, the estimation of the regression function $m(\cdot)$ at a target point $x$ is performed by minimizing

$$\sum_{i=1}^{n}\left\{Y_i-\sum_{j=0}^{p}\beta_j(X_i-x)^j\right\}^2 K\left(\frac{X_i-x}{h}\right) \tag{1.1}$$

in terms of $\beta_0,\ldots,\beta_p$, where $K$ is a kernel function and $h$ a bandwidth, yielding an estimate $\hat{m}(x)=\hat{\beta}_0$ according to Taylor's theorem. The bandwidth $h$ may either be constant, or depend on the predictors $X_i$ (global variable bandwidth) or the target point (local variable bandwidth). The bandwidth can be selected using classical methods such as cross-validation, the AIC criterion, etc., or plug-in-methods, which perform a pilot estimate of the mean squared error, which is then minimized in terms of the bandwidth.

However, the existence of this method did not guarantee its acception in the statistical community, and the path leading to the breakthrough was everything but smooth. In the early eighties the scientific interest once again left the local polynomials and attention turned to the development of modified kernel type methods, in particular to improve the poor performance of the Nadaraya-Watson estimator at the boundary (Gasser, Müller & Mammitzsch, 1985, Müller, 1991). In that period it seemed to be more elegant to modify the kernel function of local constant estimators than to work with polynomial degrees $p>0$. Chu & Marron (1991) even discarded local polynomial regression as an *"obscure alternative method"*. Fan & Marron (1993) wondered *"why it took so long for the smoothing community ... to understand fully the benefits of local polynomials"*. They speculate that the reasons for this were the "equivalence results" (Müller, 1987), *"whose main intuitive message was that for equally spaced $x_j$, away from the boundary, there is essentially no difference between local polynomial and traditional kernel estimators"* for a suitably chosen (or better: designed) kernel $K$. However, the effort to obtain those kernels is extraordinarily high, and these methods are *"not worth the space they take up in the practicing statistician's toolbox"* (Hastie & Loader, 1993b), whereas local polynomial smoothers adapt to boundary points or random design automatically. It was mainly the merit of Fan (1992) and Hastie & Loader (1993a) to point this out clearly and to renew the interest in local polynomial regression, setting off a flood of publications in the middle of the last decade. As a landmark work we mention Ruppert & Wand (1994), who developed asymptotics for multivariate local polynomial regression. Indeed, it were the asymptotic results which enabled in a better way to compare local smoothers and to perform inference for them, leading to a deeper insight into the benefits of local polynomial smoothing. For an overview of the state of the art we may refer to two excellent books written by Fan & Gijbels (1996) and Loader (1999b), summing up concepts, theoretical results and applications of local polynomial modelling. The divulgence of local polynomials was accelerated by the S-Plus implementation of Cleveland's LOWESS method (for a description see Cleveland & Grosse, 1991). A more capacious software package named LOCFIT has meanwhile been developed by C. Loader and is described in his above mentioned book.

Local polynomial regression is widely used today and has proven to have nice theoretical properties and to work fine in practice. Nevertheless, it suffers from a somewhat bad reputation. Raymond Carroll (2003) stated on the occasion of a talk at the Department of Statistics, University of Munich, that *"the world is moving towards the splines"*, and that the most noteworthy benefit of local polynomial smoothing is that it is theoretically easy to analyze. A variety of other points are still discussed within the local smoothing community or challenged from outside. The remainder of this section is dedicated to the discussion of these issues.

The largest reproach to local smoothing methods is that they only work well for univariate predictors, at the most for bivariate predictors, since for higher dimensions the data get too sparse and the number of parameters rises too rapidly, commonly called the "curse of dimensionality" (Bellmann, 1961). This problem is even admitted by the most convinced local smoothing propagators. However, even for high dimensions local smoothing does not necessarily need to fail. Cleveland & Devlin (1988) state that *"some have mistakenly supposed that the curse makes multivariate smoothing"* - that is, smoothing with $d > 1$ independent variables, *"a method to avoid. What must be avoided is allowing the bandwidth to remain fixed as $d$ increases..."*. They demonstrate that local smoothing with three independent variables makes sense, and Fowlkes (1986) even successfully employs local smoothing for more than three independent variables. Of course, the bandwidth cannot be arbitrarily increased, since then the curvature of the underlying function cannot be recovered properly. Hastie, Tibshirani & Friedman (2001) note that *"the complexity of functions of many variables can grow exponentially with the dimension, and if we wish to be able to estimate such functions with the same accuracy as functions in low dimensions, then we need the size of our training set to grow exponentially as well."* However, citing again Cleveland & Devlin (1988), *"this is not a defect of the method but a statement that the more complicated the regression surface becomes, the larger $n$ must be to get good estimates of it. Exactly the same considerations obtain whatever the method of estimation."* Thinking this point further, if local methods fail for a certain multivariate data problem, this is due to a lack of relevant information around the target point. This information is lacking regardless of the smoothing method, and any smoothing method yielding a result for the same data arouses the suspicion that it relies on data points that are *not* relevant for the target point. It is a philosophical question whether a possibly *unreliable* estimation is preferred to none at all. We will not judge this here, but keep in mind that a perpetuum mobile does not exist, neither in physics nor in statistics. As a solution to those problems, one has to *"attack the curse of dimensionality by implying some structure in the predictor space"* (Fahrmeir & Tutz, 2001) and therefore some kind of foregoing dimension reduction or model simplifying is necessary. Fan & Gijbels (1996, p. 264ff) and Loader (1999b, p. 53ff) describe a variety of such methods, which still enable local smoothing for high dimensional data.

Another issue that has frequently been raised is that of computational speed. It had been *"general folklore in smoothing ... that smoothing splines are much faster computationally"* (Fan & Marron, 1993) than local polynomial methods. Using fast implementations of kernel methods as binning (Härdle & Scott, 1992), local polynomial methods are, however, *"at least competitive with smoothing splines"* (Fan & Marron, 1993). For univariate predictors there is not much difference in speed, since $O(n)$ implementations exist for either method (Hastie & Loader, 1993b, Fan & Marron, 1994, Fan & Gijbels, 1996). For multivariate predictors, the local polynomial approach even seems to be superior. According to Cleveland & Devlin (1988), their `loess` method achieves $O(n)$ speed in multiple dimensions, while fitting thin plate splines to two or more independent variables is an $O(n^3)$ computation (Wahba, 1984). However, this comparison, also given in Hastie & Loader (1993b), is not quite fair: The $O(n)$ complexity for local polynomial smoothing does *not* imply bandwidth selection, but the $O(n^3)$ time for splines includes the selection of parameters. A recent work by Wood (2003) even shows that, using thin plate regression splines, the complexity can be reduced to nearly $O(n^2)$. Including bandwidth selection, local smoothers might achieve magnitudes about $O(n^2)$ as well, but will hardly be faster.

Bandwidth selection has been one of the most discussed problems in the last years, and the controversy is still ongoing with undiminshed eagerness. In the nineties, a large number of authors, e.g. Park & Marron (1990), tried to show that the classical methods are inferior to plug-in-methods. Ruppert, Sheather & Wand (1995) conclude that *"one of the main findings of this research is that traditional smoothing methods, such as those based on cross-validation, exhibit very inferior asymptotic and practical performance"*. This, however, is questionable, as Loader (1999a) demonstrates in a startling work. He argues that *"plug-in methods are heavily dependent on arbitrary specification of pilot bandwidths and fail when this specification is wrong. The often-quoted variablility and undersmoothing of cross-validation simply reflects the uncertainty of bandwidth selection"* and is not attributable to deficiencies of classical methods. Asymptotically, plug-in based estimates are even beaten by their own pilot estimates, and the allegedly better asymptototic performance in comparison to classical methods is due to the fact that some extra assumptions are made, which the classical methods do not need. Local methods have been criticized that bandwidth selection is practically cumbersome, since for a flexible modelling of the regression function a variable bandwidth has to be applied. Using plug-in methods, variable bandwidth selection is admittedly awkward and compuationally demanding. However, variable bandwidth selection is possible straightforwardly within classical methods. For example, cross-validation may easily be extended to local cross-validation (Loader, 1999b, p. 198). This thesis addresses bandwidth selection for special local smoothing settings in Sections 2.1.6, 3.3.5, 4.3.2 and 5.7.

Local polynomial methods have proven to have excellent asymptotic properties. In particular, Fan (1993) showed that the local linear smoother achieves 100% minimax efficiency (which means that the local linear smoother minimizes the mean squared error over all linear smoothers, i.e. smoothers of the form $\sum w_i Y_i$, simultanously for all smooth regression functions $m$). It is often questioned whether these theoretical results have significance in reality, since for real data the condition $n \longrightarrow \infty$ is never fulfilled, and the condition $h \longrightarrow 0$ in practice implies that *no* data are situated in the support of the kernel. Matt Wand (2001), who surely cannot be charged with being reluctant to develop asymptotics, finished his talk at the Euroworkshop of Statistical Modelling in Höhenried, Germany with the words: *"Don't dwell in asymptotics"*. Asymptotics do not provide any reliable rule of how real data behave or how their properties are. Asymptotics are a *helpful tool* to compare smoothing procedures by means of the mean squared error or minimax properties (whereby local polynomial smoothers are competitive or superior compared to any other smoothing method), and they provide some hints on how to choose tuning parameters which are otherwise possibly hard to find. This thesis contains a big part of asymptotics, with five theorems included in Sections 2.1.3, 2.2.2, 2.2.3 and 3.6. The latter one however typifies the discrepance which might occur between theory and practice: The asymptotical result is even contrary to our practical recommendation given in Section 3.3.

Smoothing methods in general are weak in applications as structure or pattern recognition, in particular edge detection and the like, since they smooth about edges and thus destroy the structure which was to be detected. However, this is fortunately not a particular problem of *local* smoothing methods. Localization even enables a more flexible handling of sudden changes of the data structure than other concepts. Chu, Glad, Godtliebsen & Marron (1998) describe two variants of the Nadaraya-Watson estimator which are feasible for edge detection and applications such as image processing. Kauermann (2001) recently provided a related concept based on local mixture modelling. In this thesis, two chapters deal explicitely with problems related to pattern recognition. In Chapter 4 we provide an algorithm for the detection of breakpoints of online monitored time series, and in Chapter 5 we use local methods to introduce a new method to find principal curves.

Reflecting all these points, there seems to be no reason to discriminate local polynomial regression in comparison to other smoothing techniques. Theo Gasser, quoted by Fan & Marron (1993), even states that *"we have not found any disadvantages of the local polynomial method as yet. It should become a golden standard non-parametric technique"*. The path to acceptance as a golden technique will continue to be laborious. In this thesis, we will try to sweep some obstructions away by applying local (polynomial) methods on topics which are not yet known to be a domain of local smoothing (online breakpoint detection, outlying predictors, principal curves), and partly construct new obstacles by developing improved local smoothing methods which show that the local polynomial approach is not always the optimal solution.

## 1.2 Guideline through the thesis

This thesis is divided into four substantial chapters, which address various issues concerning data analysis with local smoothing methods. Chapters 2 and 3 treat extensions of local polynomial modelling. In particular, in Chapter 2 the linear basis $X_i - x$ in (1.1) is substituted by general basis functions $\phi(X_i) - \phi(x)$ and it is shown that this leads to an improvement of the performance of the estimates in certain situations. In Chapter 3 an additional weight function $\alpha(\cdot)$ is employed in the minimizing problem (1.1) with the objective of achieving robustness against outlying predictors. Chapter 4 describes an application of local smoothing techniques. In Chapter 5 we abandon the restriction of fitting *functions*, i.e. we fit spirals and other complex patterns, via a new approach for finding principal curves.

**Local Fitting with General Basis Functions**

In this chapter the method local poynomial fitting, which can be seen as a local fit of the data against the basis functions $1, x, \ldots, x^p$, is extended to a more general class of basis functions. In the univariate case we focus on the power basis, i.e. a basis which consists of the powers of an arbitrary smooth function, and derive an extension of Taylor's theorem for this basis. Using this theorem, some asymptotic properties of the new estimator are derived. It is shown by means of simulated examples and theoretical considerations that significant bias reduction is possible if the basis carries some information about the underlying function. Finally, some remarks about bandwidth selection are given and the method is applied to real data. Similar work is done in the multivariate setting. The asymptotic variance is calculated for arbitrary smooth multivariate basis function. Computation of the asymptotic bias, however, again requires to generalize Taylor's theorem, which is only possible by a confinement to basis functions without interactions. We provide an algorithm to construct a data-adaptive basis and demonstrate (more impressively than in the univariate setting) that local fitting with general basis functions makes sense and may lead to significant improvements of the results. The content of Section 2.2 has been published in "Computational Statistics" (Einbeck, 2003). In addition, some remarks concerning the generalized Taylor theorems are given in Section 2.3.

**Local Smoothing with Robustness against Outlying Predictors**

Outlying pollutant concentration data are frequently observed in time series studies conducted to investigate the effects of atmospheric pollution and mortality/morbidity. These outliers may severely affect the estimation procedures and even generate unexpected results like a protective effect of pollution. Although robust methods have been proposed to downweight the effect of outliers in the response variable distribution, little has been done to handle outlying explanatory variable values. We consider a robust local polynomial smoothing technique, based on downweighting points with a small design density by plugging the estimated density into the minimizing problem (1.1), which may be useful for such purposes. Using data from a study conducted in São Paulo, Brazil, it is shown how an unexpected form of the relative risk curve of mortality attributable to pollution by $SO_2$ obtained via nonrobust methods may be completely reversed when the proposed technique is employed. We focus on univariate local linear smoothing, though the method is not restricted to this case, since the weights may be plugged into any loss or likelihood function which is to be minimized or maximized, respectively. The last two sections are dedicated to providing some deeper insight into the theoretical properties of the new estimator, unveiling a surprising analogy to the Horvitz-Thompson estimator.

**Online Monitoring with Local Smoothing Methods and Adaptive Ridging**

In this chapter an application of standard local smoothing procedures is provided. The objective is to detect jumps or breaks of trends in time series which are recorded online, as met e.g. in clinical information systems. "Online" thereby implies that at each time point the future (i.e. all data points to the right) is unknown and one has to make a decision based on observations from the past. At a time point $t$, we therefore consider a long term estimate and a short term estimate (where the long term estimate uses information over a larger time span than the short term estimate) at the value $t$. The essential idea is that these estimates yield similar results when no break point has occured in the early past, but will differ when there has been a break of trend or jump shortly before. The long term estimate is calculated by combining local constant and local linear smoothing in a ridge estimate. The methodology is demonstrated by various examples. The paper is to appear in "Journal of Statistical Computation and Simulation" (Einbeck & Kauermann, 2003).

**Local Principal Curves**

Principal Curves are a relatively new topic, firstly introduced by Hastie & Stuetzle (1989) as a generalization of principal components, allowing for general smooth curves which pass through the "middle" of a multidimensional data cloud. Since principal curves are not restricted to fit *functions*, concepts which are based on minimizing a vertical residual criterion as in (1.1) are not applicable, and consequently completely new strategies are necessary. Today, a variety of definitions of principal curves exist and accordingly a variety of algorithms to estimate them. However, the large majority of them are *global* methods, i.e. all data points are used to estimate the principal curve at any location. Apart from some technical problems, global approaches lead to exploding computational costs for highdimensional data. A new tool to estimate *local* principal curves is provided, which is based on localized principal components. Tuning parameters can easily be selected regarding the coverage, i.e. the proportion of the data "covered" by the principal curve. The method is applied on simulated and real data. In particular, we show that the concept is useful to reconstruct geographical structures as river outlines.

# Chapter 2

# Local Fitting with General Basis Functions

## 2.1 Univariate predictors

### 2.1.1 Introduction

In the last years a huge amount of literature about local polynomial modelling has been published. An overview of the current state of the art was given in Fan & Gijbels (1996). Various extensions followed, in particular about ridging (Seifert & Gasser, 2000), bias reduction (Choi & Hall, 1998), the treatment of measurement errors (Carroll, Maca & Ruppert, 1999 and Lin & Carroll, 2000) and improvements concerning the shape of the smoothing matrix (Zhao, 1999) of the local linear smoother.

Local polynomial modelling is proposed for fitting data which cannot be modelled satisfactory by global polynomials, as for example the famous motorcycle data (see Section 2.1.7). In the following a short review of this method is given. Consider bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which form an i.i.d. sample from a population $(X, Y)$. We assume the data to be generated from a model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{2.1}$$

where $E(\varepsilon) = 0, Var(\varepsilon) = 1$, and $X$ and $\varepsilon$ are independent. Of interest is to estimate the regression function $m(x) = E(Y|X = x)$ and its derivatives $m'(x), m''(x), \ldots,$ $m^{(p)}(x)$. A Taylor expansion yields

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j(x)(z-x)^j, \tag{2.2}$$

given that the $(p+1)^{\text{th}}$ derivative of $m(\cdot)$ in a neighborhood of $x$ exists. We define $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$, where $K$ is a kernel function which is usually taken to be a non-negative density symmetric about zero. $h$ denotes the bandwidth which determines the size of the neighborhood of which the covariate values which shall influence the fit are chosen. The task of finding the appropriate bandwidth is the crucial point of local polynomial fitting. See Section 2.1.6 for an overview on literature concerning bandwidth selection. Minimizing

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j(x)(X_i - x)^j \right\}^2 K_h(X_i - x)$$

leads to the locally weighted least squares regression estimator $\hat{\beta}(x) = (\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))^T$ and the corresponding estimation functions

$$\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x) \tag{2.3}$$

for $m^{(j)}(x), j = 0, \ldots, p$. Alternative approaches focussed on robust nonparametric regression (Fan, Hu & Truong, 1994) or on estimating the conditional median or quantiles instead of the mean function (Honda, 2000 and Yu & Jones, 1998).

According to equation (2.2), we model data pairs $(X, Y)$ locally around $x$ by

$$Y = \beta_0(x) + \beta_1(x)(X - x) + \ldots + \beta_p(x)(X - x)^p + \sigma(X)\varepsilon.$$

By transforming parameters, this can be written as

$$Y = \alpha_0(x) + \alpha_1(x)X + \ldots + \alpha_p(x)X^p + \sigma(X)\varepsilon.$$

Thus, local polynomial modelling can be interpreted as fitting the data locally against the basis functions $1, X, X^2, \ldots, X^p$. An obviously arising question is now: Why should just these basis functions be the best possible ones? In the sequel we will extend the theory of local polynomial fitting, which is restricted to polynomial basis functions, to a wide range of other basis functions, and give an idea of the advantages and problems coming up by using arbitrary basis functions. A possible choice of basis functions are e.g. Gaussian kernels or the trigonometric functions.

In a general framework one may use the basis functions $\phi_0(X)$, $\phi_1(X), \ldots, \phi_p(X)$, which are arbitrary differentiable functions $\phi_i : \mathbb{R} \mapsto \mathbb{R}, i = 0, \ldots, p$. This can lead to very good - and tremendously bad - results, as is shown in Section 2.1.8. However, theoretical results for the local estimator are only available under some restrictions on the basis functions. Regarding (2.2) and (2.3), it is seen that estimation is based on Taylor's expansion. Taylor's theorem is also necessary for the derivation of important asymptotic results as the asymptotic bias of the local polynomial estimator. Asymptotics are a useful tool to find bandwidth selection rules etc., so that they play an important role for the use of the estimator in practice.

Thus, if we want some theoretical background, we need to develop a new Taylor expansion for every basis we want to use. Of course this will not be possible for all choices of basis functions. In the following section we focus on a special case, namely the power basis, where this is in fact possible and describe the estimation methodology. In Section 2.1.3 we provide some asymptotics for estimating the conditional bias and variance of this estimator. In Section 2.1.4 we apply this method to a simulated data set and compare the results for various basis functions. We improve these results significantly in Section 2.1.5 by using data-adaptive basis functions. In Section 2.1.6 we give some remarks on bandwidth selection. We apply the method on a real data set in Section 2.1.7, and in Section 2.1.8 we show some results obtained when using the most general model.

### 2.1.2 The power basis

**Definition 2.1.**
*Let $\phi : \mathbb{R} \longrightarrow \mathbb{R}, z \mapsto \phi(z)$ be a differentiable function. Then the functions*

$$1, \phi(z), \dots, \phi^p(z)$$

*are called a power basis of degree $p$.*

Taylor's theorem, as found for example in Lay, 1990 (p. 211), can be extended as follows:

**Theorem 2.1 (Taylor expansion for a power basis).**
*Let $I$ be a non-trivial interval, $m, \phi : I \longrightarrow \mathbb{R}$ $p+1$ times differentiable in $I$, $\phi$ invertible in $I$, and $x \in I$. Then for all $z \in I$ with $z \neq x$ a value $\zeta \in (x, z)$ resp. $(z, x)$ exists so that*

$$m(z) = \sum_{j=0}^{p} \frac{\psi_{(j)}(x)}{j!}(\phi(z) - \phi(x))^j + \frac{\psi_{(p+1)}(\zeta)}{(p+1)!}(\phi(z) - \phi(x))^{p+1} \tag{2.4}$$

*with*

$$\psi_{(j+1)}(\cdot) = \frac{\psi'_{(j)}(\cdot)}{\phi'(\cdot)}, \psi_{(0)}(\cdot) = m(\cdot), \tag{2.5}$$

*holds.*

The proof of this theorem is found in the appendix. Let $\phi, m$ as in Theorem 2.1 and $g(\cdot) = (m \circ \phi^{-1})(\cdot)$. Applying Taylor's theorem on $m = g \circ \phi$ and comparing the result with (2.4) yields

$$\psi_{(j)}(\cdot) = (m \circ \phi^{-1})^{(j)}(\phi(\cdot)). \tag{2.6}$$

Assuming the underlying model (2.1), Theorem 2.1 suggests to model the data in a neighborhood of $x$ by

$$Y = \gamma_0(x) + \gamma_1(x)(\phi(X) - \phi(x)) + \ldots + \gamma_p(x)(\phi(X) - \phi(x))^p + \sigma(X)\varepsilon \tag{2.7}$$

where

$$\gamma_j(x) = \frac{\psi_{(j)}(x)}{j!}.$$

One might find the constants $\phi(x)$ disturbing. However, (2.7) can easily be transformed to the model

$$Y = \delta_0(x) + \delta_1(x)\phi(X) + \ldots + \delta_p(x)\phi^p(X) + \sigma(X)\varepsilon \tag{2.8}$$

by setting

$$\begin{aligned} \delta_j &= \gamma_j - \gamma_{j+1}\binom{j+1}{1}\phi(x) + \gamma_{j+2}\binom{j+2}{2}\phi^2(x) \mp \ldots + \\ &\quad + (-1)^{p-j}\gamma_p\binom{p}{p-j}\phi^{p-j}(x), \end{aligned}$$

(where for ease of notation $\gamma_j \equiv \gamma_j(x), \delta_j \equiv \delta_j(x)$ ). Thus model (2.7) and (2.8) yield the same computational results when used for fitting the function $m$. The advantage of working with model (2.7) is that its theoretical properties are easier to derive, since the theorem given above can be applied. Moreover, computation is faster and more stable, because very large values are avoided by subtracting $\phi(x)$ under the powers.

Since the parameters $\gamma_j$ are constructed more complex than the parameters $\beta_j \equiv \beta_j(x)$ for local polynomial fitting, the simple relationship $m^{(j)}(x) = j!\beta_j$ cannot be retained. However, by using the simple recursive formula

$$\gamma_j(x) = \frac{1}{j\phi'(x)}\gamma'_{j-1}(x), \quad \gamma_o(x) = m(x),$$

the parameters $\gamma_j$ can be calculated and thus the following relations between parameters and the underlying function and their derivatives are derived for the power basis:

$$\begin{aligned} m(x) &= 0!\,\gamma_0 && (2.9)\\ m'(x) &= 1!\,\phi'(x)\,\gamma_1 && (2.10)\\ m''(x) &= 2!\,[\phi'(x)]^2\,\gamma_2 + \phi''(x)\,\gamma_1 && (2.11)\\ m'''(x) &= 3!\,[\phi'(x)]^3\,\gamma_3 + 3!\,\phi''(x)\phi'(x)\,\gamma_2 + \phi'''(x)\,\gamma_1 && (2.12)\\ &\vdots \end{aligned}$$

This indicates that for estimating the $j^{\text{th}}$ derivative of $m$, the basis function $\phi$ has to be $j$ times differentiable in an environment of $x$. In the following we will shortly describe the estimation procedure. In order to estimate $\hat{\gamma}(x) = (\hat{\gamma}_0, \ldots, \hat{\gamma}_p)^T$, a locally weighted least squares regression has to be run, i.e.

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \gamma_j (\phi(X_i) - \phi(x))^j \right\}^2 w_i \tag{2.13}$$

(with $w_i = K_h(X_i - x)$) has to be minimized in terms of $(\gamma_0, \ldots, \gamma_p)$. The design matrix and the necessary vectors are given by

$$X_x = \begin{pmatrix} 1 & \phi(X_1) - \phi(x) & \cdots & (\phi(X_1) - \phi(x))^p \\ \vdots & \vdots & & \vdots \\ 1 & \phi(X_n) - \phi(x) & \cdots & (\phi(X_n) - \phi(x))^p \end{pmatrix},$$

$$y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \gamma(x) = \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_p \end{pmatrix}, W_x = \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix}, s = \begin{pmatrix} \sigma(X_1)\varepsilon \\ \vdots \\ \sigma(X_n)\varepsilon \end{pmatrix}.$$

With this notation the local fit of (2.7) corresponds to the fit of $y = X_x\gamma(x) + s$ and the minimization problem (2.13) has the form

$$min_{\gamma(x)} (y - X_x\gamma(x))^T W_x (y - X_x\gamma(x)),$$

yielding $\hat{\gamma}(x) = (X_x^T W_x X_x)^{-1} X_x^T W_x y$, just as in the case of local polynomial fitting. Then $\hat{m}(x) = e_1^T \hat{\gamma}(x)$, where $e_1 = (1, 0, \ldots, 0)^T$, is an estimator for the underlying function $m(\cdot)$ at point $x$. Using (2.10) to (2.12), estimators for the derivatives can be obtained in a similar way. Then

$$Bias(\hat{\gamma}(x)|\mathbb{X}) = (X_x^T W_x X_x)^{-1} X_x^T W_x r_x, \tag{2.14}$$

holds, where $r_x = (m(X_1), \ldots, m(X_n))^T - X_x\gamma(x)$ is the vector of the residuals of the local fit and $\mathbb{X}$ denotes the vector $(X_1, \ldots, X_n)$. The conditional covariance matrix is given by

$$Var(\hat{\gamma}(x)|\mathbb{X}) = (X_x^T W_x X_x)^{-1} (X_x^T \Sigma_x X_x)(X_x^T W_x X_x)^{-1}, \tag{2.15}$$

where $\Sigma_x = diag(w_i^2 \sigma^2(X_i))$.

### 2.1.3 Asymptotics

Usually formulas (2.14) and (2.15) cannot be used in practice, since they depend on the unknown quantities $r_x$ and $\Sigma_x$. Consequently an asymptotic derivation is required. We will use the notations

$$\mu_j = \int_{-\infty}^{\infty} u^j K(u)\,du \quad \text{and} \quad \nu_j = \int_{-\infty}^{\infty} u^j K^2(u)\,du$$

for the $j^{\text{th}}$ moments of the kernels $K$ and $K^2$. Note that $\mu_0 = 1$ and (for symmetric kernels, which we apply here) $\mu_{2k+1} = \nu_{2k+1} = 0$ for all $k \in \mathbb{N}_0$ . Further we define the kernel moment matrices

$$\begin{array}{ll}
S = (\mu_{j+l})_{0\le j,l\le p} & c_p = (\mu_{p+1},\ldots,\mu_{2p+1})^T \\
\tilde{S} = (\mu_{j+l+1})_{0\le j,l\le p} & \tilde{c}_p = (\mu_{p+2},\ldots,\mu_{2p+2})^T \\
\bar{S} = ((j+l)\mu_{j+l+1})_{0\le j,l\le p} & \bar{c}_p = ((p+1)\mu_{p+2},\ldots,(2p+1)\mu_{2p+2})^T \\
S^* = (\nu_{j+l})_{0\le j,l\le p}\,. &
\end{array}$$

Finally we introduce the denotation $\varphi(x) = \phi'(x)$ and the matrices $H = \operatorname{diag}(h^j)_{0\le j\le p}$ and $P_x = \operatorname{diag}(\varphi^j(x))_{0\le j\le p}$ and recall that $e_{j+1} = (0,\ldots,0,1,0,\ldots,0)^T$ with 1 at $(j+1)^{\text{th}}$ position. $o_P(1)$ denotes a sequence of random variables which tends to zero in probability.

**Theorem 2.2.**
*Assume that $f(x) > 0$, $\sigma^2(x) > 0$, $\varphi(x) \neq 0$ and that $f(\cdot)$, $m^{(p+1)}(\cdot)$, $\phi^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of $x$. Further assume that $h \longrightarrow 0$ and $nh \longrightarrow \infty$. Then the asymptotic conditional covariance matrix of $\hat{\gamma}(x)$ is given by*

$$Var(\hat{\gamma}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nhf(x)} P_x^{-1} H^{-1} S^{-1} S^* S^{-1} H^{-1} P_x^{-1}(1 + o_P(1)). \tag{2.16}$$

*The asymptotic conditional bias is given by*

$$Bias(\hat{\gamma}(x)|\mathbb{X}) = h^{p+1}\varphi^{p+1}(x) P_x^{-1} H^{-1}\left(\gamma_{p+1} S^{-1} c_p + b_n)\right), \tag{2.17}$$

*where $b_n = o_P(1)$. If in addition $f'(\cdot)$, $m^{(p+2)}(\cdot)$ and $\phi^{(p+2)}(\cdot)$ are continuous in a neighborhood of $x$ and $nh^3 \longrightarrow \infty$, the sequence $b_n$ can be written as*

$$\begin{aligned} b_n \;=\; h\Bigg[&\left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\varphi(x)\right) S^{-1}\tilde{c}_p + \\ &\gamma_{p+1}\frac{\varphi'(x)}{2\varphi(x)} S^{-1}\bar{c}_p - \gamma_{p+1} S^{-1}\left(\frac{f'(x)}{f(x)}\tilde{S} - \frac{\varphi'(x)}{2\varphi(x)}\bar{S}\right) S^{-1} c_p + o_P(1)\Bigg]. \end{aligned} \tag{2.18}$$

Based on Theorem 2.2 and formulas (2.9) to (2.12) asymptotic expressions for bias and variance of the mean function and its derivatives can be derived. In particular we obtain for the variance

$$\begin{aligned} Var(\hat{m}(x)|\mathbb{X}) &= Var(e_1^T\hat{\gamma}(x)|\mathbb{X}) \\ &= \frac{\sigma^2(x)}{nhf(x)} e_1^T S^{-1} S^* S^{-1} e_1 (1 + o_P(1)) \end{aligned}$$

and

$$\begin{aligned} Var(\hat{m}'(x)|\mathbb{X}) &= Var(\varphi(x) e_2^T\hat{\gamma}(x)|\mathbb{X}) \\ &= \frac{\sigma^2(x)}{nh^3 f(x)} e_2^T S^{-1} S^* S^{-1} e_2 (1 + o_P(1)). \end{aligned}$$

Thus for $j = 0, 1$ the asymptotic variance of $\hat{m}^{(j)}(x)$ does not depend on the basis function! Now we take a look at the bias. Using (2.17) and (2.9) we arrive at

$$\begin{aligned} Bias(\hat{m}(x)|\mathbb{X}) &= bias(e_1^T\hat{\gamma}(x)|\mathbb{X}) \\ &= h^{p+1}\varphi^{p+1}(x) e_1^T \left( \frac{\psi_{(p+1)}(x)}{(p+1)!} S^{-1} c_p + b_n \right) \end{aligned} \tag{2.19}$$

and

$$\begin{aligned} Bias(\hat{m}'(x)|\mathbb{X}) &= bias(\varphi(x) e_2^T\hat{\gamma}(x)|\mathbb{X}) \\ &= h^{p}\varphi^{p+1}(x) e_2^T \left( \frac{\psi_{(p+1)}(x)}{(p+1)!} S^{-1} c_p + b_n \right). \end{aligned}$$

It is important to know that the product $e_{j+1} S^{-1} c_p$ is zero for $p - j$ even. Thus in the case $p - j$ odd, it is sufficient to work with $b_n = o_P(1)$. However, if $p - j$ takes an even value, the refined formula (2.18) for $b_n$ has to be chosen.

**Remark:** Local polynomial fitting

Since local fitting based on a power basis is a generalization of local polynomial fitting, setting $\phi(x) = x$ should correspond to the formulas given by Fan & Gijbels (1996), Theorem 3.1. Using that for local polynomial fitting $P_x = I$, $\varphi = 1$ and $\gamma_j = \beta_j$ holds, we actually find with equation (2.3)

$$\begin{aligned} Var(\hat{m}^{(j)}(x)|\mathbb{X}) &= e_{j+1}^T S^{-1} S^* S^{-1} e_{j+1} \frac{(j!)^2\sigma^2(x)}{f(x) nh^{1+2j}} \\ &\quad + o_P\left(\frac{1}{nh^{1+2j}}\right) \end{aligned}$$

and

$$\begin{aligned} Bias(\hat{m}^{(j)}(x)|\mathbb{X}) &= e_{j+1}^T S^{-1} c_p \frac{j!}{(p+1)!} m^{(p+1)}(x) h^{p+1-j} \\ &\quad + o_P(h^{p+1-j}). \end{aligned}$$

However, in the case $p-j$ even, when the deeper derivation is required, we obtain via (2.17) and (2.18)

$$\begin{aligned} Bias(\hat{m}^{(j)}(x)|\mathbb{X}) &= j!h^{p+2-j}e_{j+1}^T\left[S^{-1}\tilde{c}_p\left(\beta_{p+1}\frac{f'(x)}{f(x)}+\beta_{p+2}\right)\right. \\ &\quad \left. -\frac{f'(x)}{f(x)}\beta_{p+1}S^{-1}\tilde{S}S^{-1}c_p\right]+o_P(h^{p+2-j}), \end{aligned}$$

which is not the same as formula (3.9) given in Fan & Gijbels (1996). This difference arises because there on page 103 it is claimed that the $(j+1)^{\text{th}}$ element of $S^{-1}\tilde{S}S^{-1}c_p$ is zero for $p-j$ even, which seems to be true for $j=0$, but not in general (consider for e.g. $p=1$ and $j=1$, then $e_{j+1}S^{-1}\tilde{S}S^{-1}c_p = \mu_2$). The correct formula for general values of $p-j$ was provided in Fan, Gijbels, Hu & Huang (1996).

### 2.1.4 A simulated example

Here and in the following sections, we use the relative squared error as a measure for the error when estimating the target function:

$$\text{RSE}(\hat{m}) = \frac{\|\hat{m}-m\|}{\|m\|} = \frac{\sqrt{\sum_{i=1}^n (m(X_i)-\hat{m}(X_i))^2}}{\sqrt{\sum_{i=1}^n m(X_i)^2}}. \tag{2.20}$$

As an example, we consider the underlying function

$$m(x) = x + \frac{1}{1.2\sqrt{2\pi}}e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}}e^{-(x-0.7)^2/0.0018}, \tag{2.21}$$

which we contaminate with Gaussian noise with $\sigma = 0.05$. The 41 data points and the function are shown in Figure 2.1.

The following table shows the "best" values of the relative errors for various basis functions $\phi(\cdot)$ and degrees $p$. "Best" means that we choose the bandwidth $h_{emp}$ by

$$h_{emp} = min_h \quad \text{RSE}(\hat{m})$$

and calculate the relative error for the corresponding value. In order to avoid boundary effects, we omitted the first and last value in the calculation of the relative error.

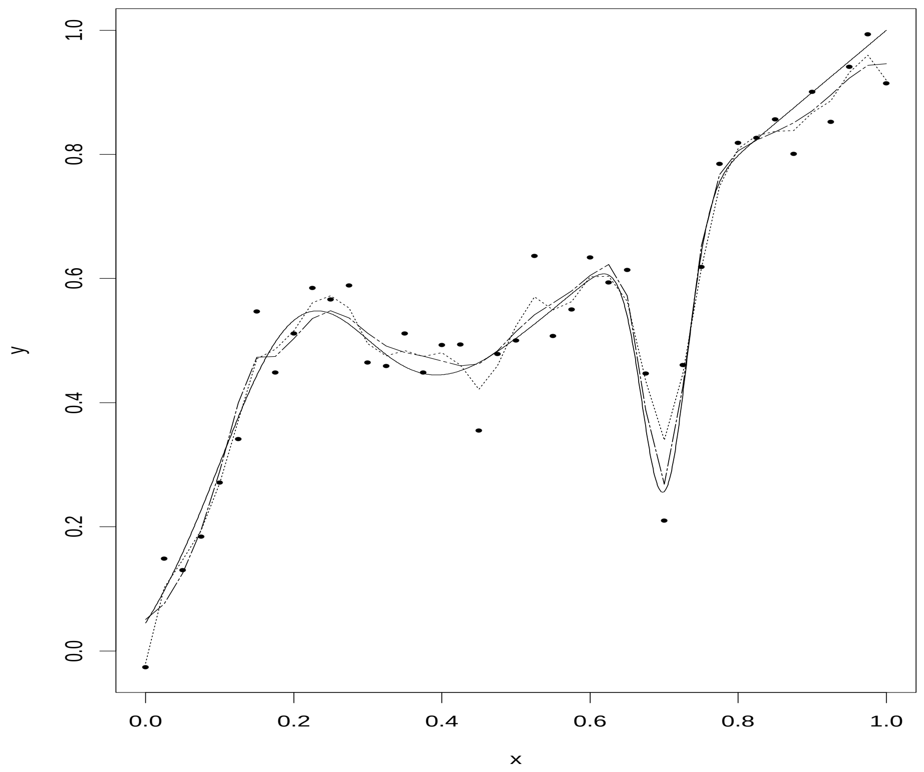



Figure 2.1: Function (2.21) (solid line), contaminated data and estimated curve with $\phi(x) = x$ (dotted line) and $\phi(x) = \tilde{m}(x)$, (dashed line, see Section 2.1.5).

| $\phi$ | $p = 1$ | $p = 2$ | $p = 3$ | $p = 8$ |
|---|---|---|---|---|
| $x$ | 0.04878 | 0.04597 | 0.04636 | 0.05116 |
| $\sin x$ | 0.04862 $*$ | 0.04649 | 0.04563 $*$ | 0.05253 |
| $\arctan x$ | 0.04861 $**$ | 0.04656 | 0.04560 $**$ | 0.05229 |
| $\operatorname{arcsinh} x$ | 0.04870 $*$ | 0.04625 | 0.04599 $*$ | 0.05038 $*$ |
| $x^2$ | 0.04996 | 0.04572 $*$ | 0.04920 | 0.04898 $*$ |
| $\cos x$ | 0.04991 | 0.04581 $*$ | 0.04903 | 0.04836 $*$ |
| $\cosh x$ | 0.05000 | 0.04564 $*$ | 0.04935 | 0.04933 $*$ |
| dnorm $x$ | 0.04982 | 0.04601 | 0.04869 | 0.04716 $**$ |
| $\exp x$ | 0.04903 | 0.04545 $***$ | 0.04710 | 0.05014 $*$ |
| $\log(x+1)$ | 0.04863 $*$ | 0.04631 | 0.04600 $*$ | 0.05170 |
| $\sqrt{x}$ | 0.04870 $*$ | 0.04599 | 0.04772 | 0.05011 $*$ |

Table 2.1: Relative squared errors for various basis functions and degrees.

In Table 2.1 dnorm$(x)$ denotes the density of the standard normal distribution. We have inserted one star (*) behind the RSE if the value is better than that for local polynomial fitting, two stars for the winner of the column, and three stars for the over-all-winner. The first four basis functions (included the linear function) are odd functions, followed by four even functions and three more functions without any symmetric properties. The table is easy to read - it shows that, at least in this example, odd basis functions perform better for odd values of $p$ and even functions perform better for even values of $p$. The non-symmetric functions show inhomogeneous behaviour: The exponentional function seems to behave like an even function and provides the over-all-winner, while the logarithmic basis performs like an odd basis. Except for large degrees of $p$, the Gaussian kernels turn out to be not a very good basis.

We did a variety of simulations with other underlying functions. The pattern could always be repeated for odd $p$, while for even values of $p$ the results differ from case to case. In all simulated examples we obtained this block pattern, i.e. for a given $p$, the group of odd basis functions generally behaves differently than the group of even basis functions. It is already well known that odd order local fits should be preferred to even order local fits (Fan & Gijbels, 1996, page 79). We add the observation that, if an odd value of $p$ is used, an odd basis should be preferred to an even basis. For even values of $p$ the results were too inhomogeneous to give a general statement.

Regarding special basis functions, in almost all simulations the best choice for odd values of $p$ was the arctan function. In addition, the computation was mostly faster and more stable than for a polynomial basis. The reason for this phenomenon is probably that the design matrix is less asymmetric for the arctan than for the linear basis. The good performance of the exp function falls outside of the tendency observed here, but can be explained by the results of the following section.

### 2.1.5 Bias reduction with data-adaptive basis functions

The results in the previous section only give a very weak inducement to change from the polynomial to another basis function. Thus, further work is required to find the optimal basis function, which will be done in this section.

Regarding formula (2.19) in the special case $p = 1$ leads to

$$Bias(\hat{m}(x)|\mathbb{X}) = \frac{h^2\mu_2}{2}\left(m''(x) - \frac{\varphi'(x)}{\varphi(x)}m'(x)\right) + o_P(h^2), \tag{2.22}$$

which reduces to the well-known formula

$$Bias(\hat{m}(x)|\mathbb{X}) = \frac{h^2\mu_2}{2}m''(x) + o_P(h^2),$$

in the special case of local linear fitting. Thus the subtraction of $\frac{\varphi'(x)}{\varphi(x)}m'(x)$ in (2.22) provides the chance to bias reduction (given that $\varphi(x) \neq 0$). Note that expression

(2.22) is independent of the design density $f$, and thus design-adaptivity, which is one of the major advantages of local linear smoothers, is retained when using a general basis. In the optimal case, the content of the bracket in (2.22) is zero. Hence the differential equation

$$m''(x)\varphi(x) - m'(x)\varphi'(x) = 0$$

has to be solved, leading to the solutions

$$\varphi(x) = c_1 m'(x) \quad (c_1 \in \mathbb{R})$$

and hence

$$\phi(x) = c_1 m(x) + c_2 \quad (c_1, c_2 \in \mathbb{R}).$$

In particular, for $c_1 = 1, c_2 = 0$ we get $\phi_{opt}(x) = m(x)$, and thus the underlying function $m(\cdot)$ is the optimal basis function. Although the function $m(\cdot)$ is always unknown, there are several ways to use this result. For example, one can calculate a pilot estimate $\check{m}(\cdot)$ by performing a local linear fit (or any other smooth estimate, e.g. with splines), and then use the estimated function as an improved basis. Or perhaps, in another situation, one may have a notion of the underlying function or know it only partly. For example, let us assume somebody gave us (wrong) information about the underlying function (2.21), namely

$$\tilde{m}(x) = x - \frac{1}{1.2\sqrt{2\pi}} e^{-(x-0.2)^2/0.02} - \frac{1}{0.9\sqrt{2\pi}} e^{-(x-0.7)^2/0.0018},$$

i.e. the first hump points down instead of up. Then one can try to use this function as basis function. Applying these approaches to the data set of the previous section leads to the following table. We tried two different bandwidths for the pilot estimate $\check{m}(x)$, once using the best bandwidth $h_{emp} = 0.020$, leading to the estimated function $\check{m}_{20}(x)$, and once for a higher bandwidth $h = 0.038$, resulting in the fit $\check{m}_{38}(x)$. For comparison, we added the results for the linear basis $\phi(x) = x$ and the optimal basis $\phi(x) = m(x)$.

| $\phi$ | $p = 1$ | $p = 2$ | $p = 3$ | $p = 8$ |
|---|---|---|---|---|
| $x$ | 0.04878 | 0.04597 | 0.04636 | 0.05116 |
| $\check{m}_{20}(x)$ | 0.04407 $*$ | 0.04310 $*$ | 0.04499 $*$ | 0.05237 |
| $\check{m}_{38}(x)$ | 0.03811 $*$ | 0.03656 $**$ | 0.03730 $**$ | 0.05202 |
| $\tilde{m}(x)$ | 0.03380 $***$ | 0.03998 $*$ | 0.04628 $*$ | 0.05593 |
| $m(x)$ | 0.00856 | 0.02587 | 0.02600 | 0.03719 |

Table 2.2: Relative squared errors for improved basis functions.

For illustration, we did 50 simulations of the contaminated function (2.21) and calculated the best relative errors for the linear, the arctan and the cosh basis as well as for $\tilde{m}(x)$ and $\check{m}(x)$ for $p = 1$. In contrast to the original example in Section 2.1.4, where the design was fixed, we now worked with random design, which was simulated anew each time. The corresponding relative errors, plotted in boxplots, are shown in Figure 2.2. The boxplots show that the linear basis yields results similar to the arctan basis, since both are odd basis functions, whereas the even cosh basis shows a lesser performance. The performance of the estimation can be improved significantly by using either a data-adaptive basis $\check{m}(x)$ (taking an optimized pilot bandwidth in each run), or the "guessed" basis $\tilde{m}(x)$. The application of the (in practice unavailable) optimal basis $m(x)$ for $p = 1$ leads to a nearly perfect fit. For higher degrees, results get worse as the basis was optimized for $p = 1$.

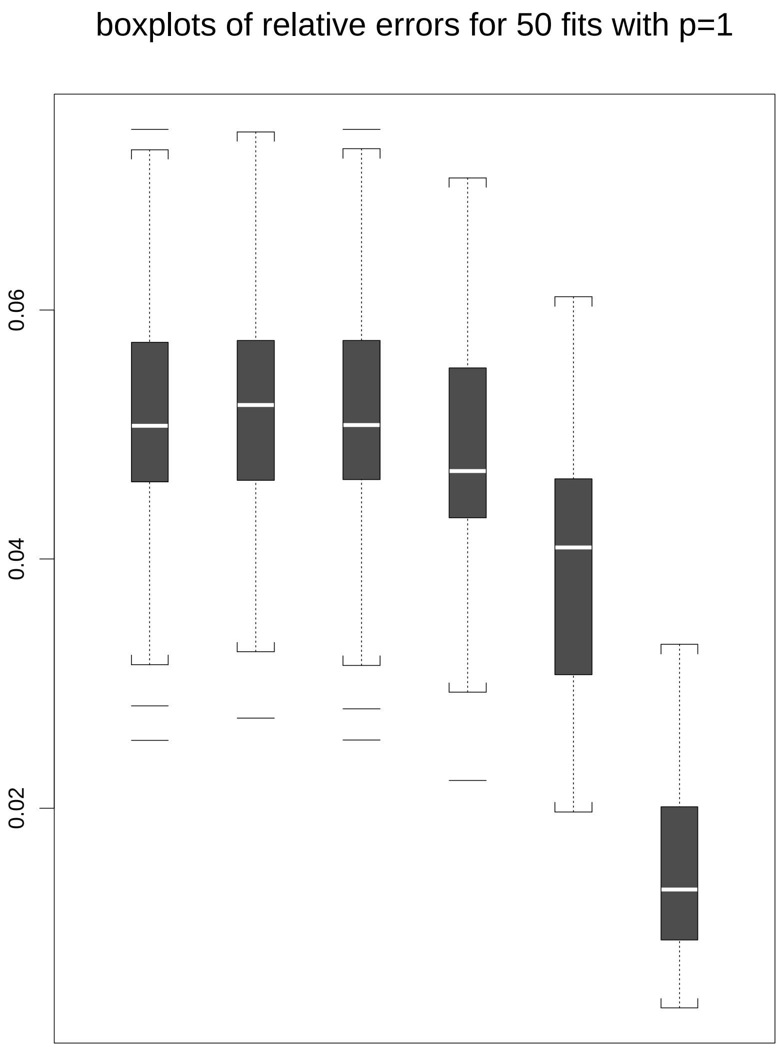



Figure 2.2: Boxplots of the relative errors using the basis functions $\phi(x) = x, \cosh(x), \arctan(x), \check{m}(x), \tilde{m}(x), m(x)$ (from left to right) with $p = 1$.

What does a basis function actually effect? For a given basis, the smoothing step in fact balances between the information given by the basis and the data. A similar concept is well-known from Bayesian statistics. Though the Bayesian prior does not contain a basis function but an assumption about the distribution of unknown parameters, the principle, boldly compared, is the same, since the *"posterior [distribution] is a trade-off between information in the data and prior knowledge"* (Raßer, 2003). Thus, in a way, one might name our basis function a "prior" basis for a "posterior" estimate of the regression function.

### 2.1.6 Notes about bandwidth selection

For bandwidth selection, one has the general choice between classical methods and plug-in methods. For an overview of bandwidth selection routines, we refer to Fan & Gijbels (1996), p. 110 ff. Classical methods as cross-validation or the AIC criterion can be applied directly to fitting with general basis functions. Promising extensions of CV and AIC have been given by Hart & Yi (1998) and Hurvich, Simonoff & Tsai (1998), respectively. Classical approaches compete with plug-in methods, as treated by Fan & Gijbels (1995), Ruppert, Sheather & Wand (1995) and Doksum, Petersen & Samarov (2000), to name a few. For a comparison of classical and plug-in methods see Härdle, Hall & Marron (1988) and Loader (1999a). Plug-in estimators perform a pilot estimate in order to estimate the asymptotic mean squared error, which is then minimized in terms of the bandwidth. Each plug-in-estimator is designed exclusively for a special smoothing method, so that application of these estimators for general basis functions requires some extra work.

Using Theorem 2.2, plug-in formulas for bandwidth selection can be straightforwardly derived by extending the corresponding methods for local polynomial fitting. We will not provide a complete examination of this topic now, but only give some impressions of the results. Let us therefore consider the case that one is interested in deriving the asymptotically optimal variable bandwidth $h_{opt}(x)$, which varies with the target value $x$. Minimizing the asymptotic mean squared error $MSE(\hat{m}(x)|\mathbb{X}) = Bias^2(\hat{m}(x)|\mathbb{X}) + Var(\hat{m}(x)|\mathbb{X})$, whereby (2.17) and (2.16) are respectively employed for the bias and variance, we arrive at an asymptotically optimal bandwidth

$$h_{opt}^{(\varphi)}(x) = C_{0,p}(K) \left[ \frac{\sigma^2(x)}{\psi_{(p+1)}^2(x) f(x) \varphi^{2p+2}(x)} \right]^{\frac{1}{2p+3}} \cdot n^{-\frac{1}{2p+3}},$$

where the constant $C_{0,p}$, which depends only on $p$ and the kernel $K$, is the same as in Fan & Gijbels (1996), page 67. Setting this result for $p = 1$ in relation to the optimal bandwidth $h_{opt}^{LL}(x)$ for a local linear fit, we obtain

$$\frac{h_{opt}^{(\varphi)}(x)}{h_{opt}^{LL}(x)} = \left[1 - \frac{\varphi'(x)}{\varphi(x)} \cdot \frac{m'(x)}{m''(x)}\right]^{-2/5}.$$

Note that this expression tends to infinity if $\phi(x)$ approximates $m(x)$. This conforms to the observations that can be drawn from Figure 2.3.

Bandwidth selection is especially difficult for data-adaptive basis functions as presented in the previous section: In this case one needs two bandwidths, one for the first and one for the second fit. This problem is so far not entirely solved. However, we can give some general statements concerning this. If the optimal bandwidth is used for the first fit, then bandwidth selection is not very crucial in the second fit, since this fit is more a correction of the first fit than a localization. In this case the optimal (second) bandwidths are very high (in our example $h_{emp} = 0.190$ for $p = 1$) and the minimas are very flat. Hence every large bandwidth will do a good job. This is illustrated in Figure 2.3. However, it is not necessary to find the optimal bandwidth in the first fit. Our simulations showed that the results can even be improved when the optimal bandwidth is *not* met, i.e. somewhat higher bandwidths are used. This is seen in Table 2.2 and Figure 2.3. From our experience, working with about the double optimal bandwidth (of a local linear fit) for both fits leads to satisfactory results.

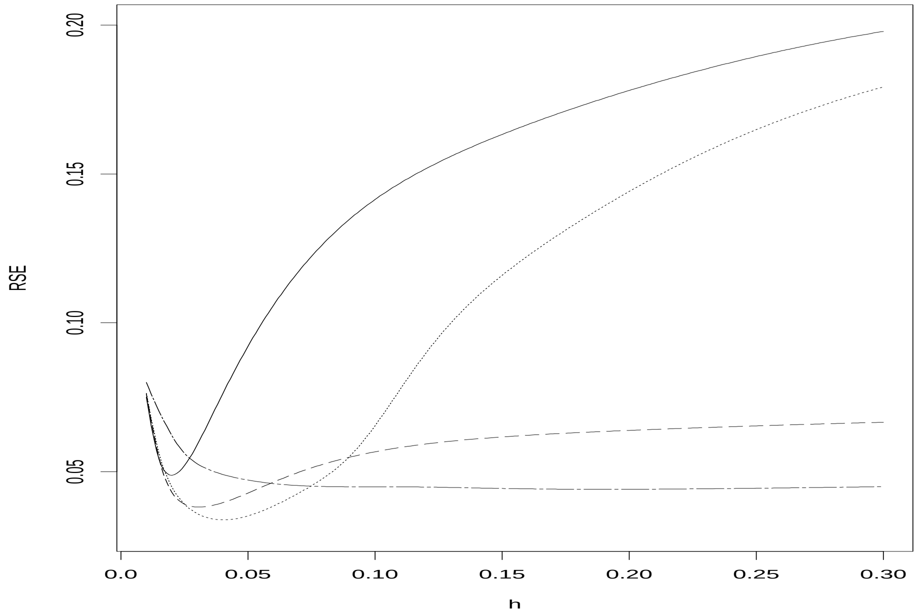



Figure 2.3: Relative squared errors as functions of the bandwidth for the linear basis (solid line), the data-adaptive basis $\check{m}_{20}(x)$ (dashed-dotted line) and $\check{m}_{38}(x)$ (dashed line), and for the guessed basis $\tilde{m}(x)$ (dotted line).

### 2.1.7 A real data example

In this section we consider the motorcycle data, firstly provided by Schmidt, Mattern & Schüler (1981), which has been widely used in the smoothing literature to demonstrate the performance of nonparametric smoothing methods, as e.g. in Fan & Gijbels (1996), page 2. The data was collected by performing crash tests with dummies sitting on motorcycles. The head acceleration of the dummies (in g) was recorded at a certain time (measured in milliseconds) after they had hit a wall. Figure 2.4 shows the motorcycle data with a local linear fit (solid line), identical to the fitted curve shown in Fan & Gijbels (1996), page 5. The bandwidth value $h = 1.48$ is obtained by cross-validation. There might be something wrong about the value $h = 3.3$ given by Fan & Gijbels - maybe it is measured on another scale.

Strictly considered, the observations are not independent, since several measurements were taken from every dummy at different time points, so that a mixed model (Wand, 2003) might be more appropriate. Though Opsomer, Wang & Yang (2001) show that correlation may seriously affect bandwidth selection, this problem is mostly ignored for the motorcycle data. Probably the correlation is negligible in this case, since the dummies are identical and do not characteristically behave as different persons.

For calculation of the pre-fit, the bandwidth $h = 2.7$ was applied (dotted line in Fig. 2.4). The long-dashed line is the local fit obtained using the pre-fit as basis function, applying the same bandwidth $h = 2.7$. For comparison, we provide also the result of a fit with smoothing splines. With real data it is hard to judge which fit might be the best one - but at least we can say that the fit applying a local pre-fit basis represents the first hump better and is smoother in the outer right area than a local linear fit. The performance now seems comparable to a spline fit.

### 2.1.8 Outlook

In Section 2.1.1 we already mentioned that the most general model for fitting with an arbitrary basis is

$$Y = \alpha_0(x)\phi_0(X) + \alpha_1(x)\phi_1(X) + \ldots + \alpha_p(x)\phi_p(X) + \sigma(X)\varepsilon. \tag{2.23}$$

However, theoretical properties of this fit can only be analyzed insufficiently. Nevertheless, as will be shown in this section, it is worth investigating this approach, because the results are somewhat impressive. We analyze the same data set as in Sections 2.1.4 and 2.1.5. We choose $p = 2$ and use $\phi_0(x) = 1$ and two Gaussian kernels as basis functions $\phi_{1,2}(x)$. There are two parameters to be adjusted: The

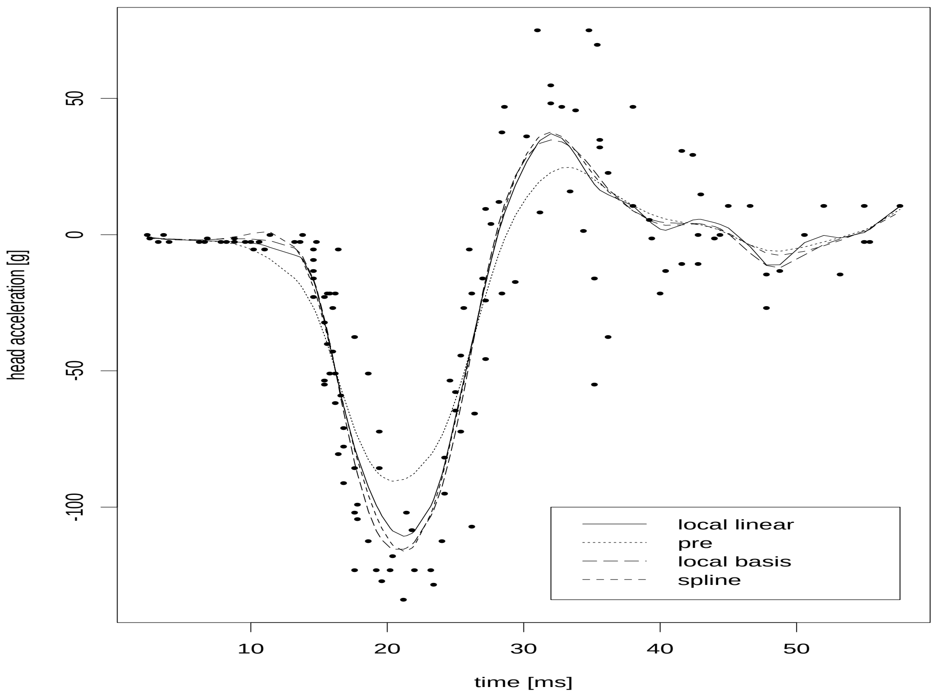



Figure 2.4: Motorcycle data with a local linear fit, a local pre-fit, a local fit using the latter fit as basis function, and a spline fit.

distance $d$ between the centers of the kernels and their width, i.e. the standard deviation $\sigma$. Simply taking the arbitrary values $\sigma = 0.4$ and $d = 0.35$ yields the relative error 0.04598 for the minimizing bandwidth $h_{emp} = 0.033$, similar to the result for the polynomials. However, the variation of $d$ and $\sigma$ leads to significant variations of the relative error, as shown in Figure 2.6 for the bandwidth $h = 0.033$ suggested above.

There is an absolute minimum at $\sigma = 0.04$ and $d = 0.39$, yielding a relative error of 0.03132. Using the best bandwidth for this basis function, $h_{emp} = 0.032$ leads to a relative error of 0.03128. Again, it is obvious that general basis functions can yield much better results than the polynomial basis. For comparison, see the two pictures in top of Figure 2.5. The fit with the Gaussian basis is smoother and closer to the underlying function, especially in the area of the cusps. The optimal basis functions, namely $\phi_1(x) = \frac{1}{0.04\sqrt{2\pi}} \exp(-(x-0.305)^2/(2 \cdot 0.04^2))$ and $\phi_2(x) = \frac{1}{0.04\sqrt{2\pi}} \exp(-(x-0.695)^2/(2 \cdot 0.04^2))$ are shown in the right bottom. An interesting observation is that the centers of the Gaussian kernels are situated near the humps of $m(\cdot)$, which is simply explicable because we showed in the previous section that the better the basis functions model the underlying function, the better are the results.

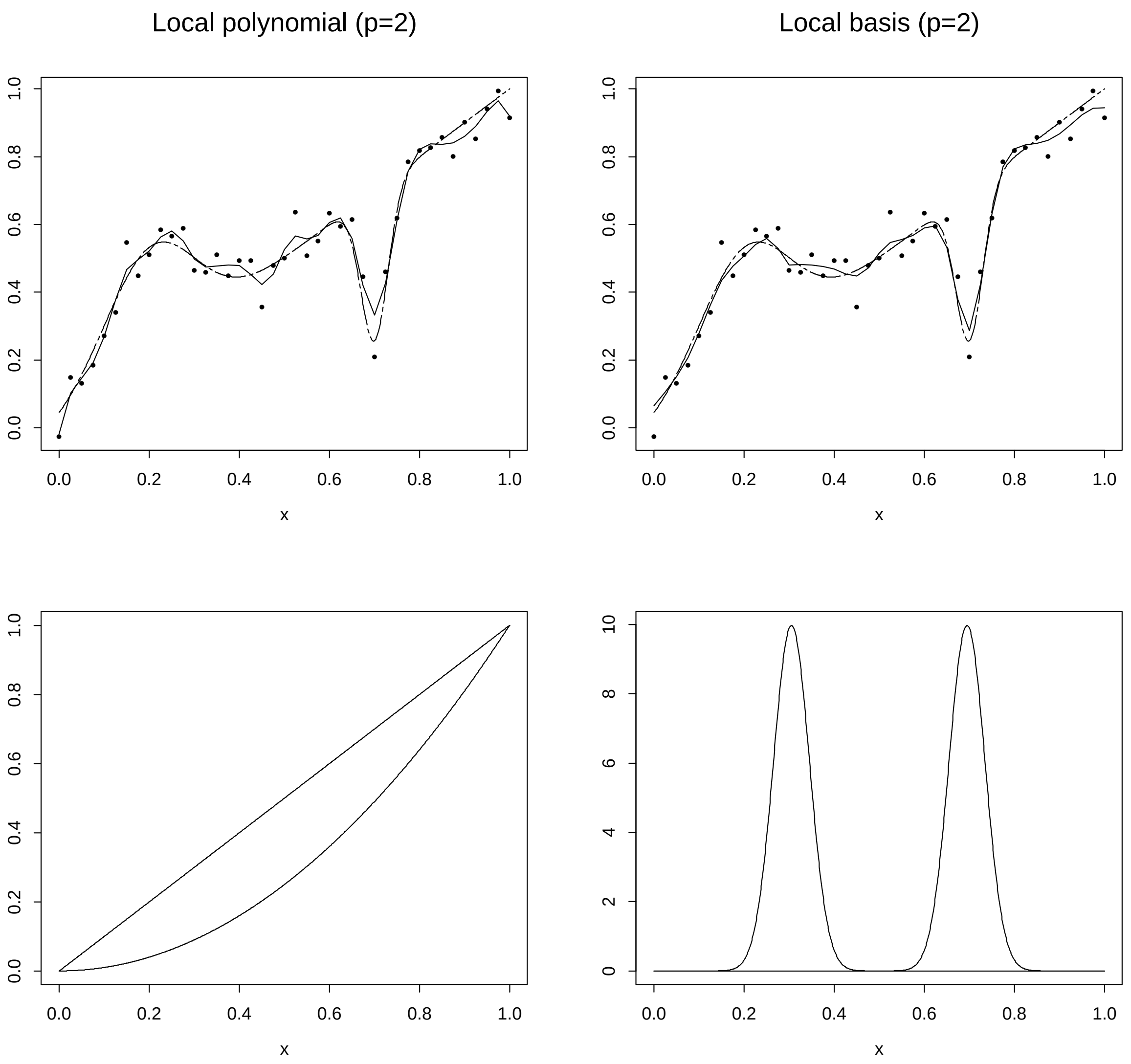



Figure 2.5: Top: Data, underlying function (dashed lines) and estimated functions (solid lines) for a polynomial (left) and a Gaussian basis (right); bottom: Basis functions $\phi_{1,2}(x)$ used for the corresponding fit.

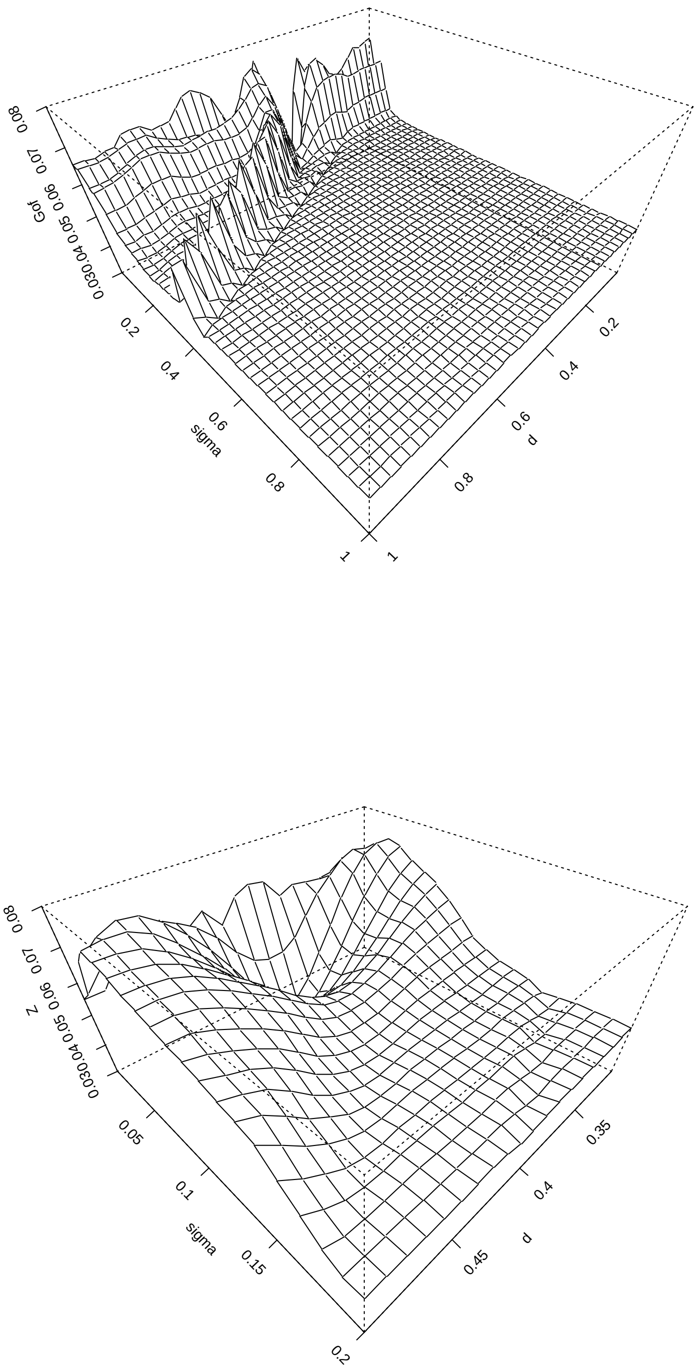



Figure 2.6: Top: Relative squared errors as function of $d$ and $\sigma$ (for an estimate of function (2.21) with $h = 0.033$); bottom: The interesting part in the area of the crater is shown in more detail.

Unfortunately, the minimum of the relative squared error in Figure 2.6 is impossible to find without knowledge of the underlying function. Nevertheless, it is an interesting observation that the structure of the mountain landscape is more or less similar for all underlying functions $m(\cdot)$. There always appears a large plain, yielding relative errors similar to those obtained by local polynomial fitting. In addition to the ridge at $\sigma = 0$, there often appears a diagonal ridge along the line $d = 2\sigma, d \gtrsim 0.4$, resulting from a singularity of the matrix $X_x^T W_x X_x$, and mostly some craters are situated along the line $d = 2\sigma, d \lesssim 0.4$. A large number of further craters often appear anywhere in or between the mountain ridges. Their locations result from a subtle interplay between the data set, the basis functions and the underlying function. It may be doubted that an analytical solution to this problem exists.

### 2.1.9 Appendix

**Proof of Theorem 2.1**

Define

$$\begin{aligned} g(y) &= m(z) - m(y) - \psi_{(1)}(y)(\phi(z) - \phi(y)) - \frac{\psi_{(2)}(y)}{2!}(\phi(z) - \phi(y))^2 \\ &- \ldots - \frac{\psi_{(p)}(y)}{p!}(\phi(z) - \phi(y))^p - \frac{M}{(p+1)!}(\phi(z) - \phi(y))^{p+1}, \end{aligned} \tag{2.24}$$

where $M \in \mathbb{R}$ is chosen so that $g(x) = 0$ is fulfilled.

Using $g(x) = g(z) = 0$ Rolle's theorem (see Lay (1990), p. 196) yields that there exists a $\zeta \in (x, z)$ resp. $(z, x)$ with $g'(\zeta) = 0$. Since

$$g'(y) = -\frac{\psi'_{(p)}(y)}{p!}(\phi(z) - \phi(y))^p - \frac{M}{p!}(\phi(z) - \phi(y))^p(-\phi'(y))$$

it follows that

$$0 = -\psi'_{(p)}(\zeta) + M\phi'(\zeta)$$

and thus $M = \psi_{(p+1)}(\zeta)$. The theorem is obtained by setting $y = x$ in (2.24).

**Proof of Theorem 2.2**

*I. Asymptotic conditional variance*

Recall that $O_P(1)$ denotes a sequence of random variables which is bounded in probability. Whenever there appears an integral, the borders $-\infty$ and $\infty$ are omitted. We denote further $S_{n,j} = \sum_{i=1}^n w_i(\phi(X_i) - \phi(x))^j$ and $S^*_{n,j} = \sum_{i=1}^n w_i^2\sigma^2(X_i)(\phi(X_i) - \phi(x))^j$. Then $S_n := (S_{n,j+l})_{0 \le j,l \le p} = X_x^T W_x X_x$ and $S_n^* := (S^*_{n,j+l})_{0 \le j,l \le p} = X_x^T \Sigma_x X_x$

hold, and the conditional variance (2.15) can be written as

$$Var(\hat{\gamma}(x)|\mathbb{X}) = S_n^{-1} S_n^* S_n^{-1} \tag{2.25}$$

and thus approximation of the matrizes $S_n$ and $S_n^*$ is required. Using that

$$\int K(u)u^j g(x+hu)\, du = \mu_j g_x + o(1) \tag{2.26}$$

for any function $g : \mathbb{R} \mapsto \mathbb{R}$ which is continuous in $x$, we obtain

$$\begin{aligned} ES_{n,j} &= n \int K(u)(\phi(x+hu) - \phi(x))^j f(x+hu)\, du = \\ &= nh^j \int K(u)u^j \varphi^j(\zeta_u) f(x+hu)\, du \\ &= nh^j (f(x)\varphi^j(x)\mu_j + o(1)) \end{aligned} \tag{2.27}$$

where $\zeta_u \in (x, x+hu)$ exists according to Taylor's theorem. Similar we derive

$$\begin{aligned} VarS_{n,j} &= nE(w_1^2(\phi(X_1) - \phi(x))^{2j}) - nE^2(w_1(\phi(X_1) - \phi(x))^j) \\ &= nh^{2j-1}(f(x)\varphi^{2j}(x)\nu_{2j} + o(1)) \\ &= n^2 h^{2j} O\left(\frac{1}{nh}\right) & (2.28) \\ &= o(n^2 h^{2j}). & (2.29) \end{aligned}$$

Since for every sequence $(Y_n)_{n\in\mathbb{N}}$ of random variables

$$Y_n = EY_n + O_P\left(\sqrt{VarY_n}\right) \tag{2.30}$$

holds (what can be proven with Chebychev's inequality), we can proceed with calculating

$$\begin{aligned} S_{n,j} &= ES_{n,j} + O_P\left(\sqrt{VarS_{n,j}}\right) \\ &= nh^j f(x)\varphi^j(x)\mu_j(1 + o_P(1)), \end{aligned} \tag{2.31}$$

which leads to

$$S_n = nf(x)P_x HSHP_x(1 + o_P(1)). \tag{2.32}$$

In the same manner, we find that

$$\begin{aligned} S^*_{n,j} &= ES^*_{n,j} + O_P\left(\sqrt{VarS^*_{n,j}}\right) \\ &= nh^{j-1}(\varphi^j(x)\sigma^2(x)f(x)\nu_j + o(1)) + O_P\left(\sqrt{o(n^2h^{2j-2})}\right) \\ &= nh^{j-1}\varphi^j(x)\sigma^2(x)f(x)\nu_j(1 + o_P(1)) \end{aligned}$$

and thus

$$S_n^* = \frac{n}{h} f(x)\sigma^2(x)P_x HS^* HP_x(1 + o_P(1)) \tag{2.33}$$

and finally assertion (2.16) by plugging (2.32) and (2.33) into (2.25).

*II. Asymptotic conditional bias*

Finding an asymptotic expression for

$$Bias(\hat{\gamma}|\mathbb{X}) = S_n^{-1} X_x^T W_x r_x \tag{2.34}$$

still requires to approximate $r_x \equiv (r_i)_{1\le i\le n}$. For all data points within the kernel support we obtain

$$\begin{aligned} r_i &= m(X_i) - \sum_{j=0}^{p} \gamma_j (\phi(X_i) - \phi(x))^j \\ &= \frac{\psi_{(p+1)}(\zeta_i)}{(p+1)!} (\phi(X_i) - \phi(x))^{p+1} \\ &= \gamma_{p+1}(x)(\phi(X_i) - \phi(x))^{p+1} + o_P(1) \frac{(\phi(X_i) - \phi(x))^{p+1}}{(p+1)!} \end{aligned}$$

where $\zeta_i \in (X_i, x)$ resp. $(x, X_i)$ exists according to Theorem 2.1. Note that the invertibility demanded for $\phi(\cdot)$ in Theorem 2.1 is already guaranteed locally around $x$ by the condition $\varphi(x) \neq 0$. Finally we calculate

$$\begin{aligned} Bias(\hat{\gamma}(x)|\mathbb{X}) &= S_n^{-1} X^T W_x \left[ (\phi(X_i) - \phi(x))^{p+1} (\gamma_{p+1} + o_P(1)) \right]_{1\le i\le n} \\ &= S_n^{-1} c_n (\gamma_{p+1} + o_P(1)) \\ &= P_x^{-1} H^{-1} S^{-1} H^{-1} P_x^{-1} \frac{1}{nf(x)} \left\{ \gamma_{p+1} c_n + \begin{pmatrix} o(nh^{p+1}) \\ \vdots \\ o(nh^{2p+1}) \end{pmatrix} \right\} (1 + o_P(1)) \\ &= P_x^{-1} H^{-1} S^{-1} h^{p+1} \varphi^{p+1}(x) \gamma_{p+1} c_p (1 + o_P(1)), \end{aligned}$$

by substituting the asymptotic expressions for $S_{n,j}$ (2.31) in $c_n := (S_{n,p+1}, \ldots, S_{n,2p+1})^T$, and thus (2.17) is proven.

Now we proceed to the derivation of $b_n$ which requires to take along some extra terms resulting from higher order expansions. With $(a + hb)^j = a^j + h(ja^{j-1}b + o(1))$ we find that

$$\begin{aligned} ES_{n,j} &= nh^j \int K(u) u^j \left( \varphi(x) + \frac{hu}{2} \varphi'(\zeta_u) \right)^j (f(x) + huf'(\xi_u)) du \\ &= nh^j \int K(u) u^j \left[ \varphi^j(x) + h \left( \frac{j}{2} \varphi^{j-1}(x) u \varphi'(\zeta_u) + o(1) \right) \right] (f(x) + huf'(\xi_u)) du \\ &= nh^j \left[ f(x) \varphi^j(x) \mu_j + h \left( f'(x) \varphi^j(x) + \frac{f(x)}{2} j \varphi^{j-1}(x) \varphi'(x) \right) \mu_{j+1} + o(h) \right] \end{aligned} \tag{2.35}$$

with $\zeta_u$ and $\xi_u$ according to Taylor's theorem. Plugging (2.35) and (2.28) into (2.30) yields

$$S_{n,j} = nh^j\varphi^j(x)\left[f(x)\mu_j + h\left(f'(x) + \frac{f(x)}{2}\frac{\varphi'(x)}{\varphi(x)}j\right)\mu_{j+1} + o_n\right], \tag{2.36}$$

where $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right)$, and further

$$S_n = nP_xH\left(f(x)S + hf'(x)\tilde{S} + h\frac{f(x)}{2}\frac{\varphi'(x)}{\varphi(x)}\bar{S} + o_n\right)HP_x. \tag{2.37}$$

The next task is to derive a higher order expansion for $r_x$. With Theorem 2.1 we obtain

$$\begin{aligned} r_i &= \frac{\psi_{(p+1)}(x)}{(p+1)!}(\phi(X_i)-\phi(x))^{p+1} + \frac{\psi_{(p+2)}(\zeta_i)}{(p+2)!}(\phi(X_i)-\phi(x))^{p+2} \\ &= \gamma_{p+1}(\phi(X_i)-\phi(x))^{p+1} + \gamma_{p+2}(\phi(X_i)-\phi(x))^{p+2} + \\ &\quad +(\psi_{(p+2)}(\zeta_i) - \psi_{(p+2)}(x))\frac{(\phi(X_i)-\phi(x))^{p+2}}{(p+2)!} \\ &= (\phi(X_i)-\phi(x))^{p+1}\gamma_{p+1} + (\phi(X_i)-\phi(x))^{p+2}(\gamma_{p+2} + o_P(1)) \end{aligned}$$

with $\zeta_i \in (X_i, x)$ resp. $(x, X_i)$. Plugging this and (2.37) into (2.34) and denoting

$$T_n := f(x)S + h\left(f'(x)\tilde{S} + \frac{f(x)}{2}\frac{\varphi'(x)}{\varphi(x)}\bar{S}\right) + o_n$$

leads to

$$\begin{aligned} Bias(\hat{\gamma}(x)|\mathbb{X}) &= \left[nP_xHT_nHP_x\right]^{-1}\left[c_n\gamma_{p+1} + \tilde{c}_n(\gamma_{p+2} + o_P(1))\right] \\ &= P_x^{-1}H^{-1}T_n^{-1}h^{p+1}\varphi^{p+1}(x)\Big[\gamma_{p+1}f(x)c_p + \\ &\quad +h(\gamma_{p+1}f'(x) + \gamma_{p+2}\varphi(x)f(x))\tilde{c}_p + \\ &\quad +h\gamma_{p+1}f(x)\frac{\varphi'(x)}{2\varphi(x)}\bar{c}_p + o_n\Big], \end{aligned} \tag{2.38}$$

where the asymptotic expressions (2.36) are substituted in $c_n$ and $\tilde{c}_n = (S_{n,p+2}, \ldots, S_{n,2p+2})^T$. The matrix $T_n$ still has to be inverted. Applying the formula

$$(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$$

yields

$$T_n^{-1} = \frac{1}{f(x)}S^{-1} - h\frac{1}{f(x)}S^{-1}\left(\frac{f'(x)}{f(x)}\tilde{S} - \frac{\varphi'(x)}{2\varphi(x)}\bar{S}\right)S^{-1} + o_n, \tag{2.39}$$

and we finally obtain

$$\begin{aligned} Bias(\hat{\gamma}(x)|\mathbb{X}) &= h^{p+1}\varphi^{p+1}(x)P_x^{-1}H^{-1}\Big\{\gamma_{p+1}S^{-1}c_p + \\ &\quad +h\left[\left(\gamma_{p+1}\frac{f'(x)}{f(x)} + \gamma_{p+2}\varphi(x)\right)S^{-1}\tilde{c}_p + \gamma_{p+1}\frac{\varphi'(x)}{2\varphi(x)}S^{-1}\bar{c}_p + \right. \\ &\quad \left. +\ \gamma_{p+1}S^{-1}\left(\frac{f'(x)}{f(x)}\tilde{S} - \frac{\varphi'(x)}{2\varphi(x)}\bar{S}\right)S^{-1}c_p\right] + o_n\Big\}. \end{aligned}$$

## 2.2 Multivariate predictors

### 2.2.1 Introduction

In the last decades nonparametric smoothing has been one of the most attended and challenging fields in statistics. A widely used concept is that of localizing, where only observations in a neighborhood of the target value are used for the estimation of the regression function.

Nadaraya (1964) and Watson (1964) developed one of the earliest local estimators by simply fitting locally a constant mean value to the data. Stone (1977) was among the first to replace the constant by a line, which reduced the bias of the fit significantly, as Fan (1992) shows. Cleveland (1979) did the next extension and fitted polynomials of arbitrary degree instead of a line. Surprisingly the next step, replacing the polynomial basis $1, x, \ldots, x^p$ by an arbitrary basis $\phi_0(x), \ldots, \phi_p(x)$, as suggested briefly in Ramsay & Silverman (1997), has never been further pursued.

Since local fitting is so far only performed with the polynomial basis, the question of what is special about this particular basis arises. The answer is simple. For this basis Taylor's theorem is available which enables us to interpret the estimated parameters and to calculate the error of the approximation. According to this theorem, whose univariate version was firstly discovered by Brook Taylor (1685-1731) and published 1715 in his book *Methodus incrementorum directa et inversa*, a function $m$ at point $x$ can be approximated by a linear combination of polynomials in a neighborhood of $x$.

Local fitting with general basis functions will require to find a new Taylor theorem for every basis one wants to use, if some theoretical background is desired. Though this is certainly not possible for every basis, extensions for special cases exist. In Section 2.1.2 we provided a Taylor theorem covering the case where polynomials are replaced by the powers $\phi(x), \phi^2(x), \ldots, \phi^p(x)$ of an invertible function $\phi$. The properties of local modelling with such a power basis were examined, and it was shown that by a suitable basis the results of local polynomial fitting can be significantly improved.

Recently, the general research interest has turned from univariate to multivariate smoothing. Cleveland & Devlin (1988) gave an introduction to multivariate locally weighted regression and showed that the concept is useful in practice. Further impacts on multivariate local modelling were made by Staniswalis, Messer & Finston (1993), treating kernel estimators for multivariate regression, Wand & Jones (1993), describing bivariate kernel density estimation, and Wand (1992), calculating asymptotic mean square errors for multivariate kernel estimators. Ruppert & Wand (1994) derived asymptotic expressions for bias and variance of the multivariate local linear and quadratic fit.

In Section 2.2.2 we introduce the concept of multivariate local fitting with general basis functions. However, a fully theoretical handling of this estimator is not possible since a Taylor theorem for general basis functions does not exist. In Section 2.2.3 we focus on basis functions without interactions. We derive a new Taylor theorem which covers this case and provide asymptotic expressions for bias and variance of the corresponding local estimator. We give an example for fitting with general basis functions by means of a simulated data set in Section 2.2.4 and provide a data-driven tool to obtain a suitable basis in Section 2.2.5. We finish with the discussion in Section 2.2.6.

### 2.2.2 Multivariate locally weighted regression using a general basis

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a set of i.i.d. random variables sampled from a population $(X, Y) \in \mathbb{R}^{d+1}$. Y is a scalar response variable and X a $\mathbb{R}^d$-valued predictor variable with density $f$ having support $\text{supp}(f) \in \mathbb{R}^d$. We want to estimate the regression function

$$m(x) = E(Y|X = x) \tag{2.40}$$

at a vector $x \in \text{supp}(f)$ nonparametrically, i.e. without assuming $m$ to belong to a parametric family of functions. A model fulfilling (2.40) is

$$Y_i = m(X_i) + \sigma(X_i)\epsilon, \tag{2.41}$$

where $\sigma^2(x) = Var(Y|X = x)$ is finite, $E(\epsilon) = \mathbf{0}$, $Var(\epsilon) = \mathbf{I}_d$ and $\epsilon$ independent of all $X_i, i = 1, \dots, n$. Let $\{\phi_j : \mathbb{R}^d \longrightarrow \mathbb{R}, j = 1, \dots, q\}$ a set of multivariate continuously differentiable basis functions, $\Phi(x) = (\phi_1(x), \dots, \phi_q(x))^T$ and $\alpha_{1q}(x) := (\alpha_1(x), \dots, \alpha_q(x))^T$.

The amount of smoothing is determined by a symmetric positive definite bandwidth matrix $H \in \mathbb{R}^{d,d}$. (Often instead of $H$ a nonsingular matrix $B \in \mathbb{R}^{d,d}$ is called bandwidth matrix, where $H$ and $B$ have the relationship $H = BB^T$). Let $K : \mathbb{R}^d \mapsto \mathbb{R}$ be a multivariate kernel function and $K_H(u) = |H|^{-1/2}K(H^{-1/2}u)$. For a detailed description of multivariate kernels and bandwidth matrices see Wand & Jones (1993). The estimator of the function $m(\cdot)$ at point $x$ is $\hat{\alpha}_0(x)$, where $\hat{\alpha}(x) = (\hat{\alpha}_0(x), \hat{\alpha}_{1q}^T(x))^T$ is the minimizer of

$$\sum_{i=1}^{n} \left\{Y_i - \alpha_0(x) - \alpha_{1q}^T(x)(\Phi(X_i) - \Phi(x))\right\}^2 K_H(X_i - x). \tag{2.42}$$

The constant $\Phi(x)$, which only transforms the parameters, is useful because it makes the computation faster and the asymptotic calculations more convenient. This approach covers a wide range of well-known estimators. If, for example, $q = d$, then with $\phi_j(z_1, \ldots, z_d) = z_j$, $j = 1, \ldots, d$ we get the multivariate local linear estimator.

With

$$X_x = \begin{pmatrix} 1 & \phi_1(X_1) - \phi_1(x) & \ldots & \phi_q(X_1) - \phi_q(x) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(X_n) - \phi_1(x) & \ldots & \phi_q(X_n) - \phi_q(x) \end{pmatrix},$$

$W_x = \text{diag}(K_H(X_1 - x), \ldots, K_H(X_n - x))$ and $y = (Y_1, \ldots, Y_n)^T$ the least squares problem (2.42) can be written as

$$min_{\alpha(x)}(y - X_x\alpha(x))^T W_x (y - X_x\alpha(x))$$

and has the solution

$$\hat{\alpha}(x) = (X_x^T W_x X_x)^{-1} X_x^T W_x y, \tag{2.43}$$

provided that the matrix $X_x^T W_x X_x$ is nonsingular. Thus we obtain

$$\hat{m}(x) = \hat{\alpha}_0(x) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x y,$$

where $e_1^T = (1, 0 \ldots, 0) \in \mathbb{R}^{q+1}$. Furthermore,

$$E(\hat{m}(x)|\mathbb{X}) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x m, \tag{2.44}$$

where $m = (m(X_1), \ldots, m(X_n))^T$ and $\mathbb{X} = (X_1, \ldots, X_n)$. Finally the conditional covariance matrix is given by

$$\text{Var}(\hat{m}(x)|\mathbb{X}) = e_1^T (X_x^T W_x X_x)^{-1} (X_x^T \Sigma_x X_x)(X_x^T W_x X_x)^{-1} e_1, \tag{2.45}$$

where $\Sigma_x = \text{diag}(K_H^2(X_i - x)\sigma^2(X_i))$.

In the following we will provide an asymptotic expression for the variance of the estimator $\hat{m}(x)$. We will treat interior as well as boundary points. Thereby we call a point $x \in \text{supp}(f)$ an interior point if $\{z : H^{-1/2}(x - z) \in \text{supp}(K)\} \subset \text{supp}(f)$; otherwise, $x$ will be called a boundary point. Let

$$\mathcal{D}_{x,H} = \{u : (x + H^{1/2}u) \in \text{supp}(f)\} \cap \text{supp}(K).$$

Then $\mathcal{D}_{x,H} = \text{supp}(K)$ if and only if $x$ is an interior point. Note that we consider $x$ as a fixed point in the case of an interior point, but as a sequence $x_n$ converging sufficiently rapidly to the boundary in the case of a boundary point, ensuring that

$x$ is a boundary point for all $n$ (see (A4)). Also let

$$\begin{aligned} M_x &= \int_{\mathcal{D}_{x,H}} \begin{pmatrix} 1 \\ u \end{pmatrix} \begin{pmatrix} 1 & u^T \end{pmatrix} K(u)\,du, \\ N_x &= \int_{\mathcal{D}_{x,H}} \begin{pmatrix} 1 \\ u \end{pmatrix} \begin{pmatrix} 1 & u^T \end{pmatrix} K^2(u)\,du, \\ D_x &= (\nabla\phi_1(x), \ldots, \nabla\phi_q(x)), \\ A_{D_x} &= \begin{pmatrix} 1 & \\ & D_x \end{pmatrix}. \end{aligned}$$

$\nabla$ denotes the gradient function $(\partial_1, \ldots, \partial_d)^T = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_d}\right)^T$. Let $\nu_0 = \int K^2(u)\,du$. $o_P(1)$ denotes a sequence of random variables which tends to zero in probability. The asymptotic variance of the estimator $\hat{m}(x)$ is provided by the following theorem:

**Theorem 2.3.**
*Let $x$ be a fixed element in the interior of supp($f$). Then under regularity conditions (A1) to (A3) and (A5)*

$$Var(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2}\nu_0(1 + o_P(1)) \tag{2.46}$$

*holds. Let further $x$ be a boundary point, i.e. $x = x_b + H^{1/2}c$, where $x_b$ is a point on the boundary of supp($f$) and $c$ is a fixed element of supp($K$). Then under conditions (A2) to (A5)*

$$\begin{aligned} &Var(\hat{m}(x)|\mathbb{X}) = \\ &= \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2}e_1^T(A_{D_x}^T M_x A_{D_x})^{-1}A_{D_x}^T N_x A_{D_x}(A_{D_x}^T M_x A_{D_x})^{-1}e_1 \cdot \\ &\quad \cdot(1 + o_P(1)). \end{aligned} \tag{2.47}$$

Surprisingly, the asymptotical conditional variance of $\hat{m}(x)$ for interior points doesn't depend on the basis function (compare Ruppert & Wand (1994), Theorem 2.1). Thus, with a suitable basis, one could reduce the bias without a rise of the variance. However: For general basis functions we can't compute the asymptotical bias, since a general Taylor theorem is missing. In the next section we will focus on a case where a Taylor theorem is available.

### 2.2.3 Asymptotics for basis functions without interactions

Multivariate locally weighted polynomial regression, described in Ruppert & Wand (1994), is based on the multivariate Taylor theorem, which we will extend in the following. Let $d > 0, p \geq 0$, $U \subset \mathbb{R}^d$ open and $U_j$ the projection of $U$ on the $j^{\text{th}}$ coordinate. We impose an invertible basis function $\phi_j \in C^{p+1}(U_j)$ separately on every single coordinate, i.e.

$$\phi_j : U_j \to \mathbb{R}, z_j, \mapsto \phi_j(z_j), j = 1, \ldots, d.$$

For convenience of notation we give the same names to the functions $\phi_j : U \to \mathbb{R}, (z_1, \ldots, z_d) \mapsto \phi_j(z_j)$ picking the $j^{\text{th}}$ coordinate. Taking the notation from the previous section, it is $\Phi(z_1, \ldots, z_d) = (\phi_1(z_1), \ldots, \phi_d(z_d))^T$, and the inverse function $\Phi^{-1} : \Phi(U) \to U$ is given by $\Phi^{-1}(z_1, \ldots, z_d) = (\phi_1^{-1}(z_1), \ldots, \phi_d^{-1}(z_d))^T$. The matrix $D_x$ reduces to $P_x = \text{diag}(\phi_j'(x_j))_{1 \leq j \leq d}$. The following theorem holds.

**Theorem 2.4 (Generalized multivariate Taylor expansion).**
*Assume $U \subset \mathbb{R}^d$ open, $p \geq 0$, $m \in C^{p+1}(U)$, $\Phi : U \to \mathbb{R}^d$ like above, further assume the points $x$, $z$ and their connection curve $C_S(x, z)$, given by the function $y_\Phi(t) = \Phi^{-1}[\Phi(x) + t(\Phi(z) - \Phi(x))]$, $t \in [0, 1]$, to be in $U$. Then there exists a point $\zeta \in C_S(x, z)$ with*

$$m(z) = m(x) + \sum_{j=1}^{p} \frac{1}{j!} \left[ ((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^j m \right](x) + S_{p+1}(z), \tag{2.48}$$

*where $\nabla_\Phi m(x) = P_x^{-1} \nabla m(x)$, and*

$$S_{p+1}(z) = \frac{1}{(p+1)!} \left[ ((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^{p+1} m \right](\zeta).$$

For a better understanding and application of this theorem, we set

$$N_m(x) = H_m(x) - P_x^{-1} P_x' \text{diag}(\nabla m(x)), \tag{2.49}$$

where $H_m(x)$ is the Hessian matrix of $m$ and $P_x' = \text{diag}(\phi_j''(x_j))_{1 \leq j \leq d}$ the derivative of $P_x$. Thus $N_m(x)$ equals the Hessian matrix at all entries out of the diagonal, while the diagonal values are modified proportionally to the gradient function of $m$. Now we can write the generalized Taylor expansion in the form

$$\begin{aligned} m(z) = m(x) \quad &+ \quad (\Phi(z) - \Phi(x))^T P_x^{-1} \nabla m(x) + \\ &+ \quad \frac{1}{2} (\Phi(z) - \Phi(x))^T P_x^{-1} N_m(x) P_x^{-1} (\Phi(z) - \Phi(x)) + S_3(z, x), \end{aligned} \tag{2.50}$$

which reduces to the usual Taylor theorem by setting $\Phi = \text{id}$.

We denote $x = (x_1, \ldots, x_d)$ and $X_i = (X_{i1}, \ldots, X_{id})$ and work from now on with the design matrix

$$X_x = \begin{pmatrix} 1 & \phi_1(X_{11}) - \phi_1(x_1) & \ldots & \phi_d(X_{1d}) - \phi_d(x_d) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(X_{n1}) - \phi_1(x_1) & \ldots & \phi_d(X_{nd}) - \phi_d(x_d) \end{pmatrix},$$

where the $\phi_j$ are continuously differentiable, but not necessarily invertible. All formulas given from (2.42) to (2.45) remain thereby unchanged. Next, we derive asymptotic expressions for bias and variance of $\hat{m}(x)$ at interior as well as boundary points. Let $\mu_2$ as in (A1) and $\nu_0 = \int K^2(u)\, du$. We have the following theorem:

**Theorem 2.5.**
*Let $x$ be a fixed point in the interior of supp$(f)$. Then under regularity conditions (A1) to (A3) and (A5)*

$$Bias(\hat{m}(x)|\mathbb{X}) = \frac{1}{2}\mu_2 tr\,(HN_m(x)) + o_P(tr(H)) \tag{2.51}$$

*and*

$$Var(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2}\nu_0(1 + o_P(1)) \tag{2.52}$$

*hold.*

Note that the formula for the conditional bias only differs from the corresponding formular for $\Phi(z) = z$ by using $N_m(x)$ instead of $H_m(x)$ (compare Ruppert & Wand (1994), Theorem 2.1). In the univariate case (2.51) reduces to

$$\text{Bias}(\hat{m}(x)|\mathbb{X}) = \frac{1}{2}\mu_2 h^2 \left(m''(x) - \frac{\phi''(x)}{\phi'(x)}m'(x)\right) + o_P(h^2). \tag{2.53}$$

This result gives a hint of how to profit by general basis functions: (2.53) is minimized for $\phi(x) = m(x)$, thus the bias is reduced if the basis function is as near as possible to the underlying function $m$. In Sections 2.2.4 and 2.2.5 we will demonstrate how we can take advantage out of this result.

Now we continue with the treatment of boundary points. The following theorem can be seen as an extension of Theorem 2.5 which covers the case that the odd-order moments of $K$ (see condition (A1)) do *not* vanish.

**Theorem 2.6.**
*Let $x_b$ be a point at the boundary of supp$(f)$, $x = x_b + H^{1/2}c$, where $c$ is a fixed element of supp$(K)$. Then under conditions (A2) to (A5)*

$$\begin{aligned} &Bias(\hat{m}(x)|\mathbb{X}) = \\ &= \frac{1}{2}e_1^T M_x^{-1}\int_{\mathcal{D}_{x,H}} \begin{pmatrix} 1 \\ u \end{pmatrix} K(u)\{u^T H^{1/2} N_m(x) H^{1/2} u\}\, du + o_P(tr(H)) \end{aligned} \tag{2.54}$$

*and*

$$Var(\hat{m}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2}\left(e_1^T M_x^{-1} N_x M_x^{-1} e_1 + o_P(1)\right). \tag{2.55}$$

Again the asymptotic bias only differs from the corresponding formula for local linear fitting by the modified Hessian matrix $N_m(x)$. The asymptotic conditional variance at the boundary turns out to be independent of the basis function and is identical to the corresponding formula for a linear basis (see Ruppert & Wand (1994), Theorem 2.2). Note that this result is not self-evident, since we showed in (2.47) that for arbitrary basis functions the asymptotic variance at the boundary is *not* independent of the basis.

Finally recall that in the beginning of the section we defined the function $\Phi(\cdot)$ to be invertible on $U$. Here $U$ is a neighborhood of $x$ which becomes arbitrarily small for large $n$, see condition (A3). Thus it is sufficient if the basis functions are *locally* invertible around the target value $x$, what is already guaranteed by (A5).

### 2.2.4 Example

In this example we contaminate the underlying function $m : [0,1]^2 \longrightarrow \mathbb{R}$,

$$m(x_1, x_2) = (1 - x_1)\sin(12x_2) + x_1^2 \cos(16x_1) \qquad (2.56)$$

with Gaussian noise ($\sigma = 0.25$). The $n = 961$ design points are uniformly distributed on $[0,1]^2$. The function without and with contamination is shown in Fig. 2.7. For assessing the quality of the fit we use the relative squared error

$$\mathrm{RSE}(\hat{m}) = \frac{\|\hat{m} - m\|}{\|m\|} = \frac{\sqrt{\sum_{i=1}^{n} (m(X_i) - \hat{m}(X_i))^2}}{\sqrt{\sum_{i=1}^{n} m(X_i)^2}}. \qquad (2.57)$$

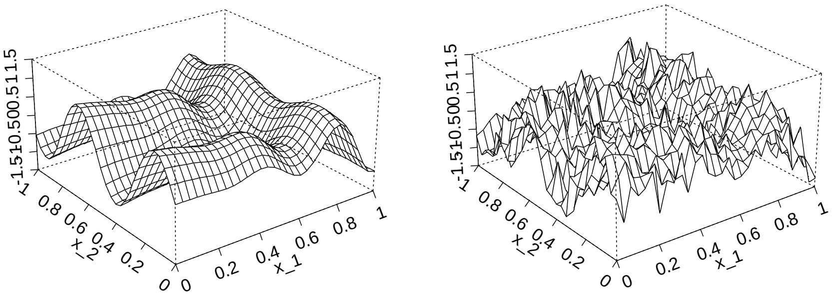



Figure 2.7: Function (2.56) without (left) and with contamination (right).

In Table 2.3 we compare the results of the local fit with various basis functions. For reasons of comparability we restrict on the case of two basis functions. From the first to the last line we will increase the amount of information which we install in the basis. Note that all basis functions fall in the general framework of Section 2.2.2, whereas only a) to e) fit the setting of Section 2.2.3. We provide the bandwidths $h_1$ and $h_2$ which minimize the RSE, and the value of RSE obtained at this minimizing bandwidth.

In g), we denote $g_1(\cdot) = \phi_{0.33,0.15^2}(\cdot)$ and $g_2(\cdot) = \phi_{0.67,0.15^2}(\cdot)$, where $\phi_{\mu,\sigma^2}(\cdot)$ is the density of a normal distribution with mean $\mu$ and variance $\sigma^2$. The bandwidth was restricted to a maximum value of 1. For a), d), f) and g) the results are illustrated in Fig. 2.8, where the plots of the basis functions and the corresponding fits are shown.

| | $\phi_1(x)$ | $\phi_2(x)$ | $h_1$ | $h_2$ | $RSE$ |
|---|---|---|---|---|---|
| $a)$ | $x_1$ | $x_2$ | 0.03 | 0.04 | 0.162 |
| $b)$ | $\cos 16x_1$ | $x_2$ | 0.14 | 0.03 | 0.132 |
| $c)$ | $x_1^2 \cos 16x_1$ | $x_2$ | 0.11 | 0.03 | 0.108 |
| $d)$ | $x_1$ | $\sin 12x_2$ | 0.03 | 1.00 | 0.080 |
| $e)$ | $\cos 16x_1$ | $\sin 12x_2$ | 0.04 | 1.00 | 0.059 |
| $f)$ | $x_1^2 \cos 16x_1$ | $(1 - x_1) \sin 12x_2$ | 1.00 | 1.00 | 0.011 |
| $g)$ | $g_1(x_1) \cdot g_2(x_2)$ | $g_2(x_1) \cdot g_1(x_2)$ | 0.04 | 0.03 | 0.164 |

Table 2.3: Relative squared errors for various basis functions.

The observations obtained from the table and the figures are the following:

- The more information the basis carries about the underlying function, the better the local fit and the higher the optimal bandwidths, see b) to f).
- If by accident a basis is used which doesn't contain any information about the underlying function, as in g), the results fortunately stay similar like for the linear basis. This result is simply explicable: The linear basis is a *wrong* basis. Mostly it does not contain any information about the true function. Thus replacing a wrong basis with another wrong basis will not make much difference.

Now we have the chance to use given information in an effective way. If one has any notion about the true function one can use this information in the basis. If the basis was more or less correct the fit can be improved tremendously.

### 2.2.5 Finding a data-driven basis function

A logical objection to this methodology will be that usually no information about the true function is available. The question is then whether a data-driven method to obtain a suitable basis exists?

We said that the fit will improve if one uses a basis which is similar to the true function. There is a well-known way to obtain a function which is similar to the true function: *Smoothing*. This gives us the following idea: We perform a simple

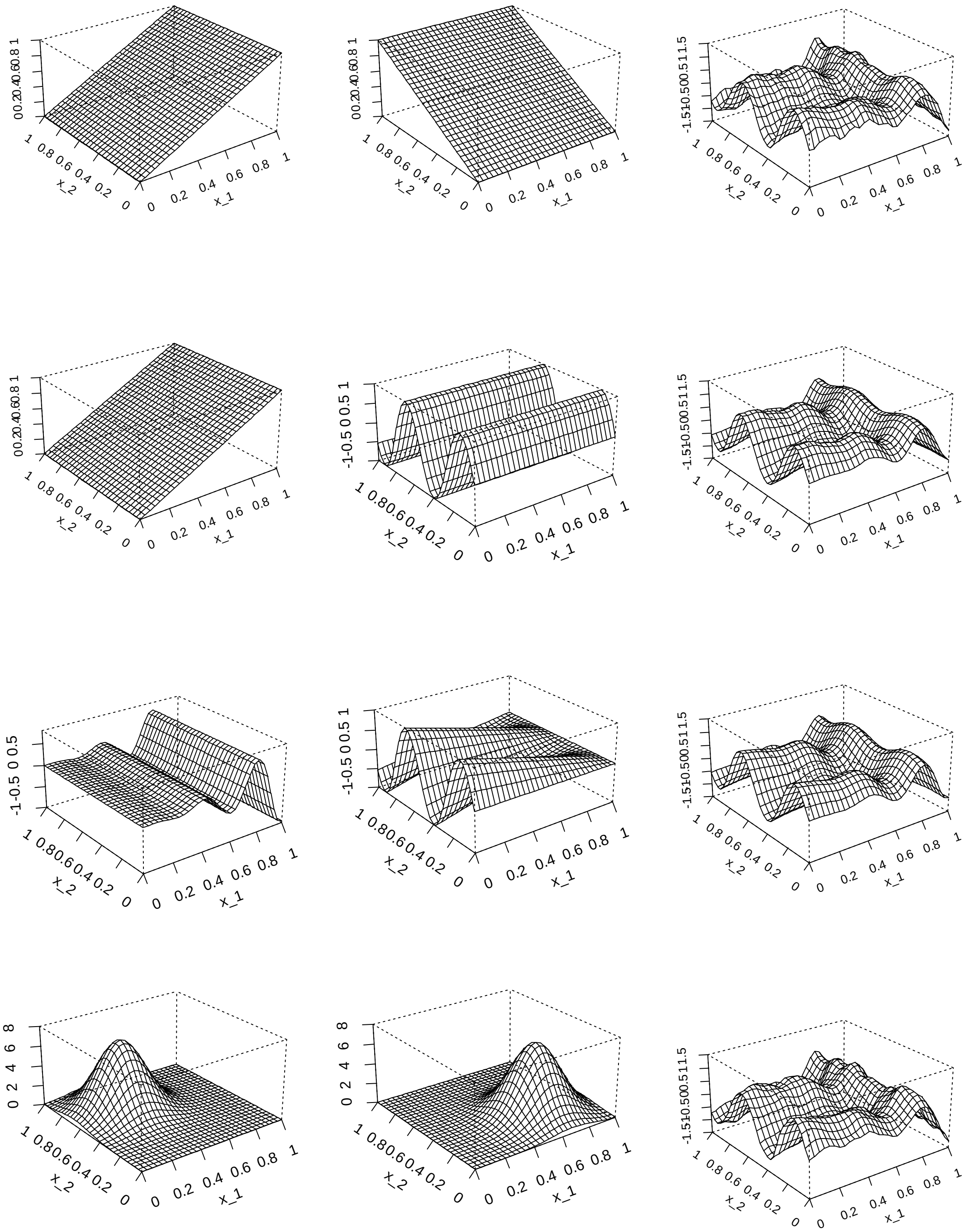



Figure 2.8: From top to bottom: Basis functions a), d), f) and g) and corresponding fits.

local linear fit and use the result as a basis function.

Returning to the previous example this means that we use the function in the top right of Fig. 2.8 as our basis function. Fitting only to this basis we obtain an $RSE$ value of 0.143 (at optimal bandwidths $h_1 = 0.29, h_2 = 0.20$), which is already a good part better than the relative error in a). However, the resulting fit in Fig. 2.9 (top right) seems to be identical to the basis we used. What is happening? The second step - smoothing with the data-driven basis - does not change the local properties of the basis. If there is a wiggly structure in the basis, this wiggly structure will be retained after the second fit. Nevertheless the fit is improved, because the global properties of the basis are modified. The range of the basis obtained from the fit in a) is $(-1.28, 0.93)$. If we smooth the data with this basis, the range blows up to $(-1.36, 1.06)$, i.e. this smoothing step is in fact a kind of backwards-smoothing of the basis, which corrects the fit where the basis was oversmoothed.

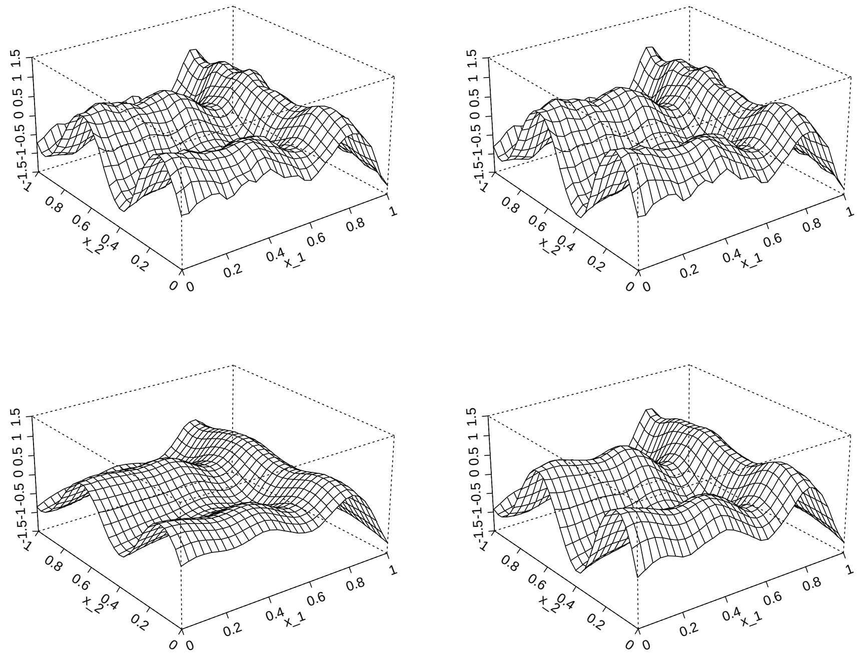



Figure 2.9: Top left: Local linear pre-fit (identical to the top right picture in Fig. 2.8); bottom left: Local linear pre-fit using the double bandwidth. On the right side is each found the fit obtained using the basis to the left.

This observation motivates us not to use the optimal bandwidths in the first fit, but somewhat higher bandwidths, in order to avoid a wiggly basis. Of course the result will be oversmoothed, but this will be corrected in the second fit. In our example calculating a local linear fit with $h_1 = 0.06$ and $h_2 = 0.08$ leads to $RSE(\hat{m}) = 0.318$, which is certainly not a very good fit, as shown in Fig. 2.9 (bottom left). However, if we use this fit as a basis for the second fit the $RSE$ can be optimized down to 0.122, what is an impressive improvement. The resulting smooth curve is shown in Fig. 2.9 (bottom right).

Summarizing the findings, we suggest the following algorithm (for $d \geq 1$ dimensions).

1. Calculate a d-dimensional local linear fit, using the double size of the optimal bandwidths. The optimal bandwidth matrix $H^{1/2} = \text{diag}(h_i)_{1 \leq i \leq d}$ can be obtained by applying usual multivariate local linear bandwidth selection routines, see e.g. Yang & Tschering (1999).

2. Use the result as a d-dimensional basis for the second fit. As a rule of thumb, use of the same bandwidths as in the first fit leads to satisfactory results.

For the verification of this algorithm we did 200 simulations of the contaminated function (2.56), and plotted the corresponding $RSE$ for the local linear fit (using the optimal bandwidth) and the fit according to the algorithm in boxplots. Since our intention was to explore the benefit of the use of a pre-fit basis and not the performance of local polynomial bandwidth selection procedures, we used the bandwidth minimizing (2.57) as optimal bandwidth - keeping in mind that this is certainly not possible for a real data set. The boxplots are shown in Fig. 2.10. The result is obvious and confirms the algorithm.

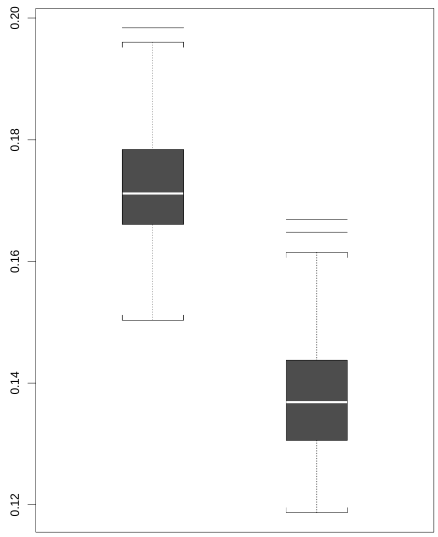



Figure 2.10: Boxplots of the $RSE$ values of 200 simulations of function (2.56); left: with local linear basis using the optimal bandwidths; right: with pre-fit basis using the double bandwidths.

Note that for $d > 1$ the multivariate pre-fit basis does not fit in the framework of Section 2.2.3. Thus the provided example shows that the idea motivated in Section 2.2.3 - to use a basis similar to the underlying function - is useful not only if the basis is free of interactions.

### 2.2.6 Discussion

We finish with some considerations about the properties that basis functions should fulfill in theory and practice. Regarding condition (A5), the theory demands the basis functions to be once or twice differentiable and to have non-vanishing gradients at the target point $x$. Differentiablility, i.e. smoothness, is also important in practice and is fulfilled by all basis functions given in this paper. In particular, the pre-fit basis in Section 2.2.5 will be sufficiently smooth for a large initial bandwidth, as proposed in the algorithm.

The condition of non-vanishing gradients, which is usually not met for the whole basis, is purely technical and of little practical relevance. The fit at points with vanishing gradients will not have apparent drawbacks compared to a fit where (A5) is fulfilled. However, only in the latter case the asymptotic bias and variance can be calculated. Taking any smooth basis, the corresponding theorems hold for all points with non-vanishing gradients of the basis functions. This will usually be fulfilled for all points except a set of measure zero.

In this paper we showed that the concepts of localization and fitting with basis functions can be combined successfully. However we stress that there exists no optimal basis function which could replace the usual polynomial basis in general. The benefit of the application of alternative basis functions depends on the amount of information which is available about the underlying function. If no information is available, we fortunately still can profit by applying the algorithm introduced in Section 2.2.5.

There is still plenty of room for further research. For example, it would be desirable to calculate more accurate estimators for the bandwidths which are used in the algorithm. In particular the factor 2, which we use to derive the initial bandwidth from the optimal bandwidth, probably can be further improved. However, a fully theoretical treatment of the pre-fit algorithm will be extremely difficult, since the basis function in the second fit is now a random variable itself.

### 2.2.7 Appendix

#### 2.2.7.1. Regularity conditions

(A1) The kernel $K$ is bounded with compact support, $\int uu^T K(u)du = \mu_2 \mathbf{I}_d$, where $\mu_2$ is a scalar and $\mathbf{I}_d$ the $d \times d$ identity matrix. In addition, all odd-order moments of $K$ vanish, i.e. $\int u_1^{l_1} \cdots u_d^{l_d} K(u)du = 0$ for all nonnegative integers $l_1, \ldots, l_d$ with an odd sum.

(A2) The point $x$ is $\in$ supp($f$). At $x$, $\sigma^2$ is continuous, $f$ is continuously differentiable and all second-order derivatives of $m$ are continuous. Further $f(x) > 0$, $\sigma^2(x) > 0$.

(A3) The sequence of bandwidth matrices $H^{1/2}$ is such that $n^{-1}|H|^{-1/2}$ and each entry of $H$ tends to zero as $n \longrightarrow \infty$ .

(A4) For a boundary point $x$, there exists a value $x_b$ on the boundary of supp($f$) with $x = x_b + H^{1/2}c$, where $c$ is a fixed element of supp(K), and a convex set $\mathcal{C}$ with nonnull interior containing $x_b$ such that $\inf_{x\in\mathcal{C}} f(x) > 0$.

(A5) At $x$, all basis functions are continuously differentiable (for variance expressions in Theorem 2.3, 2.5, 2.6) resp. twice continuously differentiable (for bias expressions in Theorem 2.5, 2.6). In either case, the point $x$ is non-singular for all basis functions, i.e. $\nabla\phi_j(x) \neq 0$ for $j = 1, \dots, q$.

For explanations and interpretations of conditions (A1) to (A4) see Ruppert & Wand (1994).

#### 2.2.7.2. Proofs

**Proof of Theorem 2.3**

Let $\mathbf{1}$ be a matrix of appropriate dimension having only entries equal to 1, further let

$$A_H = \begin{pmatrix} 1 & 0 \\ 0 & H^{1/2} \end{pmatrix} \quad \in \mathbb{R}^{d+1,d+1}, \text{ and } A_{\mathbf{1}} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{1} \end{pmatrix} \quad \in \mathbb{R}^{d+1,q+1}.$$

Note that for any $u \in \mathbb{R}^d$

$$\Phi(x + H^{1/2}u) - \Phi(x) = D_x H^{1/2}u + o(H^{1/2}\mathbf{1})$$

holds. Let $C_{x,H} = \{t : H^{-1/2}(t - x) \in \mathcal{D}_{x,H}\}$. For interior and boundary points we derive

$$\begin{aligned}
&X_x^T W_x X_x = \\
&= \sum_{i=1}^{n} K_H(X_i - x) \begin{pmatrix} 1 & (\Phi(X_i) - \Phi(x))^T \\ \Phi(X_i) - \Phi(x) & (\Phi(X_i) - \Phi(x))(\Phi(X_i) - \Phi(x))^T \end{pmatrix} \\
&= n \int_{C_{x,H}} K_H(t - x) \begin{pmatrix} 1 & (\Phi(t)-\Phi(x))^T \\ \Phi(t)-\Phi(x) & (\Phi(t)-\Phi(x))(\Phi(t)-\Phi(x))^T \end{pmatrix} f(t)\, dt \\
&\quad + n o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}}) \\
&= n f(x) \int_{\mathcal{D}_{x,H}} K(u) \begin{pmatrix} 1 & u^T H^{1/2} D_x \\ D_x^T H^{1/2} u & D_x^T H^{1/2} u u^T H^{1/2} D_x \end{pmatrix} du \qquad (2.58) \\
&\quad + n o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}}) \\
&= n f(x)(A_{D_x}^T A_H M_x A_H A_{D_x} + o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}})), \qquad (2.59)
\end{aligned}$$

and analogously

$$X_x^T \Sigma_x X_x = n|H|^{-1/2} f(x)\sigma^2(x)(A_{D_x}^T A_H N_x A_H A_{D_x} + o_P(A_{\mathbf{1}}^T A_H \mathbf{1} A_H A_{\mathbf{1}})). \quad (2.60)$$

Substituting (2.59) and (2.60) into (2.45) leads to (2.47). In the special case of an interior point we have $M_x = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \mathbf{I}_d \end{pmatrix}$. Thus (2.47) reduces to

$$\begin{aligned} \mathrm{Var}(\hat{m}(x)|\mathbb{X}) &= \frac{\sigma^2(x)}{nf(x)} |H|^{-1/2} e_1^T N_x e_1 (1 + o_P(1)) = & (2.61) \\ &= \frac{\sigma^2(x)}{nf(x)} |H|^{-1/2} \nu_0 (1 + o_P(1)). & (2.62) \end{aligned}$$

**Proof of Theorem 2.4**

We introduce the function $M : [0,1] \to \mathbb{R}$,

$$M(t) = m(y_\Phi(t)) = m(\Phi^{-1}[\Phi(x) + t(\Phi(z) - \Phi(x))]).$$

Then we have $M(0) = m(x)$ and $M(1) = m(z)$. We apply the univariate Taylor theorem on the function $M \in C^{p+1}([0,1])$ and obtain

$$M(1) = M(0) + M'(0) + \frac{1}{2!}M''(0) + \ldots + \frac{1}{p!}M^{(p)}(0) + r_{p+1}, \quad (2.63)$$

where

$$r_{p+1} = \frac{1}{(p+1)!} M^{(p+1)}(\tau) \quad (\tau \in [0,1]).$$

Using the inverse function theorem we obtain

$$y'_\Phi(t) = \left[ \frac{1}{\phi'_i\left(y_\Phi(t)_{(i)}\right)} (\phi_i(z_i) - \phi_i(x_i)) \right]_{(1 \le i \le n)}$$

Repeated application of the chain rule on $M = m \circ y_\Phi$ leads to

$$\begin{aligned} M'(t) &= \nabla m(y_\Phi(t)) \cdot y'_\Phi(t) = [((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi) m](y_\Phi(t)) \\ M''(t) &= [((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^2 m](y_\Phi(t)) \\ &\vdots \\ M^{(n)}(t) &= [((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^n m](y_\Phi(t)) \end{aligned}$$

Applying the latter formulas in (2.63) and substituting $\zeta = y_\Phi(\tau)$ proves the allegation.

**Proof of Theorem 2.5**

The proof is kept shortly since it follows mainly the ideas of the corresponding proof for multivariate local linear fitting, see Ruppert & Wand (1994).

*Asymptotic bias*

First note that, applying (2.50), we have

$$m = X_x \begin{pmatrix} m(x) \\ P_x^{-1}\nabla m(x) \end{pmatrix} + \frac{1}{2}Q_m(x) + S_m(x) \tag{2.64}$$

with

$$Q_m(x) = \left[(\Phi(X_i) - \Phi(x))^T P_x^{-1} N_m(x) P_x^{-1}(\Phi(X_i) - \Phi(x))\right]_{1\le i\le n}$$

and $S_m(x) = o(Q_m(x))$. Plugging (2.64) into (2.44) shows that

$$\text{Bias}(\hat{m}(x)|\mathbb{X}) = \frac{1}{2}e_1^T(X_x^T W_x X_x)^{-1} X_x^T W_x Q_m(x)(1 + o(1)). \tag{2.65}$$

Let $w_i = K_H(X_i - x)$. Using matrix algebra (see e.g. Fahrmeir & Hamerle (1984)) we derive

$$\begin{aligned}
&(X_x^T W_x X_x)^{-1} = \\
&= \begin{pmatrix} \sum w_i & \sum w_i(\Phi(X_i) - \Phi(x))^T \\ \sum w_i(\Phi(X_i) - \Phi(x)) & \sum w_i(\Phi(X_i) - \Phi(x))(\Phi(X_i) - \Phi(x))^T \end{pmatrix}^{-1} \\
&= n\begin{pmatrix} f(x) + o_P(1) & o_P(\mathbf{1}^T H^{1/2}) \\ o_P(H^{1/2}\mathbf{1}) & \mu_2 P_x H P_x f(x) + o_P(H) \end{pmatrix}^{-1} \\
&= \frac{1}{n}\begin{pmatrix} \frac{1}{f(x)} + o_P(1) & o_P(\mathbf{1}^T H^{-1/2}) \\ o_P(H^{-1/2}\mathbf{1}) & \frac{1}{\mu_2 f(x)} P_x^{-1} H^{-1} P_x^{-1} + o_P(H^{-1}) \end{pmatrix}
\end{aligned} \tag{2.66}$$

and

$$X_x^T W_x Q_m(x) = n\begin{pmatrix} \mu_2 f(x)\text{tr}\{H N_m(x)\} + o_P(\text{tr}(H)) \\ O_P(H^{3/2}\mathbf{1}) \end{pmatrix}, \tag{2.67}$$

so that substituting (2.66) and (2.67) into (2.65) proves (2.51).

*Asymptotic variance*

Similar like above we obtain

$$\begin{aligned}
&X_x^T \Sigma_x X_x = \\
&= \sum w_i^2 \sigma^2(X_i)\begin{pmatrix} 1 & (\Phi(X_i) - \Phi(x))^T \\ (\Phi(X_i) - \Phi(x)) & (\Phi(X_i) - \Phi(x))(\Phi(X_i) - \Phi(x))^T \end{pmatrix} \\
&= n|H|^{-1/2}\begin{pmatrix} \nu_0\sigma^2(x)f(x) + o_P(1) & \mathbf{1}^T H^{1/2}(1 + o_P(1)) \\ H^{1/2}\mathbf{1}(1 + o_P(1)) & G(x,H) + o_P(H) \end{pmatrix},
\end{aligned}$$

where

$$G(x,H) = \left(\int K^2(u)uu^T du\right) P_x H P_x \sigma^2(x) f(x).$$

Plugging this result and (2.66) into (2.45) leads to (2.52).

**Proof of Theorem 2.6**

Let

$$A_H = \begin{pmatrix} 1 & 0 \\ 0 & H^{1/2} \end{pmatrix}, \quad A_{P_x} = \begin{pmatrix} 1 & 0 \\ 0 & P_x \end{pmatrix}.$$

*Asymptotic bias*

Note that

$$\begin{aligned} & X_x^T W_x X_x = \\ &= n\int_{C_{x,H}} K_H(t-x) \begin{pmatrix} 1 & (\Phi(t)-\Phi(x))^T \\ (\Phi(t)-\Phi(x)) & (\Phi(t)-\Phi(x))(\Phi(t)-\Phi(x))^T \end{pmatrix} f(t)\,dt \\ & \quad + n o_P(A_H \mathbf{1} A_H) \\ &= nf(x)(A_{P_x} A_H M_x A_H A_{P_x} + o_P(A_H \mathbf{1} A_H)), \end{aligned} \tag{2.68}$$

where $C_{x,H}$ was defined in the proof of Theorem 2.3. Using the first step in (2.67),

$$\begin{aligned} & X_x^T W_x Q_m(x) = \\ &= nf(x) \begin{pmatrix} \int_{\mathcal{D}_{x,H}} K(u)u^T H^{1/2} N_m(x) H^{1/2} u\,du \\ P_x H^{1/2} \int_{\mathcal{D}_{x,H}} uK(u)\{u^T H^{1/2} N_m(x) H^{1/2} u\}\,du \end{pmatrix} \\ & \quad + o_P \begin{pmatrix} n\mathrm{tr}(H) \\ nH^{1/2}\mathbf{1}\mathrm{tr}(\mathbf{H}) \end{pmatrix} \end{aligned} \tag{2.69}$$

holds. Assuming (A4), $M_x$ is nonsingular and we have

$$M_x^{-1} = \begin{pmatrix} \mu_x^{11} & \mu_x^{12} \\ \mu_x^{21} & \mu_x^{22} \end{pmatrix},$$

where $\mu_x^{11} = (\mu_{x,11} - \mu_{x,12}\mu_{x,22}^{-1}\mu_{x,21})^{-1}$, $\mu_x^{12} = -(\mu_{x,12}/\mu_{x,11})\mu_x^{22}$ and $\mu_x^{22} = (\mu_{x,22} - \mu_{x,21}\mu_{x,12}/\mu_{x,11})^{-1}$. Then substituting (2.68) and (2.69) into (2.65) and noticing that

$$e_1^T A_{P_x}^{-1} A_H^{-1} M_x^{-1} A_H^{-1} A_{P_x}^{-1} = \begin{pmatrix} \mu_x^{11} & \mu_x^{12} P_x^{-1} H^{-1/2} \end{pmatrix}$$

yields formula (2.54).

*Asymptotic variance*

Similar considerations like in (2.68) lead to

$$X_x^T W_x^2 X_x = nf(x)|H|^{-1/2}(A_{P_x} A_H N_x A_H A_{P_x} + o_P(A_H \mathbf{1} A_H)). \tag{2.70}$$

With (2.45), (2.68) and (2.70) we get

$$\begin{aligned}
&\mathrm{Var}(\hat{m}(x)|\mathbb{X}) = \\
&= e_1^T(X_x^T W_x X_x)^{-1}(X_x^T W_x^2 X_x)(X_x^T W_x X_x)^{-1} e_1(\sigma^2(x) + o_P(1)) \\
&= \frac{\sigma^2(x)}{nf(x)}|H|^{-1/2}\left(e_1^T M_x^{-1} N_x M_x^{-1} e_1 + o_P(1)\right),
\end{aligned}$$

what had to be proven.

## 2.3 Remarks on the generalized Taylor expansion

In Theorems 2.1 and 2.4 we provided simple extensions of Taylor's theorem. In the univariate case, we showed that it is possible to replace the polynomials by powers of an arbitary smooth function. In the multivariate case we provided a theorem which covers the case that each coordinate is modeled separately by an individual basis function.

There are some topics concerning the generalized versions of Taylor's theorem which deserve to be treated in more depth than it has been done yet. In Section 2.3.1 we repeat some properties of the univariate Taylor expansion along with some historic notes. Further we analyze how Theorem 2.1 is related to Taylor's theorem and other well-known theorems. The content of Section 2.3.2 is more technical than theoretical. Some notations and properties concerning Theorem 2.4 are explained that have been treated only rudimentarily in the previous section.

### 2.3.1 Univariate generalized Tayor expansion

One of the most widely applied mathematical theorems is that of Taylor (1715), which allows to approximate a function by a linear combination of polynomials. Taylor did not specify the remainder term, the first representation of which is due to Lagrange (1797). Today the theorem is mostly found in the following form (see e.g. Lay, 1990, page 211):

**Taylor's theorem with Lagrange's form of the remainder.**

*Let $m : [v, w] \longrightarrow \mathbb{R}$ be $p$ times continuously differentiable and $p+1$ times differentiable in $(v, w)$, and let $x \in [v, w]$. Then for each $z \in [v, w]$ with $z \neq x$ there exists a point $\zeta \in (x, z)$ respectively $(z, x)$ such that*

$$m(z) = \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j + r_{p+1}(z), \tag{2.71}$$

*with*

$$r_{p+1}(z) = \frac{m^{(p+1)}(\zeta)}{(p+1)!}(z-x)^{p+1}$$

*holds.*

Already before Gregory, Newton, Leibniz, J. Bernoulli and de Moivre had discovered variants of Taylor's theorem by providing expansions for special functions. It was Taylor's achievement to join the results in a general theorem. The importance of Taylor's theorem remained unrecognized until 1772 when Lagrange proclaimed it the basic principle of differential calculus. In general, *"Taylor was a mathematician of far greater depth than many have given him credit for"* (O'Connor & Robertsen, 2000), as explained by Jones (1951): *"A study of Brook Taylor's life and work reveals that his contribution to the development of mathematics was substantially greater than the attachment of his name to one theorem would suggest."* Some of the topics where he did innovative contributions are the law of magnetic attraction, the center of oscillation, and perspective problems as the treatment of vanishing points. Finally he found the inverse function theorem, a way of relating the derivative of a function to the derivative of the inverse function.

There exists also the following formulation of Taylor's theorem, found by Cauchy (1821):

**Taylor's theorem with exact integral representation of the remainder.**

*Let $I$ be a non-trivial interval, $m : I \longrightarrow \mathbb{R}$ be $p+1$ times continuously differentiable in $I$ and $x \in I$. Then for all $z \in I$*

$$m(z) = \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j + r_{p+1}(z), \tag{2.72}$$

*where*

$$r_{p+1}(z) = \frac{1}{p!}\int_x^z (z-t)^p m^{(p+1)}(t)dt\,, \tag{2.73}$$

*holds.*

Beside the different form of the remainder, there is a small difference in the assumptions: The second approach using the integral form requires $m$ to be $p+1$ times *continuously* differentiable. This is due to the partial integration which is done in the proof. The Lagrange approach uses Rolle's theorem, and this only requires that $m^{(p)}(\cdot)$ is differentiable. The Lagrange form can easily be derived from the integral form by using the mean value theorem for integrals. However, if this way is chosen, one requires again $m$ to be $p+1$ times *continuously* differentiable. For a deeper comparison between these two remainder representations see Firey (1960), and for an overview of other remainder forms we refer to Blumenthal (1926). Using observation (2.6), or directly via partial integration similar to the corresponding proof of Taylor's theorem (see e.g. Forster, 1999, page 226), the remainder term in Theorem 2.1 can be written in integral form:

**Theorem 2.7 (Generalized Taylor expansion - integral form).**
*Let $I$ be a non-trivial interval, $m, \phi : I \longrightarrow \mathbb{R}$ be $p+1$ times continuously differentiable in $I$, $\phi'(\cdot) \neq 0$ in $I$ and $x \in I$. Then for all $z \in I$*

$$m(z) = \sum_{j=0}^{p} \frac{\psi_{(j)}(x)}{j!}(\phi(z) - \phi(x))^j + s_{p+1}(z), \tag{2.74}$$

*where*

$$s_{p+1}(z) = \frac{1}{p!} \int_x^z (\phi(z) - \phi(t))^p \psi_{(p+1)}(t) \phi'(t)\, dt\,,$$

*holds.*

Taylor did not investigate convergency of the *Taylor series*, an expression which seems to has been used for the first time by Lhuilier (1786). The Taylor series for the function $m$ at the point $x$ is given by

$$\sum_{j=0}^{\infty} \frac{m^{(j)}(x)}{j!}(z - x)^j.$$

Under certain conditions, which are, however, hard to specify in general (Bernstein, 1914), this infinite sum equals $m(z)$. Vice versa, every power series $m(z) = \sum_{j=0}^{\infty} a_j (z-x)^j$ can be written as $m(z) = \sum_{j=0}^{\infty} \frac{m^{(j)}(x)}{j!} \cdot (z-x)^j$ within the radius of convergence (see Forster, 1999, page 230). The last property is crucial for the applicability of Taylor's theorem, because it ensures the uniqueness of parameters. Also for the generalized Taylor series the parameters are unique, as the following lemma shows:

**Lemma 2.1 (Uniqueness in the univariate case).**
*Let $m(z) = \sum_{j=0}^{\infty} b_j (\phi(z) - \phi(x))^j < \infty$. Then*

$$b_j = \frac{\psi_{(j)}(x)}{j!},$$

*with $\phi$ and $\psi_{(j)}$ as in Theorem 2.1.*

The condition $< \infty$ is important and reflects that this lemma only holds within the convergence region of the series. From (2.6) we may also conclude that the generalized Taylor series $\sum_{j=0}^{\infty} \frac{\psi_{(j)}(x)}{j!}(\phi(z) - \phi(x))^j$ is convergent on the interval $(\phi^{-1}(\phi(x) - r), \phi^{-1}(\phi(x) + r))$, where $r$ is the convergence radius of the Taylor series of $m \circ \phi^{-1}$ centered at $\phi(x)$.

**Proof of Lemma 2.1**

It is

$$\begin{aligned}
\psi_{(0)}(z) &= m(z) = \sum_{j=0}^{\infty} b_j(\phi(z) - \phi(x))^j \\
\psi_{(1)}(z) &= \frac{m'(z)}{\phi'(z)} = \sum_{j=1}^{\infty} j b_j(\phi(z) - \phi(x))^{j-1} \\
&\vdots \\
\psi_{(k)}(z) &= \frac{m^{(k)}(z)}{\phi'(z)} = \sum_{j=k}^{\infty} j(j-1)\dots(j-k+1) b_j(\phi(z) - \phi(x))^{j-k} ,
\end{aligned}$$

and thus

$$\psi_{(k)}(x) = k! b_k .$$

□

One verifies easily that Theorem 2.1 reduces for $p = 0$ to the well-known generalized mean value theorem (Cauchy, 1821, see Forster, 1999, page 163), however with the small constraint that one of the two involved functions has to be invertible in the considered interval. Theorem 2.7 reduces for $p = 0$ to the fundamental theorem of integral calculus. Summarizing, we see that the generalized Taylor theorem in Lagrange form fits in the known framework as follows:

$$\begin{tikzcd}
\boxed{\text{Theorem 2.1}} \arrow[r, "\phi = id"] \arrow[d, "p = 0"] & \boxed{\begin{array}{c}\text{Taylor's Theorem} \\ \text{(Taylor 1715,} \\ \text{Lagrange 1797)}\end{array}} \arrow[d, "p = 0"] \\
\boxed{\begin{array}{c}\text{Generalized Mean Value} \\ \text{Theorem (Cauchy 1821)}\end{array}} \arrow[r, "\phi = id"] & \boxed{\begin{array}{c}\text{Mean Value Theorem} \\ \text{(Lagrange 1797)}\end{array}}
\end{tikzcd}$$

We notice that Theorem 2.1 fits quite naturally in the existing framework. Considering the age of the theorems surrounding it, it seems to be unimaginable that it has not been introduced before. What is known about extensions of Taylor's theorem? Research in this direction started late, what gave Widder (1928) reason to state that *"in view of the great importance of Taylor's series in analysis, it may be regarded as extremely surprising that so few attempts at generalization have been made"*. However, meanwhile there exist a large variety of extensions, which are nearly uniformly named "Generalized Taylor formula" (theorem, series, expansion, ...), complicating literature scanning in that aera. This covers e.g. extensions for functions of complex variables (Walsh, 1929, Widder, 1929), generalizations based on the Kronecker formula (Hummel & Seebeck, 1949, Boas, 1970), generalizations for non-integer valued exponents (Trujillo, Rivero & Bonilla, 1999), generalizations for fractional derivatives (Raina & Koul, 1979) and incomplete differentials (Grabert, 1982), probabilistic (Massay & Whitt, 1993) and random Taylor series (Ding, 1998). Nevertheless, to our knowledge a Taylor expansion for power basis functions has not been considered yet (or, maybe, has never been regarded as sufficiently useful to be published). The generalization that is the most connected to our one is that of Widder (1928). To understand Widder's theorem, we need firstly to introduce the following definition:

**Definition 2.2.**
*The function*

$$a_p(z) = \sum_{j=0}^{p} c_j \phi_j(z) \tag{2.75}$$

*is a function of approximation for function $m(z)$ of order $p$ for the point $z = x$ if the functions $\phi_j(z)$ are of class $C^p$ (i.e. possess $p$ continuous derivatives) in a neighborhood of $z = x$, and if $a_p(z)$ has contact of order $p$ at least with $m(z)$ at $z = x$, i.e. if*

$$a_p^{(k)}(x) = g^{(k)}(x), \quad k = 0, \ldots, p.$$

It was precisely this approach which led Maclaurin (1742) to rediscover Taylor's theorem, certainly involving polynomials instead of arbitrary smooth functions. Recall now that for basis functions $\phi_0(z), \ldots, \phi_p(z) \in C^p$ Wronski's determinant is defined by

$$W_p(z) \equiv W[\phi_0(z), \phi_1(z), \cdots, \phi_p(z)] = \begin{vmatrix} \phi_0(z) & \phi_1(z) & \cdots & \phi_p(z) \\ \phi_0'(z) & \phi_1'(z) & \cdots & \phi_p'(z) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0^{(p)}(z) & \phi_1^{(p)}(z) & \cdots & \phi_p^{(p)}(z) \end{vmatrix}.$$

Now we are able to formulate

**Widder's generalization of Taylor's formula.**

*If the functions $m(z), \phi_0(z), \phi_1(z), \ldots, \phi_p(z)$ are of class $C^p$ in a neighborhood of $z = x$, and if $W_p(x) \neq 0$, then there exists a unique function of approximation*

$$a_p(z) = \sum_{j=0}^{p} c_j \phi_j(z) = -\left(\frac{1}{W_p(x)}\right) \begin{vmatrix} 0 & \phi_0(z) & \phi_1(z) & \cdots & \phi_p(z) \\ m(x) & \phi_0(x) & \phi_1(x) & \cdots & \phi_p(x) \\ m'(x) & \phi_0'(x) & \phi_1'(x) & \cdots & \phi_p'(x) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m^{(p)}(x) & \phi_0^{(p)}(x) & \phi_1^{(p)}(x) & \ldots & \phi_p^{(p)}(x) \end{vmatrix} \quad (2.76)$$

*of order $p$ for $z = x$.*

Taylor's expansion is quite easily derived from Widder's formula. Setting $\phi_j(z) = z^j, j = 0, \ldots, p$, one obtains

$$W_p(x) = p!(p-1)! \cdots 2!1!,$$

and the determinant in formula (2.76) can be written as

$$-p!(p-1)! \cdots 2!1! \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j$$

which leads immediately to Taylor's formula. Obviously Widder's generalization covers Theorem 2.1 widely. However, the fact that it theoretically *covers* Theorem 2.1, does not necessarily mean that the latter one might be directly *derived* from it. Indeed, when setting $\phi_j(z) = \phi^j(z), j = 0, \ldots, p$, with $\phi \in C^p$, one notices rapidly that the complexity of the Wronskian as well as the determinant in (2.76) makes Widder's formula hard to apply. The problem is thereby that with each iterative differentation (from one line of the Wronskian to the deeper one) the number of terms rises exponentially, leading to intractable expressions already after some few steps. Widder also provides some other representations of the presented theorem, however the problem is inherent to all of them. Nevertheless, for very small degrees $p$ the equivalence of Widder's formula and Theorem 2.1 can be shown. Let e.g. $p = 1$, then

$$W_1(x) = \phi'(x)$$

and the determinant in (2.76) turns out to be equal to

$$-m(x)\phi'(x) - m'(x)(\phi(z) - \phi(x)),$$

yielding

$$a_1(z) = m(x) + \frac{m'(x)}{\phi'(x)}(\phi(z) - \phi(x))$$

in conformity to Theorem 2.1. Note further that the demanded invertibilty of $\phi$ corresponds to the condition of a non-vanishing Wronskian.

For a given series of basis functions $\phi_j(z), j = 0, 1, 2, \ldots$ , the parameters $c_j$ of the approximation (2.76) will normally not be interpretable, the series (2.76) will not converge, and the remainder term cannot be specified. Thus, it would be desirable to rewrite the series in the manner

$$a_0(z) + [a_1(z) - a_0(z)] + [a_2(z) - a_1(z)] + \ldots$$

with differences $a_j(z) - a_{j-1}(z) \longrightarrow 0$ as $j \longrightarrow \infty$. Widder (1928, p. 138, Theorems II and III) shows that this is indeed possible. Under some conditions on the Wronskians $W_j(z), j = 0, 1, 2, \ldots$, the series

$$d_0(x)g_0(z, x) + d_1(x)g_1(z, x) + d_2(x)g_2(z, x) + \ldots,$$

where

$$g_j(z, x) = \left(\frac{1}{W_j(x)}\right) \begin{vmatrix} \phi_0(x) & \phi_1(x) & \cdots & \phi_j(x) \\ \phi_0'(x) & \phi_1'(x) & \cdots & \phi_j'(x) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0^{(j-1)}(x) & \phi_1^{(j-1)}(x) & \cdots & \phi_j^{(j-1)}(x) \\ \phi_0(z) & \phi_1(z) & \cdots & \phi_j(z) \end{vmatrix},$$

and

$$d_j(z) = \frac{W[\phi_0(z), \phi_1(z), \cdots, \phi_{j-1}(z), m(z)]}{W_{j-1}(z)},$$

converges absolutely to $m(z)$ in a symmetric interval around $x$. Using this representation, the remainder term of a $p-$th order approximation can be written as

$$\int_x^z g_p(z, t)d_{p+1}(t)\, dt.$$

The functions $g_j(z, x)$ are linear combinations of the functions $\phi_0(z), \ldots, \phi_j(z)$. As an example consider Taylor's formula: Applying the basis $\phi_j(z) = z^j$, $j = 0, 1, 2, ...$, one obtains

$$g_j(z, x) = (z - x)^j = \sum_{k=0}^{j} \binom{j}{k} (-x)^{j-k} z^k$$

and interpretable parameters $d_j(x) = \frac{m^{(j)}(x)}{j!}$. The remainder term reduces to (2.73). As further example, Widder (1928) considers a Fourier basis $1, \sin jx, \cos jx, j = 1, \ldots, \infty$, and shows with substantial effort that the generalized Taylor expansion around $z = 0$ is given by

$$m(z) = \sum_{j=0}^{\infty} \left( A_j \frac{2^j}{(2j)!}(1 - \cos z)^j + B_j \frac{2^j}{(2j + 1)!}(1 - \cos z)^j \sin z \right),$$

where

$$A_j = D^2(D^2 + 1^2)(D^2 + 2^2) \cdots (D^2 + (j-1)^2)m(0)$$

and

$$B_j = D(D^2 + 1^2)(D^2 + 2^2) \cdots (D^2 + j^2)m(0),$$

with $D = \frac{d}{dz}$ being a differential operator. We observe that the series does not use the original Fourier basis, but a somehow more complicated basis. Calculating the bias of a local approximation using a Fourier basis would, however, require to have an expansion with respect to the Fourier basis itself. Considering the case that one really plans to perform local fitting against the basis functions $(1 - \cos z)^j$ and $(1 - \cos z)^j \sin z$, this series, however, might be suitable: For $j = 0$ one has $A_0 = m(0)$ and $B_0 = m'(0)$ and thus the underlying function and its derivative can be reconstructed, at least at $z = 0$. Summarizing, though Widder's expansion allows to employ an arbitary smooth basis, these basis functions yield generalized Taylor series of functions which are probably not of interest in a real application. The deciding question for applicability of Widder's expansion on topics as treated in Section 2.1 and 2.2 is the following: Which basis $\phi_j(z), j = 0, 1, 2, \ldots$, has to be chosen to receive a particular generalized Taylor series $\sum_j d_j g_j(z, x)$, which enables us to calculate the bias of a local fit against a *prescribed* basis $g_j(z, x), j = 0, 1, 2, \ldots$? This question, which is not a problem of statistics but rather a problem of mathematics, is not yet answered. Thus, for the time being we have to be content with constraining on the analysis of local fitting against very special basis functions, as the power basis in Section 2.1.

### 2.3.2 Multivariate generalized Taylor expansion

The multivariate version of Taylor's theorem, as found e.g. in Königsberger (2000), page 65, can be written as follows:

**Multivariate Taylor expansion.**

*Assume $U \subset \mathbb{R}^d$ open, $m \in C^{p+1}(U)$ , $p \geq 0$, further $x$, $z$ and their connecting line $C_R(x, z)$ in $U$. Then there exists a point $\zeta \in C_R(x, z)$ with*

$$m(z) = m(x) + \sum_{j=1}^{p} \frac{1}{j!} \left[ ((z - x) \cdot \nabla)^j m \right](x) + R_{p+1}(z), \tag{2.77}$$

*where*

$$R_{p+1}(z) = \frac{1}{(p+1)!} \left[ ((z - x) \cdot \nabla)^{p+1} m \right](\zeta).$$

It was shown in Theorem 2.6 (Section 2.2.3) how this multivariate version of Taylor's theorem can be extended to a wider class of basis functions. The notation of Theorem 2.6 has to be handled carefully. For example, in the case $d = 2$ with $z = (z_1, z_2)$ and $x = (x_1, x_2)$ it is

$$\begin{aligned}
\left[((\Phi(z) - \Phi(x)) \cdot \nabla_\Phi)^2 m\right](x) &= \left[\left(\begin{pmatrix} \phi_1(z_1) - \phi_1(x_1) \\ \phi_2(z_2) - \phi_2(x_2) \end{pmatrix} \circ \begin{pmatrix} \frac{\partial_1}{\phi_1'(\cdot)} \\ \frac{\partial_2}{\phi_2'(\cdot)} \end{pmatrix}\right)^2 m\right](x) = \\
&= (\phi_1(z_1) - \phi_1(x_1))^2 \frac{1}{\phi_1'(x_1)} \partial_1 \left(\frac{\partial_1 m}{\phi_1'(\cdot)}\right)(x) + \\
&+ 2(\phi_1(z_1) - \phi_1(x_1))(\phi_2(z_2) - \phi_2(x_2)) \frac{1}{\phi_2'(x_2)} \partial_2 \left(\frac{\partial_1 m}{\phi_1'(\cdot)}\right)(x) + \\
&+ (\phi_2(z_2) - \phi_2(x_2))^2 \frac{1}{\phi_2'(x_2)} \partial_2 \left(\frac{\partial_2 m}{\phi_2'(\cdot)}\right)(x).
\end{aligned}$$

We see that this notation, though easily written down in the theorem, turns out to be cumbersome in practice. We therefore set

$$\eta_{ij} m(x) = \frac{1}{\phi_i'(x)} \partial_i \left(\frac{\partial_j m}{\phi_j'(\cdot)}\right)(x).$$

Then it can be easily shown that

$$\eta_{ij} m(x) = \eta_{ji} m(x) \quad (i \neq j) \tag{2.78}$$

and

$$\eta_{ii} m(x) = \frac{1}{(\phi_i'(x))^2} \partial_i^2 m(x) - \frac{\phi_i''(x)}{(\phi_i'(x))^3} \partial_i m(x) \tag{2.79}$$

holds, so that the matrix $N_m(x) := P_x(\eta_{ij} m(x))_{1 \le i,j \le d} P_x$ has the property (2.49)

$$N_m(x) = H_m(x) - P_x^{-1} P_x' \operatorname{diag}(\nabla m(x)),$$

where $H_m(x)$ is the Hessian matrix of $m$. Using this notation, the formula (2.48) can be written as (2.50).

Uniqueness of parameters in the multivariate case can be shown similarly as for the univariate case, using the the idea of the corresponding proof for a polynomial basis, see e.g. Königsberger (2000), page 67. To illustrate the multivariate generalized Taylor expansion, we provide a simple

**Example.**

Let $d = 2$, $p = 1$ and

$$m(z_1, z_2) = \log(z_1) \sin(z_2). \tag{2.80}$$

A first order Taylor approximation of $m$ around $x = (4, 7)$ approximates $m$ by the tangential plane

$$m(z_1, z_2) \approx -7.08 + 0.16 z_1 + 1.05 z_2.$$

Taking the basis functions

$$\phi_1(z_1) = \log(z_1) \quad \text{and} \quad \phi_2(z_2) = \sin(z_2),$$

the multivariate generalized Taylor theorem substitutes this tangential plane by a 2-dimensional tangential curve, i.e.

$$\begin{aligned} m(z_1, z_2) \approx m(x_1, x_2) &+ (\phi_1(z_1) - \phi_1(x_1)) \cdot \frac{\partial_1 m(x_1, x_2)}{\phi_1'(x_1)} \\ &+ (\phi_2(z_2) - \phi_2(x_2)) \cdot \frac{\partial_2 m(x_1, x_2)}{\phi_2'(x_2)} \\ = -0.49 &+ 0.64 \log z_1 + 1.39 \sin z_2. \end{aligned}$$

The function (2.80) is shown in Figure 2.11 (top). The tangential plane spanned by a usual first order Taylor approximation is shown in Figure 2.11 (middle). The approximation via the generalized Taylor expansion is depicted in Figure 2.11 (bottom). Certainly Theorem 2.4 justifies the latter expansion only in a region around $x$ where $\phi_1$ and $\phi_2$ are invertible. This region is roughly the mountain side where the point $x$ is situated, i.e. the area between the middle crest and the first valley. But, as seen in Figure 2.11, the expansion might yield reasonable results even beyond this region.

## Final remarks

The content of Section 2.2 has been published in *Computational Statistics* (Einbeck, 2003). The copyright is hold by Physica-Verlag, c/o Springer-Verlag. The article is reprinted with the kind permission of the publisher. The author is grateful to Daniel Rost and Hubert Kalf (Math. Inst., University of Munich) for helpful suggestions concerning the extendibility of Taylor's theorem.

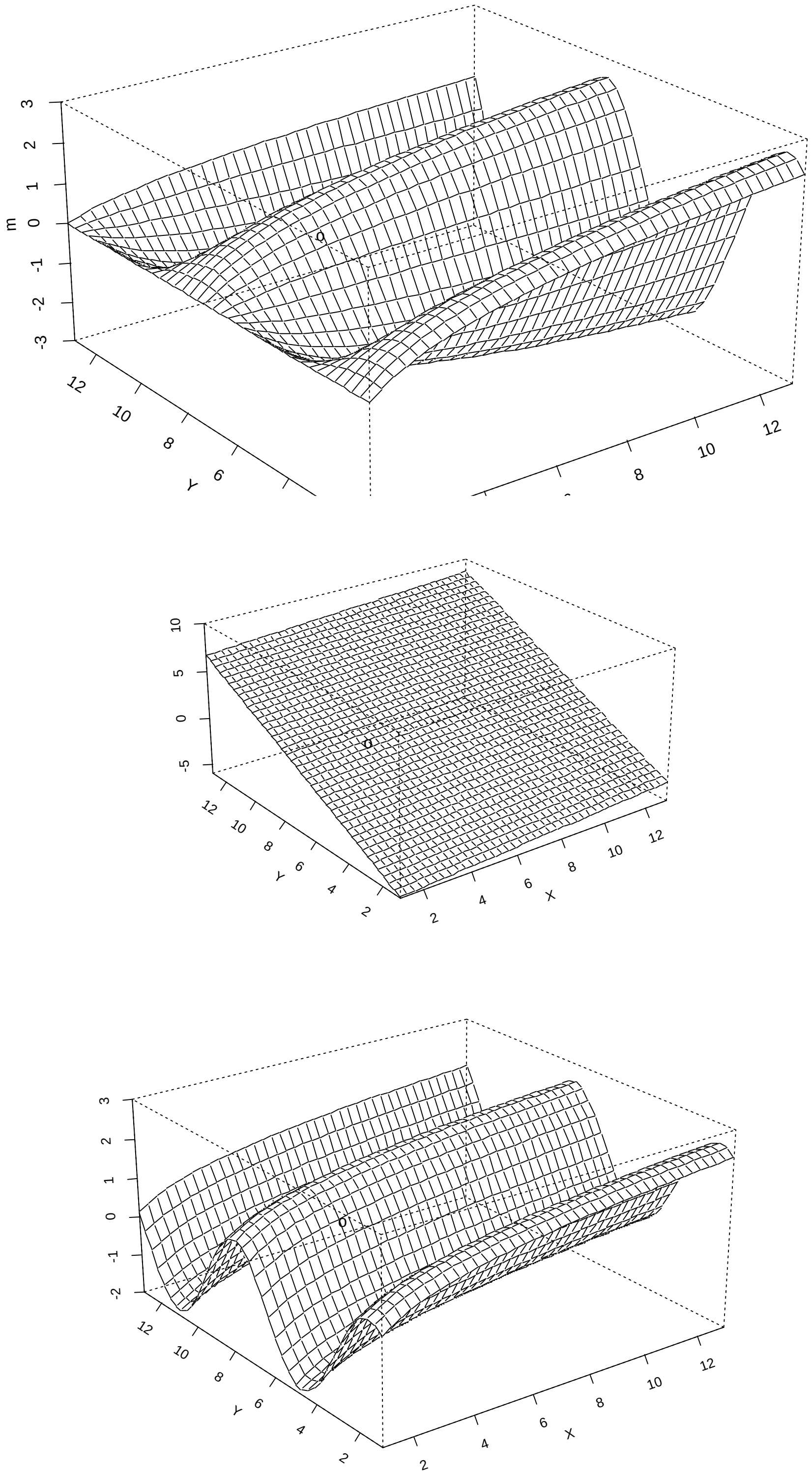



Figure 2.11: Top: function (2.80); middle: tangential plane via expansion (2.77); bottom: tangential curve via expansion (2.48). The point $x = (4, 7)$ is symbolized by "o".

# Chapter 3

# Local Smoothing with Robustness against Outlying Predictors

## 3.1 Introduction

Smoothing methods are widely used to eliminate random noise in regression problems involving time series of explanatory and response variables. Typical examples are studies of the association between daily measures of pollutant concentrations in the atmosphere and mortality as considered in Schwartz (1994), Braga et al. (2001) and Singer et al. (2002), among others. In such ecological studies, the frequent presence of outlying observations requires robust smoothing techniques and for such purposes, the LOWESS (LOcally WEighted Scatter plot Smoothing) technique has been successfully employed to downweight the effect of outliers in the response variable (Cleveland, 1979). The idea of the LOWESS technique is to carry out a series of iteratively reweighted local polynomial fits, where, in each step, the points with the largest residuals in the previous step are downweighted. Alternatively, one of the several recently published robust nonparametric methods may also be considered. For example, a common approach to robustification is to replace the quadratic loss function $l(z) = z^2$ by functions which are less sensitive to outliers, e.g. the $L_1$ norm $l(z) = |z|$ as proposed by Wang & Scott (1994).

More specifically, in the context of local constant fitting, the estimate of a function $m(\cdot)$ at point $x$, given the data $(X_i, Y_i), i = 1, \ldots, n$, is

$$\hat{m}(x) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} w_i(x) l(Y_i - a)$$

with weights $w_i(x) = K[(X_i - x)/h]$, where $K$ is a kernel function and $h$ is the bandwidth. These local $M$-estimators are discussed in Härdle & Gasser (1984), Truong (1989) and Hall & Jones (1990). An improvement on such estimators involve a local linear instead of constant fit, as discussed in Tsybakov (1986), Fan, Hu & Truong (1994) and Yu & Jones (1998). Honda (2000) enriched the concept by accounting for correlated errors. These papers, however, deal with robustness against outlying responses. The task of how to treat outliers in the predictors remains unexamined, a fact which was already noted by Hastie & Tibshirani (1990).

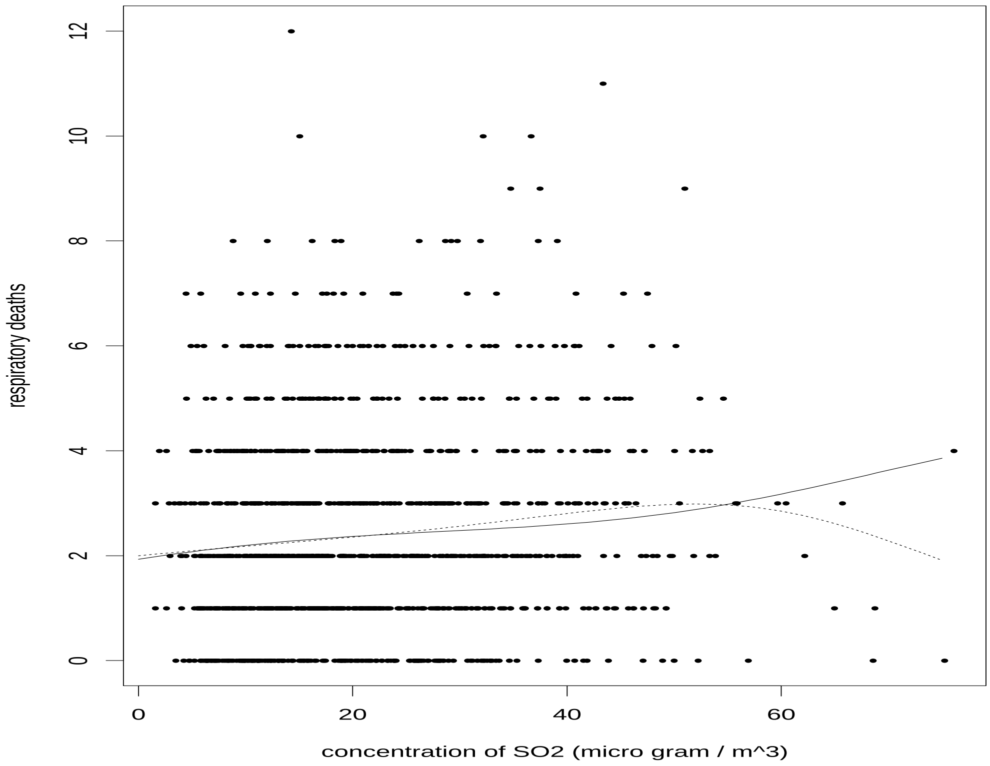



Figure 3.1: Respiratory deaths versus $SO_2$ concentration, local linear fit (dotted) and fit with robustness to horizontal outliers (solid).

To illustrate the importance of the development of such techniques we consider the data set analyzed by Conceição et al. (2001) and Singer et al. (2002) to evaluate the association between mortality attributed to respiratory causes of children under five and the concentration of $\mathrm{PM}_{10}$, $\mathrm{SO}_2$, $\mathrm{O}_3$ and CO in the city of São Paulo, Brazil, from 1994 to 1997 (the data is available in `www.ime.usp.br/~jmsinger`). The number of daily respiratory deaths as a function of the $\mathrm{SO}_2$ concentration is depicted in Figure 3.1. Days with high pollutant concentrations (as compared to the majority of the data) are clearly identified. The effect of such observations is to "pull" the fitted curve downward (dotted line), suggesting that the effect of the pollutant on children mortality decreases for concentrations beyond 50 ($\mu g/\mathrm{m}^3$), a fact that has no biological plausibility. To better understand the effect, observe the two data points at the lower right side of the picture. Although they do not seem to correspond to vertical outliers, they definitely disturb the local linear fit. A possible reason for this is that high concentrations of the pollutant are not coupled with a large number of deaths, contrary to what is expected; this is probably due to the sparse design for large concentrations. It seems clear that some robust method must be employed to bypass this inconsistency. In the light of this example, we consider a robust local polynomial smoothing technique which downweights the effect of outliers both in the response and in the explanatory variables (application of this method on the given data yields the solid line in Figure 1). It may also be used as a diagnostic tool to identify outliers in the explanatory variables.

In Section 3.2 we give a short review over some existing related concepts. In Section 3.3 we introduce the new outlier robust smoother and in Section 3.4 we return to the example addressed above. We finish with a discussion in Section 3.5.

## 3.2 An overview of related concepts

### 3.2.1 Ridging

Seifert & Gassser (1996, 2000) show that the conditional variance of a local linear fit can be unbounded in situations where the design is clustered or sparse. In many cases, the presence of sparse data is a problem highly related to outlying predictors. As a solution, they propose to use data adaptive ridging, i.e. they replace the local linear fit by a weighted sum of a local linear and a local constant fit. An appropriate choice of a data adaptive selected ridge parameter achieves a balance between these two estimators and is successful in robustifying the procedure against unbounded variance. However, this form of robustification is not exactly what we desire here. Note that outlying predictors correspond to regions with sparse design and in those situations the ridge estimator performs a local constant fit. Thus, the estimator will more or less reproduce the response value associated to the outlying observations, which is the opposite of what we expect of an outlier robust method.

### 3.2.2 Variable bandwidth

Fan & Gijbels (1992) discuss a local linear estimator based on a global variable bandwidth, i.e., a bandwidth which depends on the predictors. In particular, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a population $(X, Y)$. Assume that $m(x) = E(Y|X = x)$ is the mean regression function of $Y$ given $X$ and let $f(\cdot)$ denote the (design) density of $X$. Then

$$\sum_{i=1}^{n} \left(Y_i - a(x) - b(x)(x - X_i)\right)^2 \alpha(X_i) K\left[\frac{x - X_i}{h_n}\alpha(X_i)\right] \tag{3.1}$$

is minimized in terms of $a(x)$ and $b(x)$, applying a variable bandwidth $h(X_i) = h_n/\alpha(X_i)$, where $\alpha(\cdot)$ is some nonnegative function. This leads to the local estimator $\hat{m}(x) = \hat{a}(x)$. Fan & Gijbels (1992) show that the optimal variable bandwidth, i.e., the bandwidth minimizing the asymptotic MSE, is achieved by setting $\alpha(x)$ proportional to $(f(x)[m''(x)]^2/\sigma^2(x))^{1/5}$, where $m''$ denotes the second derivative of $m$ and $\sigma^2(x)$ denotes the conditional variance of $Y$. For $\alpha(x) = f(x)$, the estimator $\hat{m}(x)$ obtained by minimizing (3.1) corresponds approximately to a nearest-neighbour estimator.

Although not explicitly stated by the authors, this kind of estimator may be considered as a first step towards robustification against outlying predictors. Let us assume that $\alpha(\cdot)$ is any monotone increasing function of $f(\cdot)$. Then the factor $\alpha(X_i)$ in the minimization problem (3.1) downweights all points with a small design density, which is what we expect for the outlying covariates. There is, however, a serious drawback with this approach to robustification. The function $\alpha(\cdot)$ appears again in the argument of the kernel $K$ and covariates lying in sparse regions (e.g. outliers) become associated to huge bandwidths, $h(X_i)$; thus they will have a large influence on the estimation at remote (i.e., all other!) data points. This effect is also contrary to the desired one. To overcome this problem, one could either replace the function $\alpha(\cdot)$ in $K$ by a more suitable function $\beta(\cdot)$ or simply leave it out. For simplicity and transparency of the concept we will focus on the last alternative.

## 3.3 Robustness against outlying predictors

### 3.3.1 Soft robustification

In the light of the above discussion, we consider the estimator

$$\hat{m}(x, \alpha) = \hat{a}(x), \tag{3.2}$$

obtained by minimizing

$$\sum_{i=1}^{n} (Y_i - a(x) - b(x)(x - X_i))^2 \, \alpha(X_i) K\left(\frac{x - X_i}{h_n}\right) \tag{3.3}$$

where $\alpha(\cdot)$ is any monotone increasing function of $f(\cdot)$. In order to avoid singularities due to sparse designs, we propose to use kernels with unbounded support in the presence of outlying predictors. In this paper we use Gaussian kernels in all examples.

Note that asymptotically, i.e., for $h_n \longrightarrow 0$ and $nh_n \longrightarrow \infty$, the estimator (3.3) is equivalent to a local linear estimator. In fact, the factor $\alpha(\cdot)$ vanishes in the leading terms of asymptotic bias as well as asymptotic variance expressions as may be deducted from the results presented in Fan & Gijbels (1996). This however is not surprising, since design-adaptivity is one of the major advantages of local linear fitting. The weight function $\alpha(\cdot)$ essentially modifies the influence of the design density, regardless if $\alpha(\cdot)$ depends on $f(\cdot)$ or not. However, asymptotics for horizontal outliers seem not to make much sense, since for $n \longrightarrow \infty$ the data will be arbitrarily dense at any location $x$ for which $f(x) > 0$. Consequently we will not focus on asymptotic considerations in the following.

Normally the function $\alpha(\cdot)$ is unknown and we obtain the estimator $\hat{m}(x, \hat{\alpha})$ by minimizing (3.3) with $\alpha(\cdot)$ replaced by a consistent estimator, $\hat{\alpha}(\cdot)$. A near-at-hand idea is to use $\alpha(\cdot) = f(\cdot)$. We estimate the density by

$$\hat{f}(x) = \frac{1}{ng_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{g_n}\right).$$

To select the bandwidth $g_n$, we choose the modified normal reference bandwidth selector proposed by Silverman (1986), namely

$$g_n = 0.9An^{-1/5},$$

where

$$A = \min(\text{standard deviation}, \text{interquartile range}/1.34). \tag{3.4}$$

We defer the task of how to select the bandwidth $h_n$ to Section 3.5.

As a further improvement, one could imagine to use not the density $f(\cdot)$, but a power $f^k(\cdot)$ with $k > 1$ as the weight function $\alpha(\cdot)$. As we shall see, the larger the exponent, the better is the robustification. However, the exponent cannot increase arbitrarily, since then estimation becomes unstable.

In the sequel we will refer to the method introduced above as *soft robustification*. Under this method, outliers are downweighted but not eliminated. When one is convinced that the outliers do not contain useful information, it might be desirable to eliminate them from the estimation procedure. This approach, called *hard robustification*, will be introduced in the following subsection.

### 3.3.2 Hard robustification

Under soft robustification procedures, outliers still influence estimated values associated to predictors lying in their neighbourhood. To avoid this, one could consider automatically cutting off points associated to estimated density values which fall beyond a certain threshold. This threshold can be calculated data-adaptively by applying an idea similar to that of the normal reference bandwidth selector. In a (very) rough approximation, one can assume

$$\hat{f}(\cdot) \approx \phi_{\mu,A^2}(\cdot),$$

where $\phi_{\mu,A^2}$ denotes the density function of a normal distribution with $\mu = \text{median}(X_1, \ldots, X_n)$ and $A$ as in (3.4). Let $p$ denote the proportion of expected outliers (typically, $p = 0.05$ or $p = 0.01$). Then the required threshold is given by

$$\delta = \phi_{\mu,A^2}(x_{p/2}) = \phi_{\mu,A^2}(\mu + A \cdot z_{p/2}) = \frac{1}{A}\phi(z_{p/2}),$$

where $x_{p/2}$ and $z_{p/2}$ are the $p/2$ quantiles of the distribution of $X$ and of the $N(0,1)$ distribution, respectively. The estimator $\hat{m}(x)$ is now obtained by minimizing

$$\sum_{i=1}^{n} (Y_i - a(x) - b(x)(x - X_i))^2 \, \alpha(X_i) 1_{\{f(X_i) > \delta\}} K\left(\frac{x - X_i}{h_n}\right). \tag{3.5}$$

Surely the question arises whether one can rely on estimation results in areas where the data were downweighted or even cut off. This, however, is a question inherent to any robust method. In particular, when applying soft robustification techniques, we must face the question of whether it is correct to downweight the data on the one hand, i.e., to pretend not to trust the data, but to believe in the estimation results in the same region, on the other hand. Some decision has to be made and we suggest to base it on *areas of confidence*, which can be selected by means of density estimation. Within the areas of confidence, i.e., for all $x$ with $\hat{f}(x) > \delta$, the estimation is considered to be reliable. Outside these areas, the reliability of the estimation procedures is questionable and interpretation of the estimated curve must be taken cautiously.

### 3.3.3 Example

We now apply the proposed methods to a simulated data set, generated by contaminating the underlying function $m(x) = (x - 3)^2$ with Gaussian noise ($\sigma = 0.75$). The predictor values are assumed to be uniformly distributed in the interval [0,6].

In Figure 3.2 we illustrate how the soft robust fit is affected by successively adding outlying predictors. We start without outliers and finish with a cluster of seven outliers. In each case we take the estimated density and the third power of the estimated density as weights.

A similar investigation was carried out under the hard robustification method; the results are depicted in Figure 3.3. The development of the kernel density and the cut off threshold are shown in Figure 3.4. An analysis of these figures leads us to conclude that

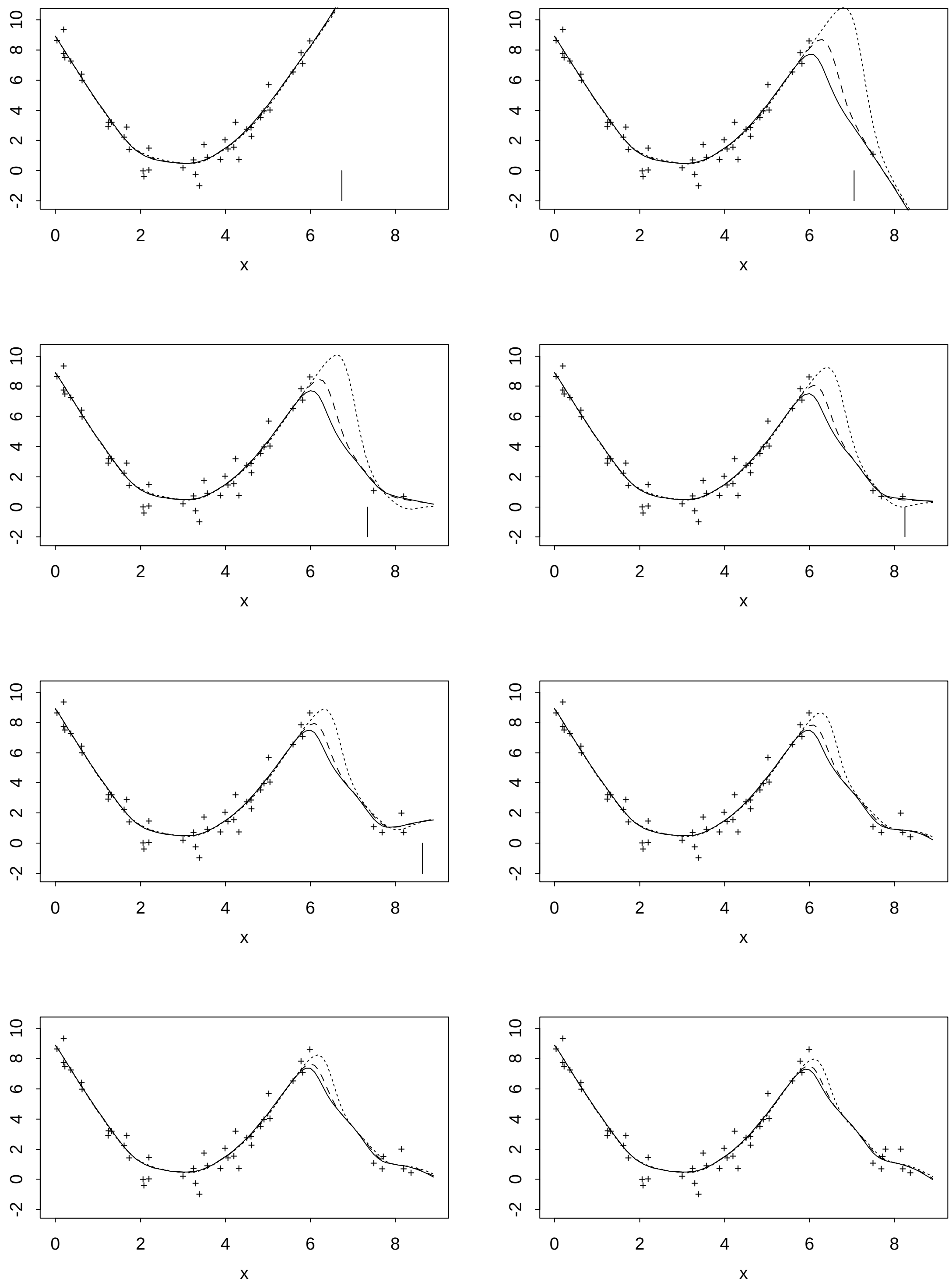



Figure 3.2: Data (+), local linear fit (solid line) and soft robustification by weighting with the density (dashed line) and third power of the density (dotted line) for varying numbers of outliers. Vertical lines indicate the end of the confidence area.

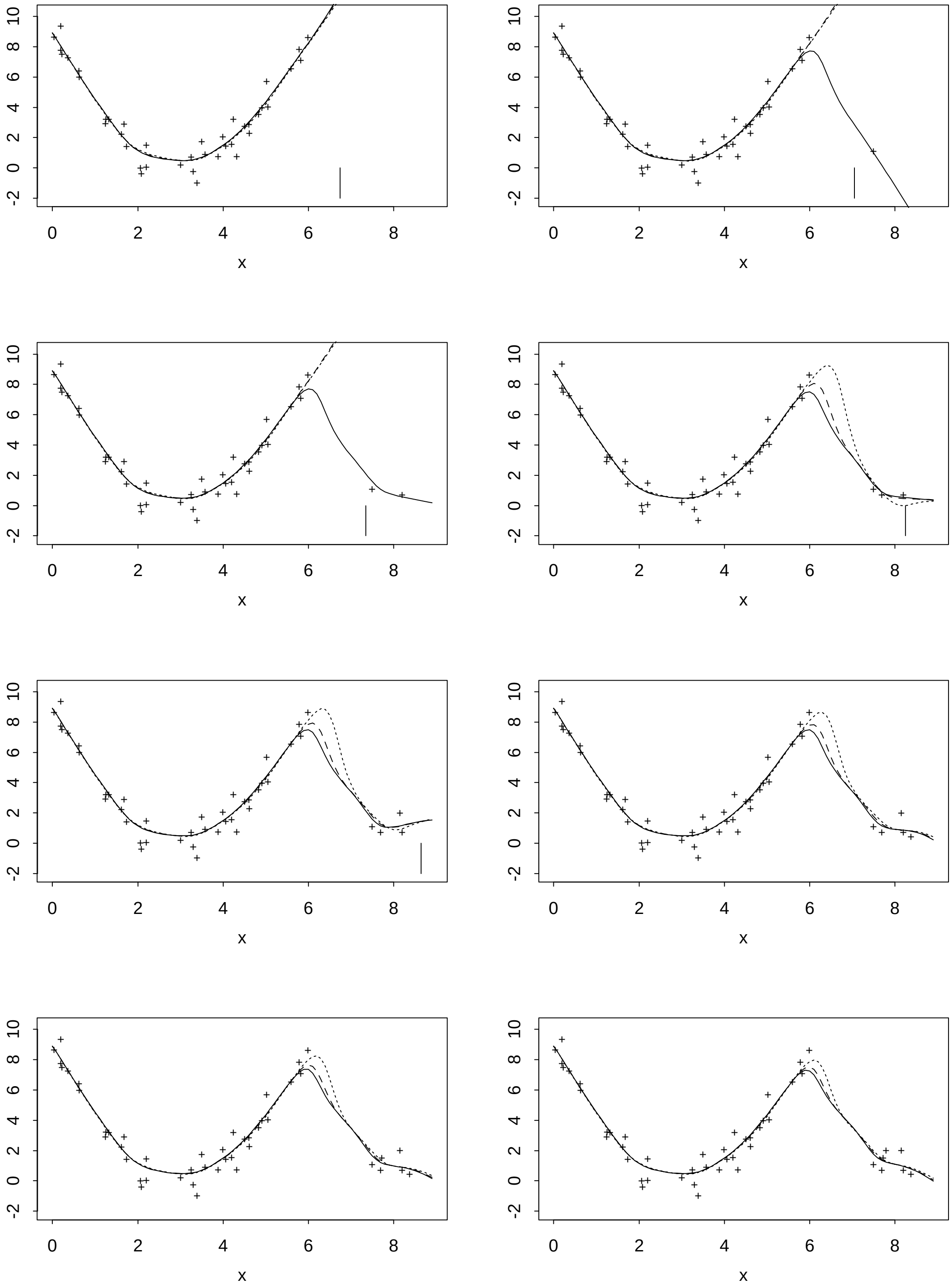



Figure 3.3: Data (+), local linear fit (solid line) and hard robustification by weighting with the density (dashed line) and third power of the density (dotted line) for varying numbers of outliers. Vertical lines indicate the end of the confidence area.

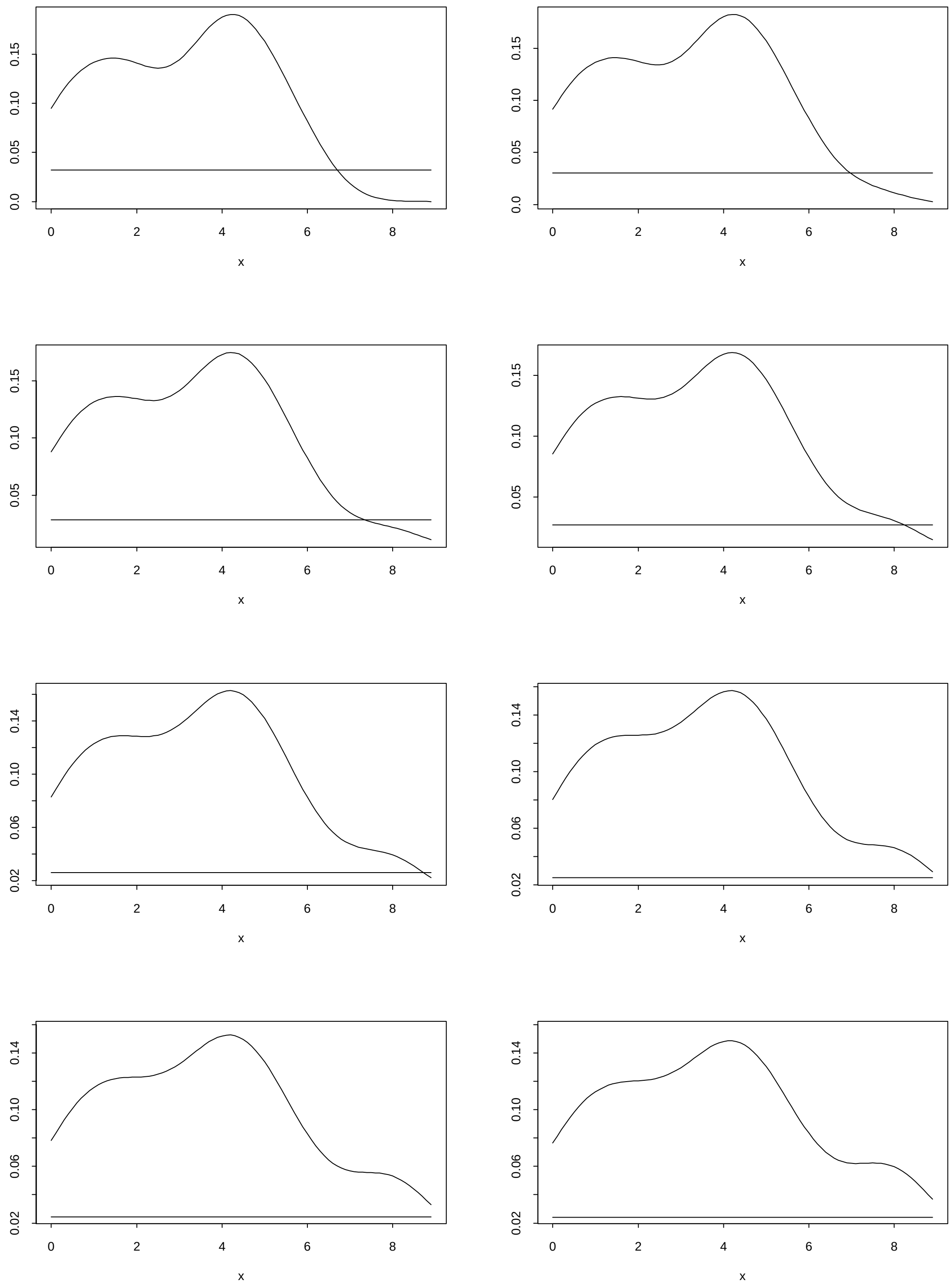



Figure 3.4: Density estimations and cut off thresholds of the data set used in Example 3.3.3.

- The proposed soft robustified fit is obviously more robust to outlying predictors than a local linear fit. The performance of the robustication procedure is thereby better inside the area of confidence than outside.
- Weighting with the third power of the density yields a better robustification effect than weighting with the density itself.
- When only one or two outliers are present, the hard robustification method produces a fit which is completely unaffected by the outliers (see top of Figure 3.3). Like desired, only the outliers are removed from the data set.
- For bigger clusters of outliers soft and hard robustification methods yield the same results since in this case the density does not get cut off. Note that the cut off threshold is decreasing with an increasing number of outliers (see Figure 3.4).
- The bigger the cluster, the smaller is the effect of robustifictaion as expected, since a big cluster is probably not just a group of outliers but rather contains genuine information.
- From Figure 3.4 we may observe that values at the boundary are downweighted in general, even if they are not outlying, due to the smoothing effect of the kernel density estimator. This is a desirable property, when the boundary points are likely to provide spurious information. From a theoretical point of view, one might have a different opinion - see Section 3.6 for a discussion.

In all estimations in this example we used the same global bandwidth $h_2 = 0.6$, motivated by the result of one-sided cross-validation (OSCV) discussed in Section 3.3.5.

Note that the outliers considered in this example could be regarded as outlying predictors as well as outlying responses. Thus methods which robustify against outlying responses should work here as well. In fact, when one outlier is present, the S-Plus function `loess` yields the same effect of a hard robustifiction procedure after two iterations. We chose this example only to demonstrate the effect with a particularly difficult kind of horizontal outliers. In Section 3.3.4 and Section 3.4 we will provide examples where vertical robustification methods fail.

### 3.3.4 Simultaneous robustness for predictor and response variables

Now we show that robust methods for outlying predictors and responses can be combined successfully. We choose the robust LOWESS method of Cleveland (1979), which is one of the most widely used robustification methods. It is implemented in S-Plus.

The data shown in Figure 3.5 were generated by contaminating the underlying function $m(x) = 6\sqrt{x}$ with Gaussian noise ($\sigma = 3$) and the predictors are uniformly distributed in the interval $[0, 6]$. One vertical outlier at point $(2; 25)$ and two horizontal outliers at $(7.5; 11)$ and $(8; 10.5)$ were intentionally added by hand, yielding a total of $n = 51$ data points. Note that the observations with outlying predictors cannot be regarded as outlying responses, since, when compared to the other data points, they are not located at a considerable distance from the function $m$.

Figure 3.5 (top) shows the result of a simple local linear fit and a LOWESS fit after four iterations. The LOWESS fit succeeds to eliminate the influence of the vertical outlier, but fails to handle the horizontal outliers. An illustration that the local linear as well as the LOWESS fit can be robustified against outlying predictors via the soft robustification method (with the estimated density as weight function) is given in the bottom part of Figure 3.5. All the estimation procedures were carried out by means of the S-Plus function `loess` with smoothing parameter equal to 0.45. This smoothing parameter corresponds to the fraction of neighboring data points used in each local fit, since `loess` utilizes by default a nearest-neighbor-bandwidth.

In Section 3.4 we will give another example for simultaneous robust smoothing against horizontal and vertical outliers, in the context of robust M-procedures for generalized additive models.

### 3.3.5 Some notes about bandwidth selection

In principle, any arbitrary local linear (constant or variable) bandwidth selection routine can be applied to select the bandwidth $h_n$. Possible methods to select constant bandwidths are, among others, cross-validation (CV), one-sided cross-validation (OSCV, Hart & Yi, 1998), plug-in methods (Ruppert, Sheather & Wand, 1995), methods based on the AIC (Hurvich, Simonoff & Tsai, 1998) or the RSC criteria (Fan & Gijbels, 1995). For local variable bandwiths, i.e., bandwidths of the form $h(x)$, we may also refer to Fan & Gijbels (1995), and further to Fan, Gijbels, Hu & Huang (1996) or Doksum, Petersen & Samarov (2000).

However, we should point out that the results of bandwidth selection routines can be seriously affected by horizontal outliers. As an example, we demonstrate the effect of outliers on cross-validation and one-sided cross-validation techniques for the simulated data examined in Section 3.3.3. The results of the selection of a bandwidth for a local linear estimator under increasing numbers of outliers is shown

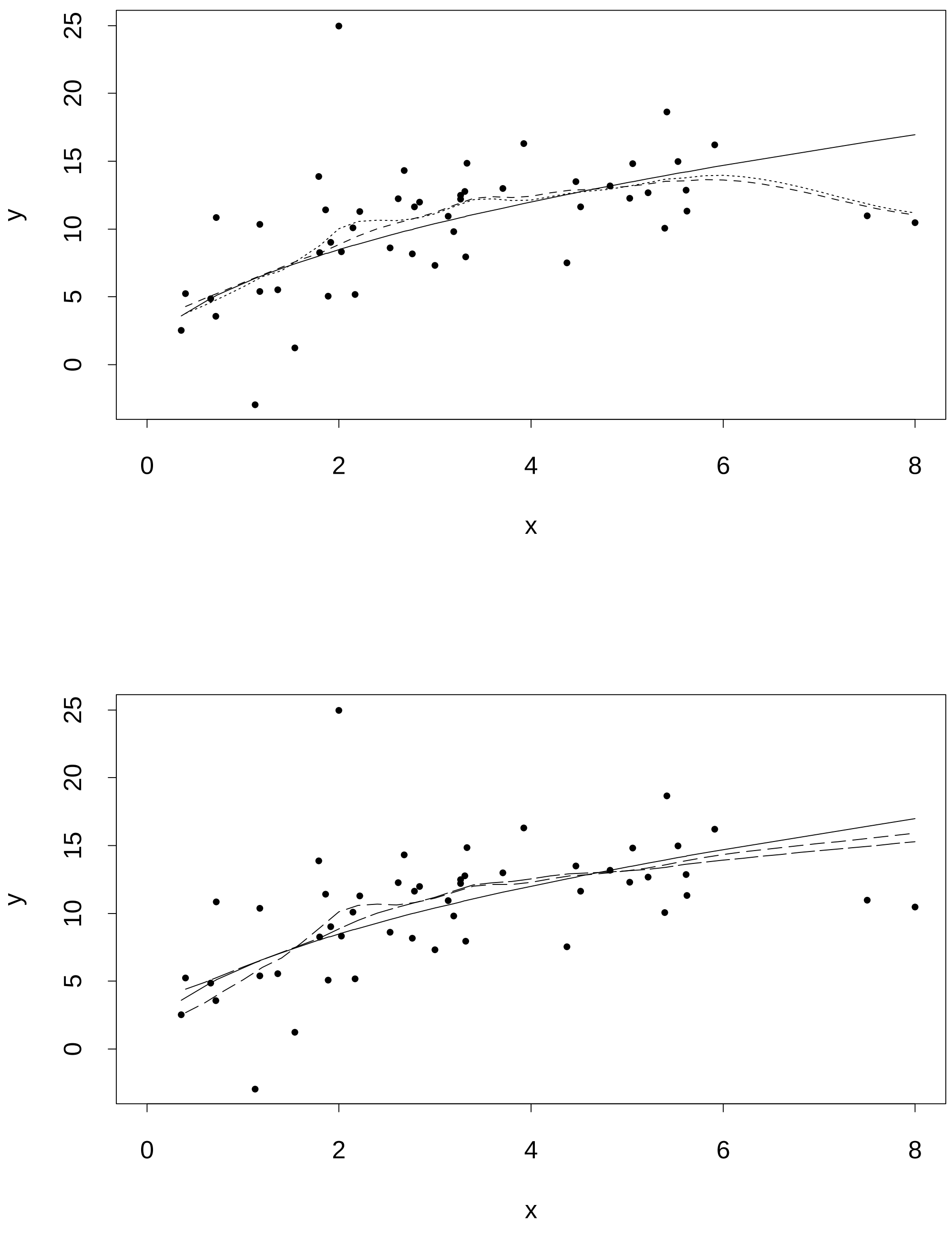



Figure 3.5: Top: Simulated data, underlying function (solid line), local linear (dotted line) and LOWESS fit (dashed line); bottom: Soft robustified local linear (short-dashed) and LOWESS fit (long-dashed).

in Figure 3.6 for each method (CV or OSCV). As may be deducted from the plot, both bandwidths obtained under CV and OSCV are considerably affected by a single outlier.

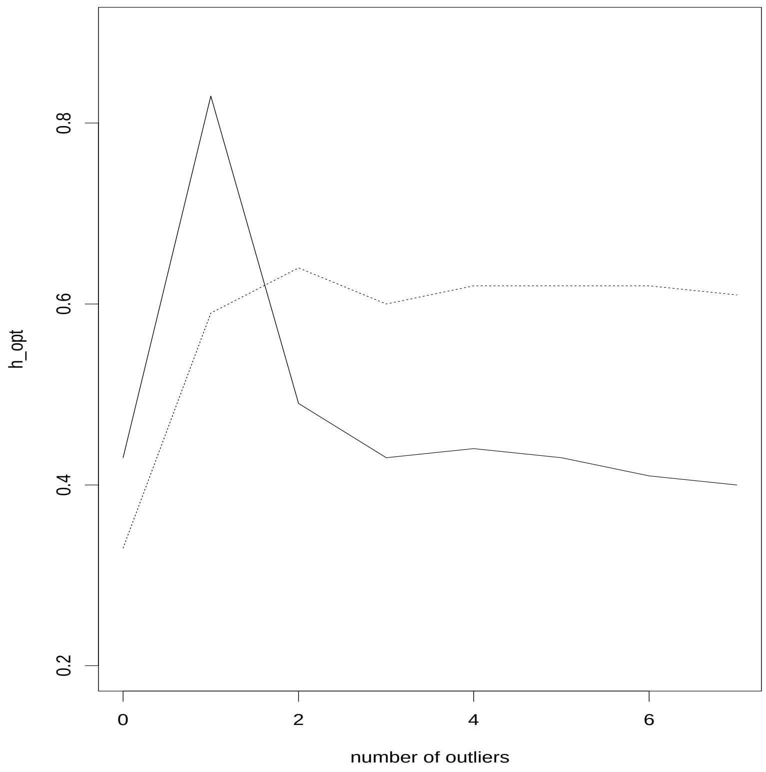



Figure 3.6: Bandwidth selected by CV (solid) and OSCV (dotted) for increasing numbers of outliers in Example 3.3.3.

The corresponding bandwidths are bigger under OSCV than under CV as the number of outliers increases. A similar analysis was conducted for the example of Section 3.3.4; the selected bandwidths under CV and OSCV for none, one and two horizontal outliers are summarized in the following table.

| Selection | Horizontal outliers | | |
|---|---|---|---|
| method | 0 | 1 | 2 |
| *CV* | 1.60 | 1.97 | 1.36 |
| *OSCV* | 1.25 | 1.28 | 1.08 |

Table 3.1: Bandwidths selected by CV and OSCV for Example 3.3.4.

One observes that OSCV yields more stable bandwidth values than CV. The seemingly better robustness of OSCV to outlying predictors is in conformity to other robustness properties of this methodology (Hart & Lee, 2002).

We finally remark that the above results do not change significantly when using a soft robustified estimator instead of a local linear estimator under the CV (OSCV) routines. The problem seems to be intrinsic to the bandwidth selector and not to the smoothing method.

## 3.4 Relative risk curves for respiratory deaths

We now return to the example addressed in the introduction. Following a standard analysis strategy for this type of data as in Schwartz (1994), a generalized additive "core" model including terms to control for trend, days of the week, seasonality, temperature, humidity and non-respiratory deaths was initially fitted.

A scatter plot of the deviance residuals from this "core" model versus the $SO_2$ concentrations is presented in Figure 3.7 (top) along with a LOWESS smoother (dotted line). The resulting curve falls for high concentrations, whereas the soft robustified fit slightly rises.

Regarding the plot, this effect does not seem to be so serious - however the misleading effect of the horizontal outliers becomes much more dramatic when regarding relative risk curves similar to those presented in Singer et al. (2002). The generalized additive model considered there is typical for count data like the ones investigated here. For our purposes it suffices to know that the model may be generally expressed as

$$\ln[E(\text{respiratory death})] = \alpha + \sum_{k=1}^{p-1} f_k(X_k) + f(SO_2)$$

where $X_k$, $k = 1, \ldots, p-1$ denote variables like temperature, humidity, etc. The relative risk of death at a concentration $\mathrm{SO_2(i)}$ of the pollutant $SO_2$ relative to the risk of death at the minimum concentration $\mathrm{SO_2(min)}$ is given by

$$RR(i) = \frac{\mathrm{E(respiratory\ death|SO_2(i))}}{\mathrm{E(respiratory\ death|SO_2(min))}} = \exp[f(\mathrm{SO_2(i)}) - \mathrm{f(SO_2(min))}].$$

In the center portion of Figure 3.7 we show the relative risk curve $(\cdot)$ and its soft robustified counterpart $(+)$. The plot shows a tremendous influence of the horizontal outliers. The unrobustified relative risk curve decreases with increasing pollutant concentration, which is obviously unacceptable. The soft robustified relative risk curve (weighted with the density) behaves as desired. The function $f(SO_2)$ has to be calculated inside the generalized additive model. To account simultaneously for outlying predictors and responses, we apply in both cases robust M procedures for generalized additive models (Hastie & Tibshirani, 1990). The soft robustified smoother is thereby easily plugged into the generalized additive model by using

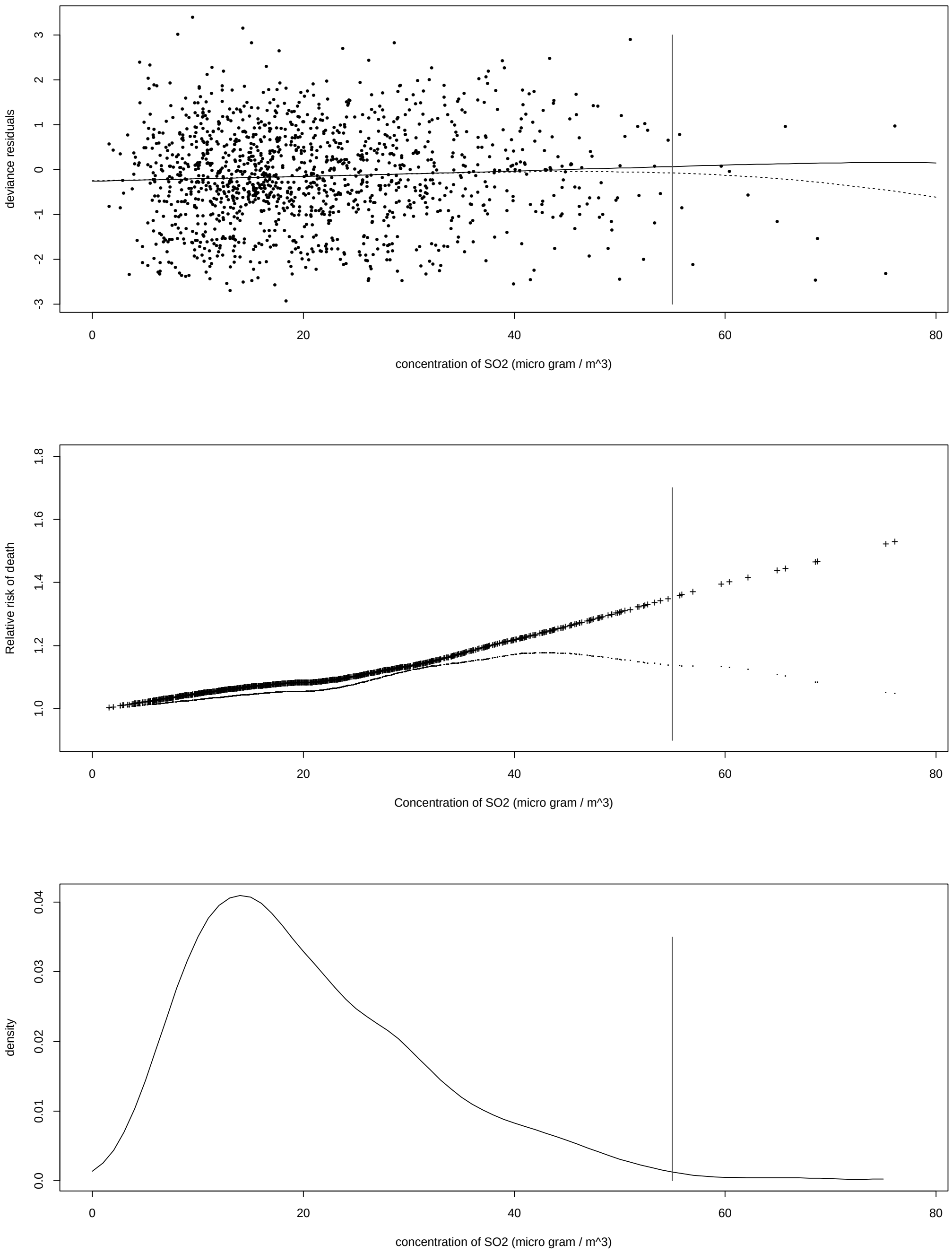



Figure 3.7: Top: Deviance residuals versus $SO_2$ concentration with a LOWESS fit (dotted line) and a soft robustified fit (solid line); middle: Relative risk curves versus $SO_2$ concentration resulting from a local linear $(\cdot)$ and a soft robustified fit $(+)$, each evaluated at all measured values of $SO_2$ concentration; bottom: Kernel density estimation of the $SO_2$ concentration. Vertical lines indicate the end of the area of confidence.

the S-Plus function `gam` applying the interfaces `lo` (LOWESS) or `lf` (LOCFIT). However, it seems that somehow soft robustification and backfitting disturb each other. This problem can be avoided by plugging the density directly into the `gam` instead into `lo` or `lf`. Figure 3.7 (middle) was obtained in this manner.

The form of the kernel density estimation for these data is presented in the bottom portion of Figure 3.7 and suggests that there is no need for a hard robustified version of the local linear fit in this case.

## 3.5 Discussion and Outlook

We showed that local linear and LOWESS smoothers can be robustified against outlying predictors. The main idea is to plug the estimated density into the minimization problem. Such an idea is not restricted to these estimators and can certainly be applied to local estimators in general and the corresponding derivative estimators. We did some further simulations with smoothing splines and this also led to the desired results. In fact, we believe that it is not an exaggeration to claim that any smoothing method which is based on minimization (maximization) of any loss (likelihood) function can be robustified against outlying predictors by applying the concept introduced here. It should also work for multivariate predictors, though more care will be necessary due to the curse of dimensionality when the data gets too sparse, as the estimates otherwise do not exist or are useless. Here the "area of confidence" plays a much more important role.

We feel that there is still necessity for further research in this area. Beyond the topics mentioned above, a challenging task seems to be the problem of bandwidth selection; in particular, we mention the problem of robustness of common bandwidth selection routines to horizontal outliers. Up to now, this task is only rudimentarily treated, even for vertical outliers. In the context of local $L_1$ regression, Wang & Scott (1994) introduced a version of CV with robustness against outlying responses.

## 3.6 Additional asymptotics

In this section we will analyze some theoretical properties of estimator (3.2). Though the application of asymptotic results on data with horizontal outliers might be limited, as already mentioned in Section 3.3.1, we will see that the results are somewhat interesting from a more general point of view. We will derive some asymptotic properties of a generalized version of (3.2), namely of the estimators

$$\hat{m}^{(j)}(x, \alpha) = j!\hat{\beta}_j(x) \qquad (0 \leq j \leq p), \tag{3.7}$$

obtained by minimizing

$$\sum_{i=1}^{n} \left( Y_i - \sum_{j=0}^{p} \beta_j(x)(X_i - x)^j \right)^2 \alpha(X_i) K_h(X_i - x) \tag{3.8}$$

in terms of $\beta(x) = (\beta_0(x), \ldots, \beta_p(x))^T$, where $p$ is any non-negative integer and $\alpha(\cdot)$ is continuous at point $x$. We use the matrices

$$X_x = \begin{pmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{pmatrix}, A = \begin{pmatrix} \alpha(X_1) & & \\ & \ddots & \\ & & \alpha(X_n) \end{pmatrix},$$

and $W_x$, $y$ as in Section 2.1.2. Then the minimization problem (3.8) has the form

$$min_{\beta(x)} \quad (y - X_x\beta(x))^T A W_x (y - X_x \beta(x)),$$

yielding

$$\hat{\beta}(x) = (X_x^T A W_x X_x)^{-1} X_x^T A W_x y.$$

Then $\hat{m}^{(j)}(x) = e_{j+1}^T \hat{\beta}(x)$, where $e_{j+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, with 1 at $(j+1)^{\text{th}}$ position, is an estimator for $m^{(j)}(\cdot)$ at point $x$. The conditional bias can be written as

$$Bias(\hat{\beta}(x)|\mathbb{X}) = (X_x^T A W_x X_x)^{-1} X_x^T A W_x r_x, \tag{3.9}$$

where $r_x = (m(X_1), \ldots, m(X_n))^T - X_x\beta(x)$ is the vector of the residuals of the local approximation and $\mathbb{X}$ denotes the vector of covariates $(X_1, \ldots, X_n)$. The conditional variance is given by

$$Var(\hat{\beta}(x)|\mathbb{X}) = (X_x^T A W_x X_x)^{-1} (X_x^T A^2 \Sigma_x X_x)(X_x^T A W_x X_x)^{-1}, \tag{3.10}$$

with $\Sigma_x = diag(K_h^2(X_i - x)\sigma^2(X_i))$. Recall the notations given in Section 2.1.3, and denote further $\tilde{S}^* = (\nu_{j+l+1})_{0 \leq j,l \leq p}$. We work with the following

**Assumptions:**

(i) The kernel $K$ is a continuous density function.

(ii) $f(x) > 0$, $f(\cdot)$ is continuously differentiable in a neighborhood of $x$;

(iii) $\alpha(x) \neq 0$, $\alpha(\cdot)$ is continuously differentiable in a neighborhood of $x$;

(iv) $\sigma^2(x) > 0$, $\sigma^2(\cdot)$ is continuously differentiable in a neighborhood of $x$;

(v) $m(\cdot)$ is $p + 2$ times continuously differentiable in a neighborhood of $x$;

(vi) The kernel $K$ is symmetric.

We have the following theorem:

**Theorem 3.1.**
*Under assumptions (i) to (v) we get for $h \longrightarrow 0$*

$$Bias(\hat{\beta}(x)|\mathbb{X}) = h^{p+1}H^{-1}\left[\beta_{p+1}S^{-1}c_p + hb^*_\alpha(x) + o_n\right] \tag{3.11}$$

*and*

$$Var(\hat{\beta}(x)|\mathbb{X}) = \frac{\sigma^2(x)}{f(x)nh}H^{-1}\left[S^{-1}S^*S^{-1} + hV^*_\alpha(x) + o_n\right]H^{-1} \tag{3.12}$$

*where $H = diag(1, h, \ldots, h^p)$, $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right)$,*

$$b^*_\alpha(x) = \left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right)\beta_{p+1}(x)\left(S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p\right) + \beta_{p+2}(x)S^{-1}\tilde{c}_p \tag{3.13}$$

*and*

$$\begin{aligned} V^*_\alpha(x) &= \left(2\frac{\sigma'(x)}{\sigma(x)} + 2\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right)S^{-1}\tilde{S}^*S^{-1} - \\ &- \frac{1}{\alpha(x)f(x)}\left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}\right)\cdot\left(S^{-1}\tilde{S}S^{-1}S^*S^{-1} - S^{-1}S^*S^{-1}\tilde{S}S^{-1}\right). \end{aligned}$$

The equations given in this theorem reduce to the expressions provided in Fan, Gijbels, Hu & Huang (1996) in the special case $\alpha(\cdot) \equiv 1$. Note that the leading bias and variance terms are independent of $\alpha(\cdot)$!

In the following we additionally suppose that assumption (vi) holds. Firstly we take a look at the variance (3.12). Note that, independent of $p$ and $j$, the expression $e^T_{j+1}S^{-1}S^*S^{-1}e_{j+1}$ is never trivially zero, while the expressions $e^T_{j+1}S^{-1}\tilde{S}S^{-1}S^*S^{-1}e_{j+1}$, $e^T_{j+1}S^{-1}S^*S^{-1}\tilde{S}S^{-1}e_{j+1}$ and $e^T_{j+1}S^{-1}\tilde{S}^*S^{-1}e_{j+1}$ are always trivially zero for a symmetric kernel. Assuming that $nh \longrightarrow \infty$, it is $o_n = o_P(1)$, and one gets

$$Var(\hat{m}^{(j)}(x)|\mathbb{X}) = e^T_{j+1}S^{-1}S^*S^{-1}e_{j+1}\frac{j!\sigma^2(x)}{f(x)nh^{1+2j}} + o_p\left(\frac{1}{nh^{1+2j}}\right),$$

which is identical to the formula for local polynomial fitting provided in Fan & Gijbels (1996), Theorem 3.1.

Secondly, we consider the bias. For $p - j$ odd, the expressions $e^T_{j+1}S^{-1}\tilde{c}_p$ and $e^T_{j+1}S^{-1}\tilde{S}S^{-1}c_p$ are zero, and we get from (3.11) for $nh \longrightarrow \infty$

$$\begin{aligned} &Bias(\hat{m}^{(j)}(x)|\mathbb{X}) = \\ &= e^T_{j+1}S^{-1}c_p\frac{j!}{(p+1)!}m^{(p+1)}(x)h^{p+1-j} + o_P(h^{p+1-j}), \end{aligned} \tag{3.14}$$

which is again the same as in Fan & Gijbels (1996), Theorem 3.1, for standard local polynomial fitting. Let additionally $nh^3 \longrightarrow \infty$ and thus $o_n = o_P(h)$. For $p-j$ even we have $e_{j+1}^T S^{-1} c_p = 0$, and we arrive at

$$\begin{aligned} & Bias(\hat{m}^{(j)}(x)|\mathbb{X}) = \\ & = e_{j+1}^T \frac{j!}{(p+1)!} \left[ \left( \frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)} \right) \left( S^{-1}\tilde{c}_p - S^{-1}\tilde{S}S^{-1}c_p \right) m^{(p+1)}(x) + \right. \\ & \quad \left. + S^{-1}\tilde{c}_p \frac{m^{(p+2)}(x)}{p+2} \right] h^{p+2-j} + \\ & \quad + o_P(h^{p+2-j}) \end{aligned} \tag{3.15}$$

This expression is somewhat interesting, because it reveals that with a suitable setting of $\alpha(x)$ the asymptotic bias may be reduced. Note that the augend in the squared bracket in (3.15) vanishes for

$$\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)} = 0,$$

and this differential equation is solved for

$$\alpha_{opt}(x) = c\frac{1}{f(x)}, \tag{3.16}$$

with $c \in \mathbb{R} \setminus \{0\}$. This is a quite surprising result, because we proposed just the opposite in the previous sections, namely to set $\alpha(x) = f(x)$ in order to weight down outlying predictors. The explanation for this apparent contradiction might be that outlying predictors are a finite sample problem, where the results of asymptotic calculations don't apply. However, it might be a bit too easy to make the asymptotics responsible for this. From a theoretical point of view, a horizontal outlier is nothing that should be discarded. In contrary, when fitting at an endpoint, the boundary point is even *"the most informative observation"* (Hastie & Loader, 1993b) for reducing the bias at the boundary. This, however, implies that the boundary point undoubtedly traces from the assumed underlying model, i.e. one assumes that the information at the boundary is as reliable as in the interior. This condition might be injured in some cases, especially when the boundary point is far away from the interior, and for these cases the methodology introduced in the previous sections was designed. Thus, we do not believe that the asymptotic result is in fact a contradiction; it only reflects another point of view of the problem.

**Proof of Theorem 3.1**

This proof is kept shortly since it mainly follows the lines of the corresponding proof for local polynomial fitting, see Fan, Gijbels, Hu & Huang (1996), or the proof of

Theorem 2.2. Let $w_i = K_h(X_i - x)$ and

$$r_{n,j} = \sum_{i=1}^{n} \alpha(X_i) w_i (X_i - x)^j; \qquad R_n = (r_{n,j+l})_{0 \le j,l \le p};$$
$$r^*_{n,j} = \sum_{i=1}^{n} \alpha^2(X_i)\sigma^2(X_i) w_i (X_i - x)^j; \quad R^*_n = (r_{n,j+l})_{0 \le j,l \le p}.$$

Then $R_n = X_x^T A W_x X_x$ and $R^*_n = X_x^T A^2 \Sigma_x X_x$.

*Bias:*

The use of standard asymptotics reveals that

$$r_{n,j} = nh^j (f_\alpha(x)\mu_j + hf'_\alpha(x)\mu_{j+1} + o_n), \tag{3.17}$$

where $f_\alpha(x) = \alpha(x)f(x)$ and $o_n = o_P(h) + O_P\left(\frac{1}{\sqrt{nh}}\right)$, and thus

$$R_n = nH[f_\alpha(x)S + hf'_\alpha(x)\tilde{S} + o_n]H \tag{3.18}$$

holds. Then, using Taylor's expansion and equation (3.9), we get

$$Bias(\hat{\beta}(x)|\mathbb{X}) = R_n^{-1}\left[\beta_{p+1}(x)d_n + \beta_{p+2}(x)\tilde{d}_n + o_P(\tilde{d}_n)\right], \tag{3.19}$$

where $d_n = (r_{n,p+1}, \ldots, r_{n,2p+1})^T$ and $\tilde{d}_n = (r_{n,p+2}, \ldots, r_{n,2p+2})^T$. We use the fact that $(B + hC)^{-1} = B^{-1} - hB^{-1}CB^{-1} + O(h^2)$ to calculate

$$R_n^{-1} = \frac{1}{n}H^{-1}\left[\frac{1}{f_\alpha(x)}S^{-1} - h\frac{f'_\alpha(x)}{f^2_\alpha(x)} + o_n\right]H^{-1}. \tag{3.20}$$

Plugging (3.20) into (3.19), and substituting (3.17) into the vectors $d_n$ and $\tilde{d}_n$, yields (3.11) via some simple matrix algebra, taking into account that

$$\frac{f'_\alpha(x)}{f_\alpha(x)} = \frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)}.$$

*Variance:*

Similar to (3.18) we find that

$$R^*_n = \frac{n}{h}H[s_\alpha(x)S^* + hs'_\alpha(x)\tilde{S}^* + o_n]H, \tag{3.21}$$

where $s_\alpha(x) = \sigma^2(x)\alpha^2(x)f(x)$. By substituting (3.21) and (3.20) in

$$Var(\hat{\beta}(x)|\mathbb{X}) = R_n^{-1}R^*_nR_n^{-1}$$

we derive (3.12) by applying matrix algebra.

## 3.7 Relation to the Horvitz-Thompson estimator

In the previous section we showed that the bias is reduced asymptotically if the observations are downweighted with their reciprocal density, at least for odd values of $p-j$. This implies that outliers and sparse data regions are upweighted, whereas the data in dense regions is downweighted. The concept of weighting up unlikely data is not new. Indeed, the phenomenon observed in the previous section seems to have a counterpart in sampling theory.

Assume a population $U$ consisting of $N$ units $u_1, \ldots, u_N$. A sample of size $n$ is to be drawn without replacement using arbitrary probabilites of selection for each draw. Suppose that a characteristic $X$ for the $n$ units is to be measured. We denote by $X_j\,(j = 1, \ldots, N)$ the value of $X$ belonging to $u_j$, and by $x_i\,(i = 1, \ldots, n)$ the value of the $i-$th *selected* unit. The objective is to estimate the population total of $X$, i.e.

$$T = \sum_{j=1}^{N} X_j.$$

Let us consider the class of estimators $\hat{T} = \sum_{i=1}^{n} \beta_{ij} x_i$, where $\beta_{ij}\,(j = 1, \ldots, N)$ is a constant to be used as a weight for the $j-$th unit whenever it is selected for the sample ($\beta_{ij}$ does not depend on $i$ itself, the subscript $i$ only indicates that it is related to the $i-$th draw). Let us assume that units $u_1, \ldots, u_n$ (suitably renamed) are selected. Horvitz & Thompson (1952) show that the only unbiased linear estimator in that class is given by

$$\hat{T} = \sum_{i=1}^{n} \frac{x_i}{P(u_i)},$$

where $P(u_i)$ is the probability that unit $u_i$ is selected in any of the $n$ draws. Thus, in other words, the estimation is best when the observations are weighted with the inverse selection probability. Furthermore, they state that if

$$P(u_j) = nX_j/T, \tag{3.22}$$

the estimator $\hat{T}$ has zero variance and the sampling will be optimal.

In the regression setting, an underlying density of the independent variable may be considered as the selection probability distribution. We realized in (3.16) that the estimation is best when the observations are weighted with the inverse density. This is now a message quite similar to that of Horvitz and Thompson.

Also for Horvitz-Thompson's estimator a paradox between theory and its practical application has been observed, firstly by Basu (1971) in his famous elephant fable:

A circus owner plans to ship 50 adult elephants and therefore needs a rough estimate of their total weight. As weighing elephants is not so easy, the owner intuitively plans to weigh only one elephant and to multiply the result with 50. However, to decide which elephant should be weighed, he (unfortunately) consults the circus statistician, who assigns a selection probability of 99/100 to a previously determined elephant ("Samba") which presumably has about the average weight of the herd. All other elephants obtain the weight 1/4900, including the elephant "Jumbo" who is biggest of all. If Samba was now selected, its weight would have to be multiplied with 100/99 according to Horvitz-Thompson, and if Jumbo was selected, his large weight would even have to be multiplied with 4900 to get the "best linear unbiased estimator" of the total weight. Certainly, after having given this advice, the circus statistician is sacked.

In a recent book by Brewer (2002), subtitled "Weighing Basu's elephants", this problem is examined in more detail. We will not pursue this topic further at this place, but finish with the statement that in sampling theory and local smoothing similar theoretical results lead to similar phenomena. We note furthermore that in Basu's fable the property (3.22) was completely ignored. These observations suggests that generally care has to be taken when applying a theoretical *bias*-minimizing criterion, as given by (3.16) or Horvitz and Thompson.

## Final remarks

Sections 3.1 to 3.5 of this chapter are joint work with Carmen D. S. de André and Julio M. Singer (Universidade de São Paulo). That part of this chapter is accepted in *Environmetrics*.

# Chapter 4

# Online Monitoring with Local Smoothing Methods and Adaptive Ridging

## 4.1 Introduction

A considerable number of papers in the last years focussed on modelling and testing of edges and jumps in smooth functions, see e.g. McDonald & Owen (1986), Hall & Titterington (1992), Chu, Glad, Godtliebsen & Marron (1998), Müller & Stadtmüller (1999). These methods are however preferably or exclusively designed for data which are analyzed "offline". This means the entire data set is available for the analysis. In contrast, "online" monitoring is required if observations arrive successively in time. Then at each time point a decision is required whether a jump or edge has occurred. In this paper we will extend some of the "offline" tools above for monitoring data online. We develop an online test checking for breakpoints.

The analysis of data occurring online is an important issue in various fields of science and industry. This includes quality control management, time series in finance or online monitoring of clinical information systems. A general overview of existing procedures for online monitoring is found in Basseville & Nikiforov (1993). The use of online methods in clinical information systems has been focussed e.g. by Daumer & Falk (1998), who make use of a Kalman filter to detect jumps and thresholds in the (online) ECG profile of a patient after surgery. Imhoff & Bauer (1996) and Bauer, Gather & Imhoff (1999) make use of a time series approach for online monitoring while Daumer (1997) uses an adaptive control chart based on moving averages. In all these papers the general focus is to detect sudden structural changes in order to give alarm.

The general problem for online monitoring we are considering here can be described as follows. Assume that at time-point $t$ the measurement $y_t$ is observed. It is assumed that $y_t$ follows the stochastic model

$$y_t = \mu(t) + \varepsilon_t \tag{4.1}$$

where $\mu(t)$ is the mean function in time, which possibly also depends on other covariates, and $\varepsilon_t$ is a random noise, which is allowed to be correlated with previous observations. Both, $y_t$ and hence $\varepsilon_t$ are allowed to be multivariate, but we restrict to the univariate case here. Based on the information available at time-point $t$, i.e. based on $y_1, \ldots, y_t$, it is to decide whether $\mu(t)$ has a breakpoint at time-point $t$. A breakpoint here means that $\mu(t)$ is discontinuous, i.e. there is a jump at $t$, or $\mu(t)$ has a discontinuous first derivative, i.e. there is an edge or sharp bend at $t$. Online monitoring of the data should give alarm if a breakpoint occurs at time-point $t$.

A convenient approach is to compare the observed value $y_t$ with a predictor $\hat{y}_t$. Alarm is given if $y_t$ differs from the predictor by more than the threshold $A_t$, say, i.e. if

$$|y_t - \hat{y}_t| > A_t. \tag{4.2}$$

The threshold $A_t$ is thereby chosen such that sensitivity of the alarm rule is achieved while the probability of false alarms is small. The prediction $\hat{y}_t$ is calculated from previous values $y_{t-h}, \ldots y_{t-1}$, with $h$ as time lag. Daumer (1999) suggests to calculate $\hat{y}_t$ by a running mean calculated from $y_{t-h}, \ldots y_{t-d}$, where $d$ is a second time lag with $1 < d < h$. Hence observations in the near past are left unconsidered. The time lag $d$ serves as delay for the running mean and Daumer shows that for $d > 1$ the alarm rule (4.2) improves its performance compared to taking $d = 1$. In this paper we apply more sophisticated smoothing techniques instead of a simple running mean. We make use of local polynomial fitting (see e.g. Fan & Gijbels, 1996) which reacts better on structural changes and moreover can cope for smooth shifts, unlike the running mean.

Considering (4.2) it becomes obvious, that the alarm rule basically depends on the value of $y_t$. This in turn implies a high variance of the procedure. We therefore replace $y_t$ in (4.2) by a smooth estimate of $\mu(t)$. In the same way we replace the predictor by a second smooth estimate. This means we consider the alarm rule

$$|\hat{\mu}_1(t) - \hat{\mu}_2(t)| \quad > \quad A_t \tag{4.3}$$

where $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$ are two estimates of $\mu(t)$. The first estimate $\hat{\mu}_1(t)$ is thereby calculated as long term estimate from $y_{t-h_1}, \ldots, y_t$ while $\hat{\mu}_2(t)$ is a short term smoother obtained from $y_{t-h_2}, \ldots, y_t$, where $h_2 < h_1$. The major difference of (4.3) compared to (4.2) is, that we do not compare the current observation with its predictor, but we compare two estimates of the mean function. The basic idea behind this is that if $\mu(t)$ has a jump or a sharp bend at $t$, the long term estimate $\hat{\mu}_1(t)$ and the short term estimate $\hat{\mu}_2(t)$ will essentially differ. If in contrast $\mu(t)$ is smooth, both smooth estimates will basically be the same. Hence the alarm rule (4.3) can be seen as smooth test statistic, where large values indicate a violation in the smoothness of $\mu(t)$.

The bandwidth $h_2$ which is chosen for the short term estimate mainly determines the speed of reaction of the alarm. Taking a large value for $h_2$, the reaction time and the specitivity of the alarm rule increases while the variance of the alarm rule decreases so that false alarms are less probable. Using a small bandwidth $h_2$ on the other hand improves the reaction time and the sensitivity of the alarm rule (4.3) but the variability increases. The second tuning parameter $h_1$ decides in which depth the method is searching for breakpoints. For small values of $h_1$ mainly short term changes will be detected, while with a large value of $h_1$ the focus is on detecting long term breaks of the structure of the time series. Beside the choice of the two bandwidths $h_1$ and $h_2$ the fixing of the threshold $A_t$ is required which however results from simple variance calculations.

The choice of the applied smoothing method is thereby essential. Generally speaking, smoothing methods are weak in detecting jumps since they smooth over edges or jumps. Once a jump occurs and is detected, it is therefore necessary that the smooth estimates adjust quickly for the new level or shift. It is well known that local linear smoothing and local constant smoothing, which is a simple running mean, react quite differently at the boundary of the support points. Note that by definition, the online estimates are calculated at the boundary. We will combine both estimates using a ridge regression, as suggested in Seifert & Gasser (2000) for "offline" analysis. The ridge regressor thereby results as weighted mean of the local linear and the local constant estimate.

## 4.2 Local linear smoothing and breakpoint detection

We calculate the long term estimate by fitting a local linear model to the data pairs $(t-i, y_{t-i})$ for $i = 0, 1, \ldots, h_1$. Let therefore $K_1(\cdot)$ denote a kernel function with support $[0, h_1]$. An example is found by taking $K_1(\cdot)$ as the truncated normal density with mean $h_1/2$ and variance $(h_1/4)^2$. The estimate $\hat{\mu}_1(t)$ is then obtained by fitting a weighted linear model using the kernel $K_1(\cdot)$ as weight function. It is not difficult to show that the resulting estimate is the weighted mean

$$\hat{\mu}_1(t) = \sum_{i=0}^{h_1} w_{i,1} y_{t-i} \tag{4.4}$$

with weights

$$\begin{aligned} w_{i,1} &= (1,0)K_1(i)\left(\sum_{j=0}^{h_1} K_1(j)\begin{pmatrix}1\\-j\end{pmatrix}(1,-j)\right)^{-1}\begin{pmatrix}1\\-i\end{pmatrix} \\ &= \frac{K_1(i)(S_{h_1,2}+iS_{h_1,1})}{S_{h_1,0}S_{h_1,2}-S_{h_1,1}^2} \end{aligned} \tag{4.5}$$

where $S_{h_1,j} = \sum_{i=0}^{h_1} K_1(i)(-i)^j$ for $j = 0,1,2$. It is important to note that the weights do not change in $t$ and hence they can be calculated once and no updating is required.

In the same fashion one obtains the short term estimate $\hat{\mu}_2(t)$ as local linear fit to the data $(t-i, y_{t-i})$, $i = 0,\ldots,h_2$. Let therefore $K_2(\cdot)$ be a kernel density with support $[0,h_2]$, e.g. a half sided normal distribution.

Our experience is that the particular choice of the kernel functions $K_1$ and $K_2$ is not very crucial as long as they follow the setting shown in Figure 4.1 and are not truncated too roughly on the left hand side. For example, setting $K_1(\cdot)$ as the uniform kernel with $K_1(x) = 1/h_1$ for $x \in [0,h_1]$ might cause an artificial alarm signal when the left border of the support of $K_1$ is passing a jump or bend which should already have been detected the time span $h_1$ before.

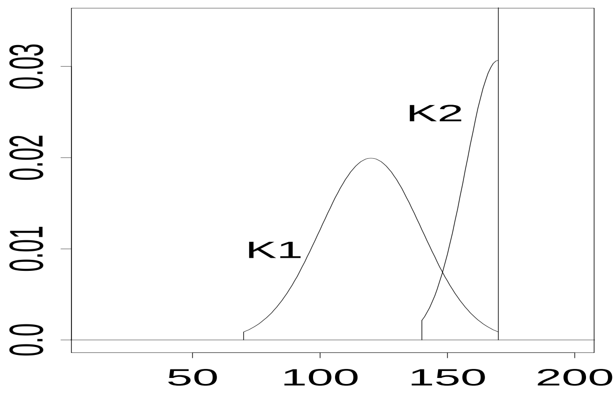



Figure 4.1: Kernel positions for an estimate at t=170.

For $i = 0,\ldots,h_2$ we set

$$w_{i,2} = \frac{K_2(i)(S_{h_2,2}+iS_{h_2,1})}{S_{h_2,0}S_{h_2,2}-S_{h_2,1}^2}$$

with $S_{h_2,j} = \sum_{i=0}^{h_2} K_2(i)(-i)^j$ for $j = 0,1,2$, while $w_{i,2} = 0$ for $i > h_2$. The short term estimate is then available through

$$\hat{\mu}_2(t) \;=\; \sum_{i=0}^{h_1} w_{i,2} y_{t-i} = \sum_{i=0}^{h_2} w_{i,2} y_{t-i}. \tag{4.6}$$

The weights are for convenience constructed such that the vectors $\boldsymbol{w}_1 = (w_{0,1}, \ldots, w_{h_1,1})^T$ and $\boldsymbol{w}_2 = (w_{0,2}, \ldots, w_{h_1,2})^T$ have equal length. We now combine the two estimates in the alarm rule (4.3). If $\mu(t)$ is smooth in $[t-h_1, t]$ the bias of $\hat{\mu}_1(t) - \hat{\mu}_2(t)$ can be approximated by

$$\begin{aligned} E\{\hat{\mu}_1(t) - \hat{\mu}_2(t)\} &= \mu''(t)/2 \sum_{i=0}^{h_1} (w_{i,1} - w_{i,2})(-i)^2 + \ldots \\ &= \mu''(t)/2 \left( \frac{S_{h_1,2}^2 - S_{h_1,1} S_{h_1,3}}{S_{h_1,0} S_{h_1,2} - S_{h_1,1}^2} - \frac{S_{h_2,2}^2 - S_{h_2,1} S_{h_2,3}}{S_{h_2,0} S_{h_2,2} - S_{h_2,1}^2} \right) + \ldots \end{aligned} \tag{4.7}$$

The approximation is based on a simple Taylor series and is heuristic in nature. Rigorous quantification of the bias would require a number of assumptions to hold, most of which are not met in practice. For instance in standard smoothing literature theoretical developments are based on the assumption that values $t$ are getting infinitely dense. In our online scenario however we assume that time $t$ is realised on an equidistant grid. For this reason we do not investigate the bias from a theoretical point of view. However, considering (4.7) shows that the bias gets large if $\mu''(t)$ is large, which is the case if $\mu(\cdot)$ rapidly changes its direction at $t$. As extreme case this results in a jump or sharp bend. The quantity $\hat{\mu}_1(t) - \hat{\mu}_2(t)$ in the alarm rule (4.3) can therefore be seen as an empirical estimate for the second order derivative of $\mu(\cdot)$. If the resulting value is large in absolute terms the resulting function is likely to be rough or unsmooth in $t$.

The choice of the threshold $A_t$ in (4.3) requires the estimation of the variability of $\hat{\mu}_1(t) - \hat{\mu}_2(t)$. We rewrite $\hat{\mu}_1(t) - \hat{\mu}_2(t)$ as

$$\hat{\mu}_1(t) - \hat{\mu}_2(t) = \sum_{i=0}^{h_1} w_i y_{t-i} \tag{4.8}$$

where $w_i = w_{i,1} - w_{i,2}$. Assuming local stationarity, simple calculation leads to

$$var\{\hat{\mu}_1(t) - \hat{\mu}_2(t)\} \;=\; \sum_{i=0}^{h_1} w_i^2 \gamma(0) + 2 \sum_{i=0}^{h_1} \sum_{j>i}^{h_1} w_i w_j \gamma(i-j)$$

where $\gamma(d) = cov(y_{l-d}, y_l)$ is the covariance function and $\gamma(0) = var(y_l)$ with $l = t-h_1, \ldots, t$.

Estimation of (4.9) can then be done by the simple moment based estimate (see Brockwell & Davis, 1987)

$$\hat{\gamma}(d) = \frac{c_d}{h+1-d} \sum_{i=t-h}^{t-d} \{y_i - \hat{\mu}_2(i)\}\{y_{i+d} - \hat{\mu}_2(i+d)\}. \tag{4.9}$$

where $h > h_1$ is some timelag expressing the local stationarity of the process. In the following we provide some heuristics to find a suitable value of $c_d$. Assuming $y_l$, $l = 1, 2, \ldots$ to be independent one finds for $d = 0$ in (4.9) by taking expectation

$$E\left[\sum_{i=t-h}^{t} \{y_i - \hat{\mu}_2(i)\}^2\right] = \gamma(0)(h+1)\left(1 - 2w_{0,2} + \sum_{j=0}^{h_2} w_{j,2}^2\right).$$

Skipping the assumption of independence, one gets for $0 < d \leq h_2$

$$\begin{aligned} E\left[\sum_{i=t-h}^{t-d} \{y_i - \hat{\mu}_2(i)\}\{y_{i+d} - \hat{\mu}_2(i+d)\}\right] &= \\ = (h+1-d)\left[\gamma(d)\left(1 - 2w_{0,2} + \sum_{j=0}^{h_2} w_{j,2}^2\right) + \ldots\right] \end{aligned}$$

with $\ldots$ standing for a collection of terms build from $\gamma(i)$, $i \neq d$. Detailed consideration shows that the terms not explicitly listed are of order $1/h_2$ and for simplicity of calculations they are neglected subsequently. This suggests to set $c_d = 1/(1 - 2w_{0,2} + \sum_{j=0}^{h_2} w_{j,2}^2)$ for all $d = 0, \ldots, h_2$, to achieve a bias reduced variance estimate. Usually the constant $c_d$ obtained in this manner is slightly bigger than 1. For $d > h_2$, we suggest to set $c_d = 0$ and thus $\gamma(d) = 0$.

The computation of (4.9) in every timepoint can be accelerated by making use of the following iterative update scheme. Let $\boldsymbol{d}_{t,h} = \{y_t - \hat{\mu}_2(t), y_{t-1} - \hat{\mu}_2(t-1), \ldots, y_{t-h} - \hat{\mu}_2(t-h)\}^T$ and

$$\boldsymbol{D}_{t,h} = \begin{pmatrix} \frac{\boldsymbol{d}_{t,h}}{h+1} & \begin{matrix} \mathbf{0}_1 \\ \frac{\boldsymbol{d}_{t,h-1}}{h} \end{matrix} & \begin{matrix} \cdots \\ \cdots \end{matrix} & \begin{matrix} \mathbf{0}_{h_1} \\ \frac{\boldsymbol{d}_{t,h-h_1}}{h-h_1+1} \end{matrix} \end{pmatrix}$$

where $\mathbf{0}_d$ are column vectors of zeros with length $d$. The covariance vector at time point $t$ can then be estimated by $\hat{\boldsymbol{\gamma}}_t = \boldsymbol{d}_{t,h}^T \boldsymbol{D}_{t,h} \boldsymbol{C}$, where $\hat{\boldsymbol{\gamma}}_t = \{\hat{\gamma}_t(0), \ldots, \hat{\gamma}_t(h_1)\}$, $\boldsymbol{C} = \mathrm{diag}(c_i)_{0 \leq i \leq h_1}$ and the subscript $t$ indicates that information available at timepoint $t$ is used. Simple matrix algebra (see appendix) provides the approximative recursive formula

$$\hat{\boldsymbol{\gamma}}_{t+1} \approx \frac{1}{h+1}(y_{t+1} - \hat{\mu}_2(t+1))\boldsymbol{d}_{t+1,h_1}^T \boldsymbol{C} + \frac{h}{h+1}\hat{\boldsymbol{\gamma}}_t. \tag{4.10}$$

Defining the covariance matrix $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}]_{ij} = [\gamma(|i-j|)]_{ij}$ for $i, j = 0, \ldots, h_1$ one gets the variance estimate

$$\widehat{\mathrm{var}}(\hat{\mu}_1(t) - \hat{\mu}_2(t)) = \boldsymbol{w}\hat{\boldsymbol{\Gamma}}\boldsymbol{w}^T \tag{4.11}$$

where $\boldsymbol{w} = (w_0, \dots w_{h_1})$ and $\hat{\boldsymbol{\Gamma}}$ is a plug in estimate of $\boldsymbol{\Gamma}$. This suggests the alarm threshold

$$A_t = a\sqrt{\hat{\text{var}}(\hat{\mu}_1(t) - \hat{\mu}_2(t))} \tag{4.12}$$

where $a$ is chosen such that the alarm rule is sensitive and false alarms are less probable. This provides a simple test on presence of a breakpoint: We reject the hypothesis "No breakpoint at time t" if

$$|T_t| = \left|\frac{\hat{\mu}_1(t) - \hat{\mu}_2(t)}{\sqrt{\hat{\text{var}}(\hat{\mu}_1(t) - \hat{\mu}_2(t))}}\right| > u_{1-\frac{\alpha}{2}} \tag{4.13}$$

where $\alpha$ is the error probability and $u_{1-\frac{\alpha}{2}}$ is the $1 - \alpha/2$ quantile of the $N(0,1)$-distribution.

## 4.3 Practical adjustments

### 4.3.1 Ridging

In Section 4.2 we suggested to use local linear fitting to calculate the long and short term estimates. All estimates are calculated at the boundary, where local polynomial smoothers are known to be more variable than local constant smoothers. In terms of variability one therefore has to consider the Nadaraya-Watson estimate

$$\hat{\mu}_{1,NW}(t) \;=\; \sum_{i=0}^{h_1} w_{i,1,NW}\, y_{t-i} \tag{4.14}$$

with $w_{i,1,NW} = K_1(i)/S_{h_1,0}$ as a competitor to $\hat{\mu}_1(t)$.

Figure 4.2 shows the behavior of the local estimates when used with the alarm rule (4.3) for independent Gaussian errors. Both estimates detect the jump at 200 and the bend at 400, but the bend at 600 is only found by the local linear fit, since this adopts the inclination. Hence, one should use a local linear fit when there is a slope in the data while local constant appears more appropriate if the data are flat. Considering the local linear fit in more depth uncovers a further drawback. The local linear fit adjusts for the model violation shortly after the jump, while the local constant fit reacts delayed. Thereafter however the local linear fit over-steers the shift and the local constant gets superior. Figure 4.3 gives a tutorial to demonstrate this point. In order to balance local linear and local constant fitting we propose to use ridging as suggested in Seifert & Gasser (2000). This means we replace the long term estimate by

$$\hat{\mu}_{1,ridge}(t) = \lambda_t \hat{\mu}_{1,NW}(t) + (1 - \lambda_t)\hat{\mu}_1(t) \tag{4.15}$$

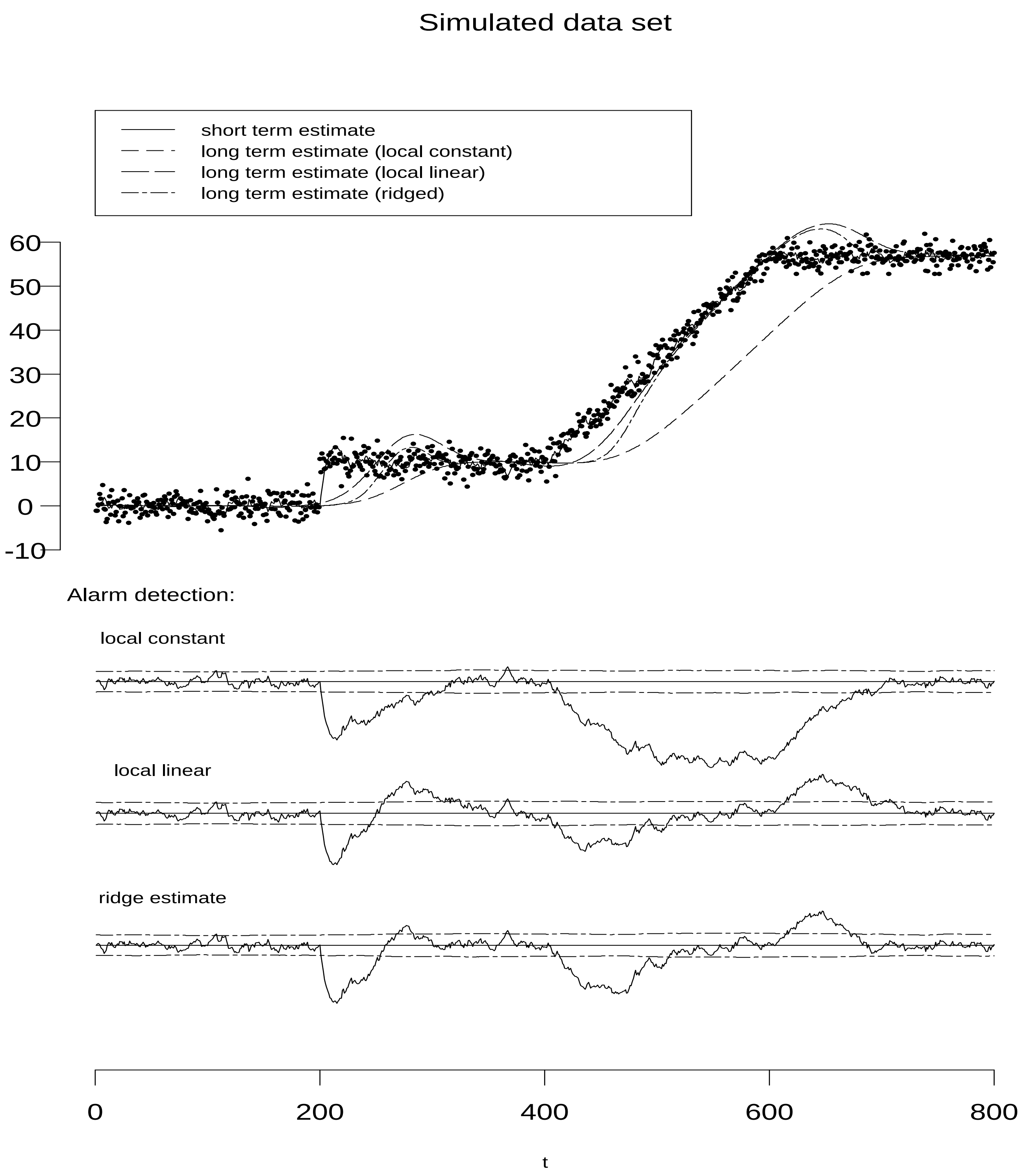



Figure 4.2: Simulated data set with local constant, local linear and ridged long term estimate using the alarm detection rule (4.3).

where $\lambda_t \in [0,1]$ is the ridge parameter. The ridge estimate again results as a weighted sum of the observations $y_i$ so that variance calculations for the ridge estimate are straight forward.

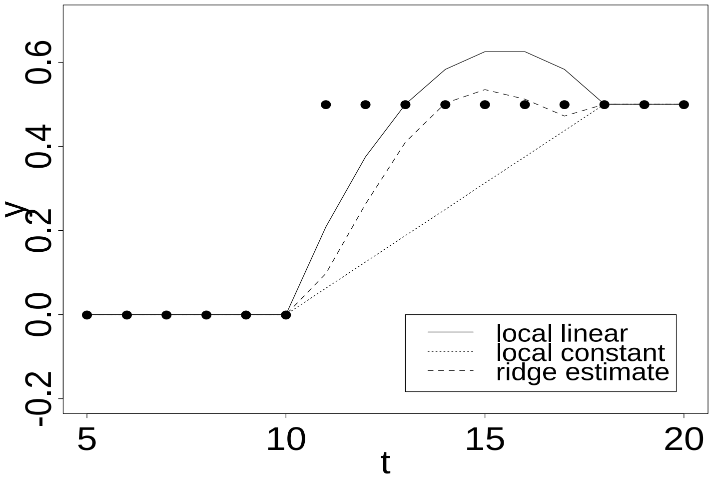



Figure 4.3: Tutorial on different behavior of local constant, local linear and ridge estimate after a jump.

The ridge parameter $\lambda_t$ in (4.15) is allowed to depend on time $t$. Seifert & Gasser (2000) suggest a rule of thumb to use design adaptive ridging. This is of little use for our scenario since the design is fixed and regular and observations are recorded at equidistant time intervals. For online monitoring it is more reasonable to use data adaptive ridging by considering the shape of the mean function $\mu(t)$. The general idea is to work with local constant smoothers if the mean is constant while local linear smoothing should be used if there is a drift. We incorporate this by estimating the slope of $\mu(t)$ via the local linear estimate

$$\hat{\beta}(t) = \sum_{i=0}^{h_1} v_i y_{t-i}$$

with $v_i = K_1(i)(S_{h_1,1} + iS_{h_1,0})/(S_{h_1,1}^2 - S_{h_1,0}S_{h_1,2})$. The principle is now that small squared slope estimates $\hat{\beta}(t)$ should lead to local constant fitting, i.e. large values of $\lambda_t$. We achieve this by setting

$$\lambda_t = e^{-c\hat{\beta}^2(t)} \qquad (c \geq 0). \tag{4.16}$$

Setting the parameter c equal to zero leads to local constant fitting while $c \to \infty$ gives local linear smoothing. In Figure 4.2 we use $c = 50$. It becomes obvious that the ridge estimate combines the advantages of local linear and local constant fitting. Figure 4.4 shows the value of $\lambda_t$ over time in this example. We pick up the task of how to select the constant $c$ at the end of example 4.4.2.

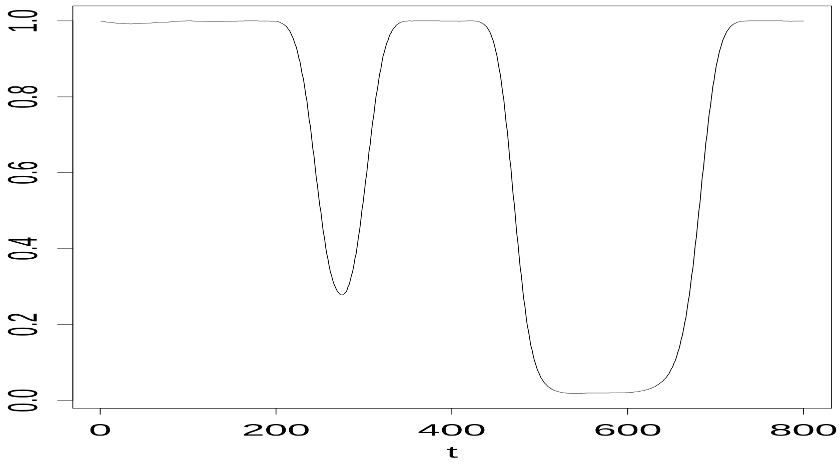



Figure 4.4: Development of the ridge parameter $\lambda_t$ over time for the simulated data set analyzed in Figure 4.2.

Criterion (4.16) is a suggestion but in no way unique. Other choices for choosing $\lambda_t$ can easily be constructed. We experimented with the Seifert & Gasser suggestion of design adaptive ridging. Not surprisingly this did not convince since it led to constant ridging, i.e. non adaptive and independent of $t$. To let the ridging parameter depend on the slope any monotonically decreasing function in $|\beta(t)|$ would work. The chosen form (4.16) however proved to work satisfactory in practice, in particular by applying the squared slope a clear distinction between the states $\lambda_t = 0$ and $\lambda_t = 1$ can be made.

### 4.3.2 Choice of $h$, $h_1$ and $h_2$.

In this section we will give some guidelines concerning the choice of the window sizes $h$, $h_1$ and $h_2$. The major importance of them has $h_1$, since this constant defines if breaks of short- or long term trends shall be detected. Thus what will be chosen firstly is $h_1$, and the other constants will be selected according to this choice.

For the selection of $h_1$ we provide the following rule of thumb: If the main focus is to detect breaks of trends with length down to a value $D$, one has to choose $h_1 \approx D$. We illustrate this point in an example: We simulated a time series of length 500 from the AR(2) process $Y_t = 0.55Y_{t-1} + 0.45Y_{t-2} + \epsilon_t$, where $\epsilon_t \sim N(0, 0.3^2)$. In Figure 4.5 (top) we perform alarm detection with $h_1 = 120$, $h_2 = 25$, $h = 200$. Apart from some alarm signals in the warming-up-period in the beginning, only breaks of long term trends at $t = 160, 237, 339$ and 386 (there a long term falling trend starting at about $t = 205$ is broken) are detected.

However, setting $h_1 = 60$, $h_2 = 15$ and $h = 100$ yields a very different picture: Now

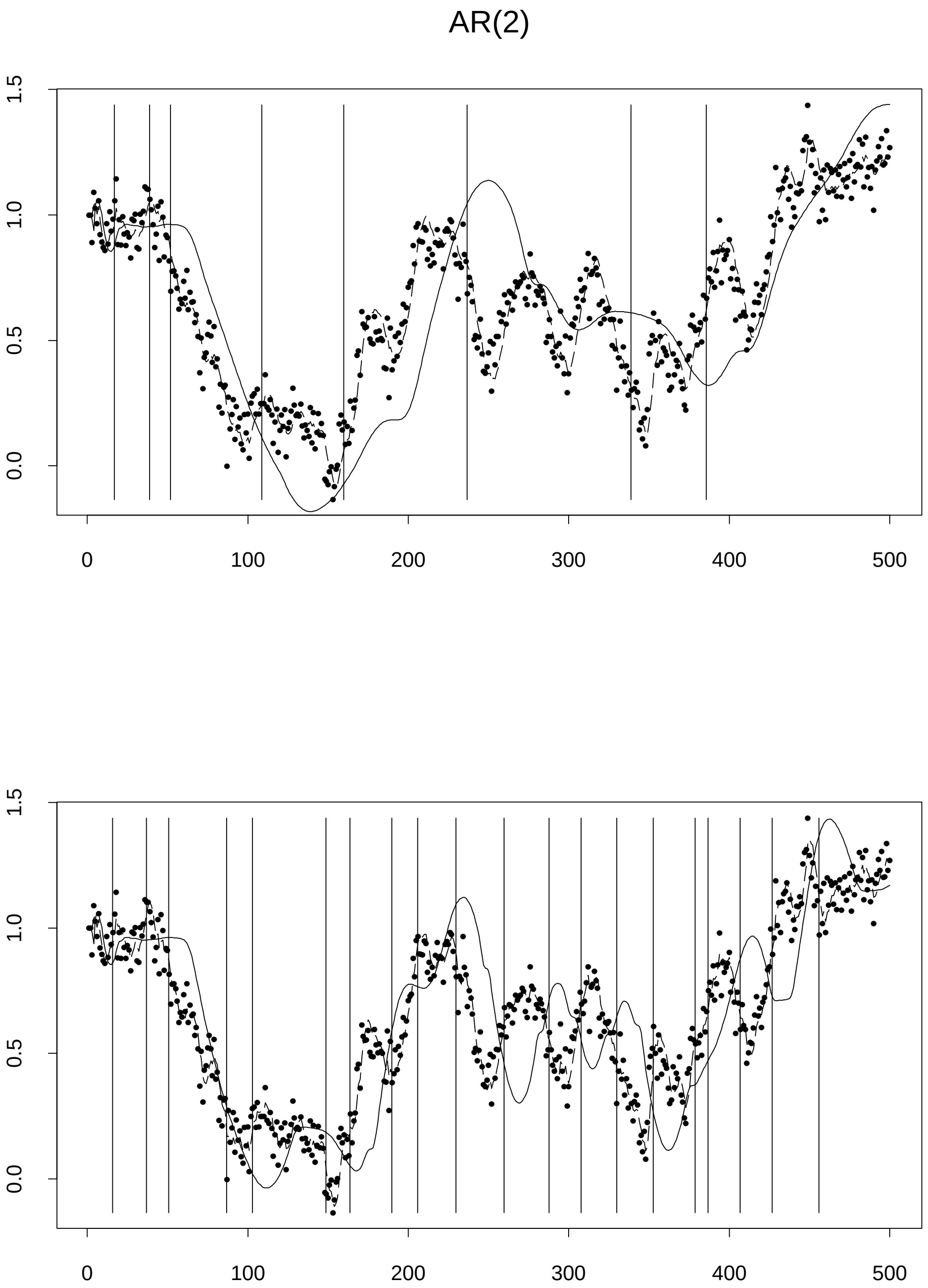



Figure 4.5: Alarm detection for simulated AR(2) process using $h_1 = 120$ (top) resp. $h_1 = 60$ (bottom). $\hat{\mu}_1$ and $\hat{\mu}_2$ are represented by the solid line resp. the dashed line. Vertical lines indicate the detection of breakpoints.

a large amount of breaks of short term trends are detected, like demonstrated in Figure 4.5 (bottom). Thus, the suitable value of $h_1$ depends less on the data than on the intentions of the data analyst.

What concerns the choice of $h_2$, we already mentioned that this value influences the speed of the detection, in the sense that small values of $h_2$ lead to a short reaction time, but to a high variance of the procedure. From our experience, we suggest to set $h_2 \approx h_1/5$, but not smaller than 15.

We already stated that the window size $h$, responsible for the amount of data used to estimate the autocorrelation function, is reflecting the local stationarity of the process, i.e. by imposing a certain value of $h$ we assume that the process is more or less stationary over a distance $h$. Therefore $h$ may not be too big, to retain sufficient flexibility of the variance estimation, but should be bigger than $h_1$ to get reliable results. We suggest to set $h \lesssim 2 \cdot h_1$.

### 4.3.3 Missing values and outliers

In practical applications one is often faced with outliers or missing data which disturb the performance of the alarm rule. We suggest the following adjustments. If observation $y_t$ is missing or outlying we impute a predicted value $\hat{y}_t$ calculated from the previous observations. A simple setting is to use $\hat{y}_t = \hat{\mu}_2(t-1)$. This setting works fine to overcome both the missing values in Example 4.4.1 as well as the artificial outlier in Example 4.4.2.

In the presence of sloping data the method can be more sophisticated to cover possible shifts. We therefore predict $y_t$ by using a linear extrapolation from the previous short term estimates via $\hat{y}_t = \sum_{i=1}^{h_2} \nu_i \hat{\mu}_2(t-i)$. The weights $\nu_i$ for this extrapolation can be calculated like $w_{i,2}$ in Section 4.2, but applying values $S_{h_2,j}(j = 0,1,2)$ constructed by sums starting at $i=1$ instead of $i=0$. These weights have to be calculated only once, so that extrapolation is numerically simple.

It remains the question of how to detect outliers. An outlier is classified as a single or small group of observations which do not follow the model. A detection rule for outliers is for instance

$$|y_t - \hat{y}_t| > k\sqrt{\hat{\gamma}_{t-1}(0)} \tag{4.17}$$

where $\hat{y}_t$ is a predictor for $y_t$ calculated as above and $k$ is some positive constant. In the data examples we collected good experiences with the setting $k = 10$, even though different values can be more suitable in other data scenarios. If $y_t$ is classified as outlier, its value is substituted by its predictor. Moreover, if (4.17) holds for a number of consecutive time-points alarm should be given.

### 4.3.4 Variance calculation

The moment based variance estimator described in the previous section can be inefficient if the data are uncorrelated or the errors trace from a model with parametric dependence pattern, e.g. an AR(1) process. In the first case one can set $\gamma(i) = 0$ for $i > 0$. In the latter case one could use the assumed dependence process to improve the variance estimation. For the AR(1) process for instance one fits locally the regression model

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \tag{4.18}$$

to the residuals $\varepsilon_t$, where $\nu_t$ are uncorrelated white noise errors. This yields the covariance function $\gamma(d) = \sigma^2\rho^d$ for $d = 0, \ldots, h$. In practice (4.18) is fitted to the fitted residual $\hat{\varepsilon}_t = y_t - \hat{\mu_2}(t)$ and one obtains

$$\hat{\rho} \;=\; \sum_{i=1}^{h} \hat{\varepsilon}_{t-i}\hat{\varepsilon}_{t-i+1} / \sum_{i=1}^{h} \hat{\varepsilon}_{t-i}^2.$$

The coefficient $\hat{\rho}$ can thereby again be updated recursively from previous values as shown in the appendix.

Variance calculation suffers from jumps and edges since both estimates, the short term and the long term estimate are biased at the jumps and residuals are overfitted. It is therefore advisable to pause online updating of the variance once a jump or outlier has been detected. This means in this case one sets $\hat{\gamma}_t = \hat{\gamma}_{t-1}$ until the alarm is stopped.

## 4.4 Examples

### 4.4.1 Cardio beats

In a hospital the cardio beats per minute of the mother before the confinement are monitored. It is of interest to detect sudden changes in the recorded data. Figure 4.6 shows the data and the resulting short and long term estimates. We use bandwidths $h_1 = 160, h_2 = 30, h = 300$ and a ridging constant $c = 80$.

A special property of this data set is the large amount of missing values, displayed as data points with $Y = 0$. However, the algorithm manages to outnumber these values and hence the estimated curves are not affected as seen in the first period of missing values from $t = 176$ to $t = 196$. The bottom graph in Figure 4.6 shows the standardized test statistic $T_t$ and bands given by the 99.5%- quantile of the standard normal distribution. It is seen that all jumps are detected quickly and significantly.

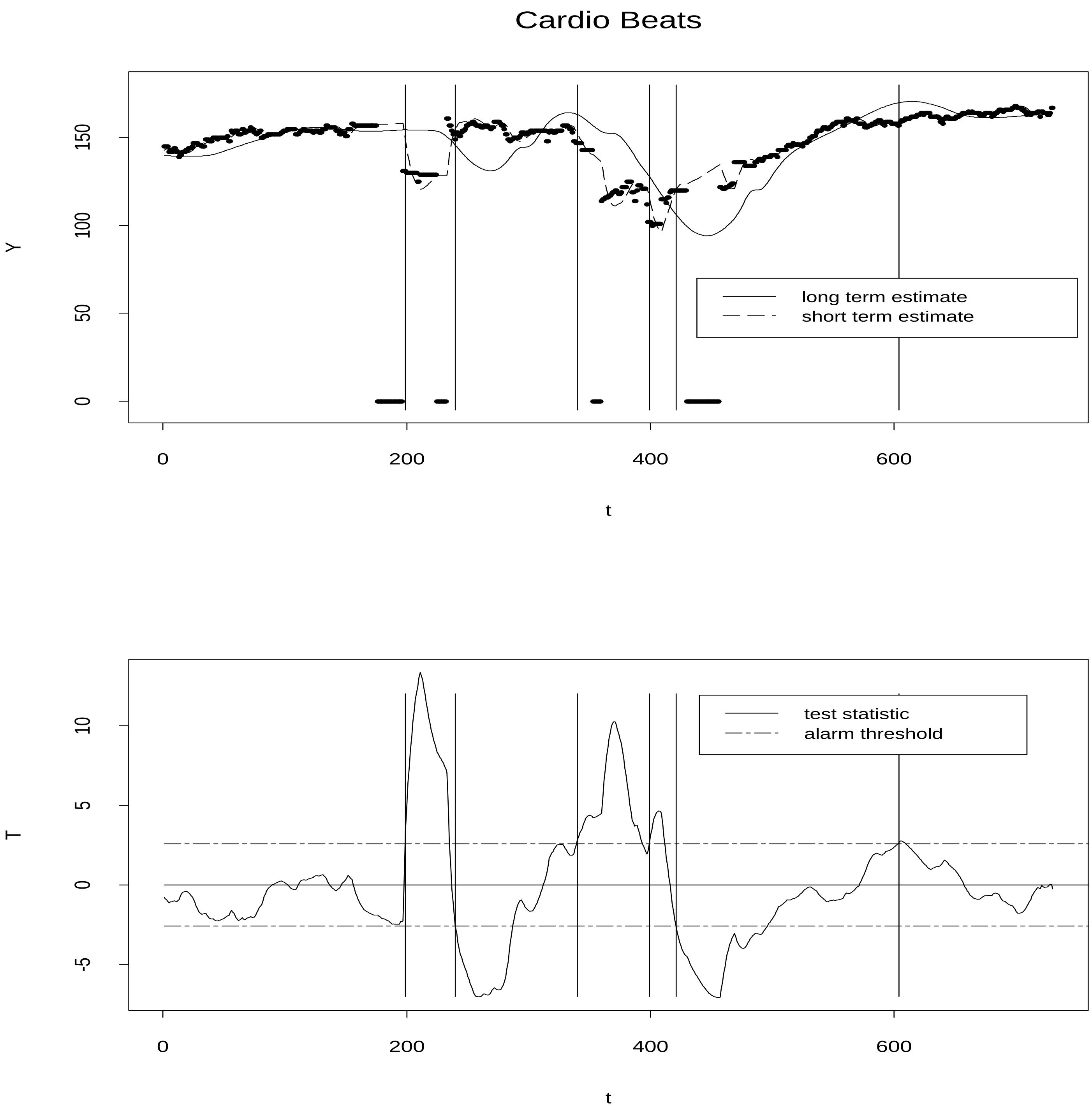


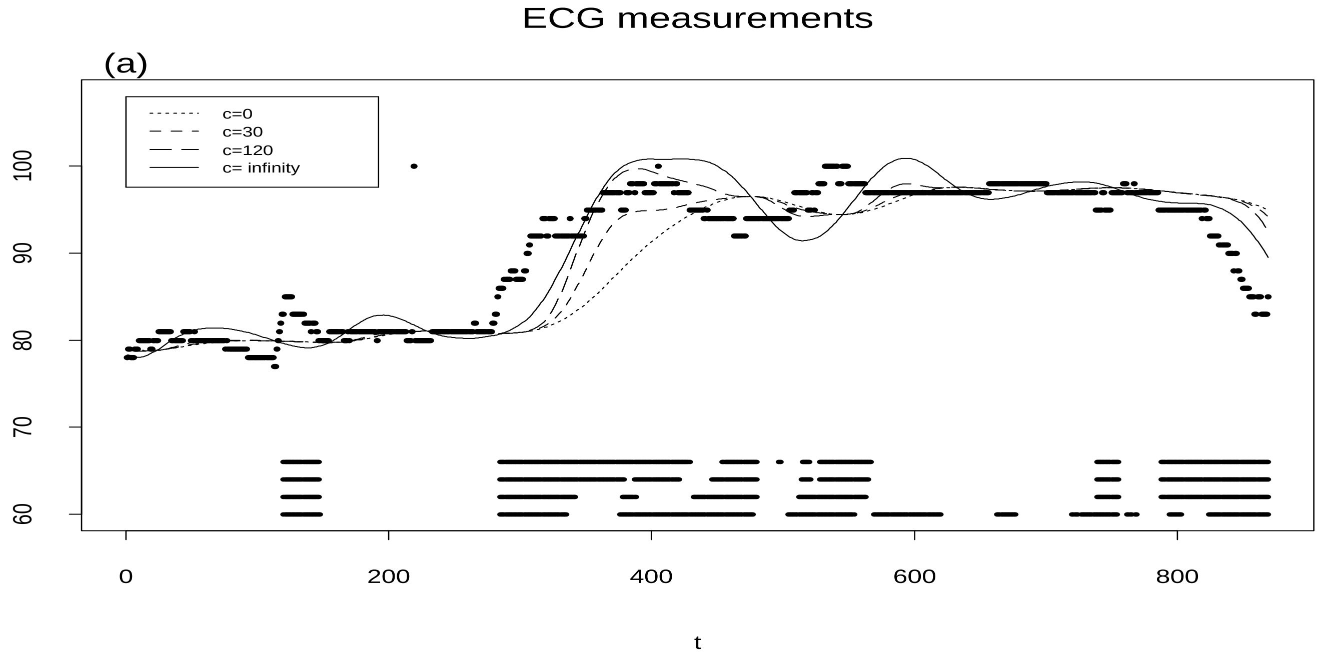

Figure 4.6: Cardio data with long and short term estimates. In the bottom the test statistic $T_t$ is compared with the quantile $u_{0.995} = 2.58$. Vertical lines indicate the detection of breakpoints.

At points 199 and 240 a shift in the level is found while at 340 the cardio beats decrease abruptly with level changes detected at 399 and 421. Afterwards the cardio beat frequency increases slowly until it reaches the plateau. The end of the increase is detected at 604.

### 4.4.2 ECG measurements

In the second example we apply the method to data which have been previously used in Daumer & Falk (1998) for the demonstration of their online monitoring algorithm. The data are ECG measurements taken every five seconds from a patient undergoing a skin transplantation. At $t = 219$ an artificial outlier is added. Figure 4.7(a) and 4.7(b) show the long term estimates and test statistics for different settings of the ridging parameter c in (4.16). For all settings the breakpoints at timepoint 120 and 285 are detected. Afterwards however the estimates behave differently. For $c = 0$ one obtains a local constant estimate. This is unable to adjust for the slope and does not find the end of the slope area at 378. Afterwards the local constant detects small level changes at 454, 497, 515 and 739. On the contrary the local linear fit, obtained for $c = \infty$, gives the end of the slope area but oversteers afterwards so that some small level changes are not uncovered, but some spurious alarm signals are given. In contrast, setting the ridging parameter $c = 120$ compensates the problem of oversteering and detects both, the end of the slope area as well as the small level changes afterwards. We return to this data example in the next section and compare our method with the procedure suggested in Daumer & Falk (1998).

Speaking more generally, we conclude that if the data describes more or less a step function and the intention is mainly to detect jumps, we recommend $c = 0$, which corresponds to a local constant long term estimator. If however mainly breaks of trends shall be detected, one should set $c = \infty$ and thus use a local linear long term estimator. Varying $c$ between 0 and $\infty$ means balancing between these two goals, and the appropriate value of $c$ depends on the kind of breakpoints which shall be detected. A general guideline or rule of thumb on how to choose $c$ is therefore difficult. Nevertheless, due to (4.16) it is observed that if the process $Y_t$ is multiplied by a constant $\delta$, say, coefficient $c$ should be updated to $c/\delta^2$ to have the same amount of ridging. Hence, the choice of $c$ depends on the overall variability of the process, which includes areas of shifts as well. We experimented a little and found that $c = 10^4/\text{var}(Y)$ is a reasonable starting point for fine tuning of $c$. Apparently the variance of $Y$ is normally unknown, since we record data online. Usually, however, one should have a notion about this value which allows to set $c$ at the starting value for further tuning. Again, the right choice of $c$ depends in



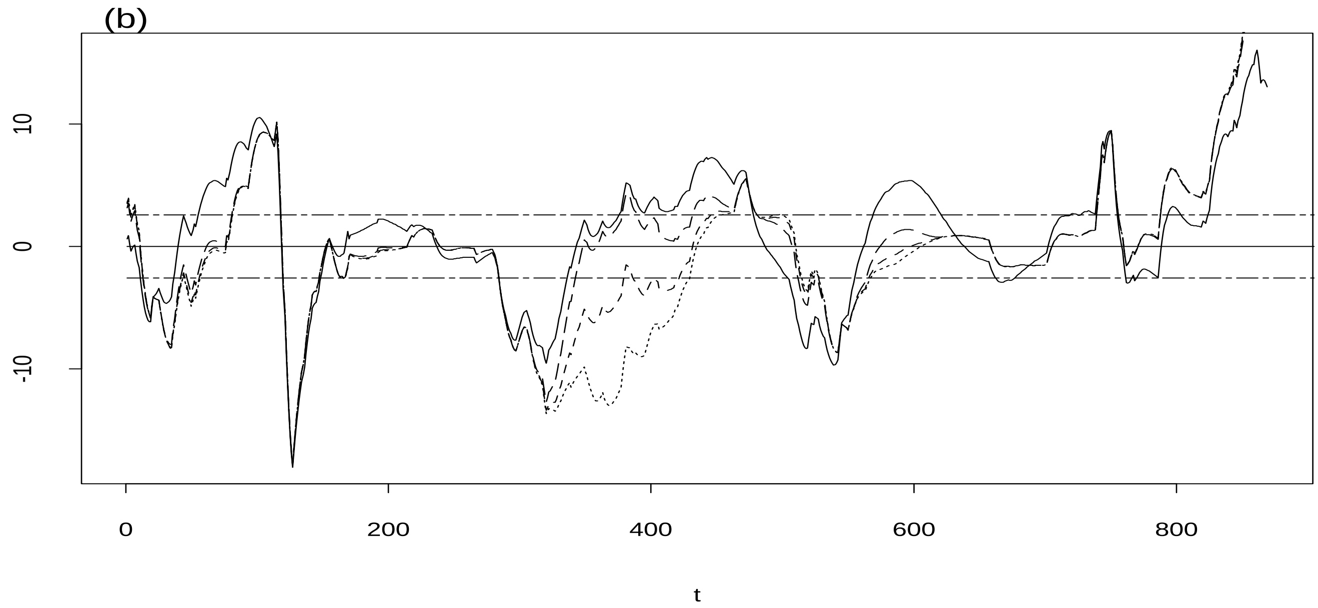




Figure 4.7: (a) ECG data with long term estimates for different degrees of ridging, using $h_1 = 150, h_2 = 25, h = 200$. The lines in the bottom indicate the alarm periods for $c = 0$ (top), $c = 30, c = 120$, $c = \infty$ (bottom). Alarm signals for $t < 100$ are ignored, since the algorithm needs sufficient data points to work. (b) Test statistic $T_t$ for $c = 0, \dots, c = \infty$, degrees of ridging symbolized like in (a). Alarm thresholds (horizontal lines) at $\pm 2.58$.

particular on the substance matter of the online monitoring, i.e. whether one is interested in detecting jumps or breaks in trends.

## 4.5 Comparison to other methods

### 4.5.1 Autoregressive models

In Gather, Bauer, Imhoff & Löhlein (1998) it is assumed that the data follow an autoregressive model. We illustrate their method at the cardio data from above. Before applying the method, we substitute the missing values by short-term-predictors. Then the data set is divided in an estimation period and a prediction period. Since the data have to be more or less stationary during the estimation period, we choose the estimation period $t = 1, \ldots, 180$. In order to obtain a nearly balanced proportion between the amount of data in the two periods we reduce the data set to the first 380 data points.

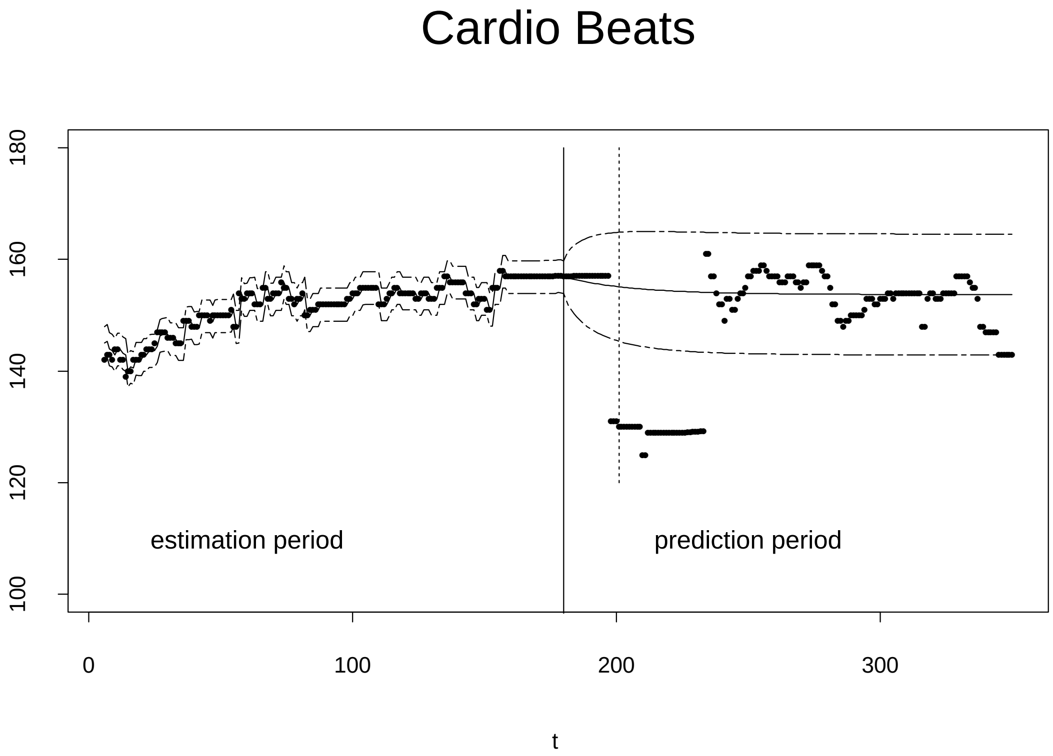



Figure 4.8: Cardio data with predicted values (solid line), 95% confidence bands (dashed lines) and alarm detection at t=201 (dotted line).

During the estimation period the parameters for the AR-process are estimated (the AIC criterion suggested an AR(1) model). In the prediction period it is observed whether the data points are inside or outside a confidence band surrounding the predicted values. The result is shown in Figure 4.8 for a 95% confidence interval. If less than five consecutive observations are outside of the confidence interval, then they are classified as outliers, whereas a level change is detected for more than five points out of the confidence interval. Thus, the jump detected at $t = 197$ is classified as a level change at $t = 201$, compared to a detection at $t = 199$ with our adaptive ridging method.

We conclude that the AR method by Gather et al. is a fast and reliable method, but is probably not suitable for every kind of data, especially data with quickly changing rising or falling trends, whereas the local estimation method we proposed in this paper adapts to a wide range of data situations.

### 4.5.2 Phase space models

Another idea of Gather, Bauer, Imhoff & Löhlein (1998) is to plot every data point against its previous data point in a phase space. We tried this method for the first 380 cardio data points (see Figure 4.9). Gather et. al. move a time window of length 60 through the data and alarm is given if the next five consecutive observations are in a different region than the previous 60 ones. In the plot, the way of the data in the phase space can be followed. Starting somewhere in the left down area of the big cluster, the line climbs up to the right top edge, then falls down and turns left to the small cluster (representing the data points $t = 197, \ldots, t = 232$) and finally climbs up again to the big cluster. Every change of the cluster represents a jump in the data. This means that alarm is given at the timepoints $t = 201$ and 236, compared to alarm signals at $t = 199$ and 240 with the adaptive ridging method.

Though this method is very useful for visualizing the structure of the data, we think it might be difficult to use it online for reliable alarm detection, especially for sloping data, where the dividing lines between single clusters become foggy.

### 4.5.3 State space models

Daumer & Falk (1998) and Fahrmeir & Künstler (1999) use state space models for filtering time series. A linear state space model is given by a linear observation equation

$$y_t = z_t'\beta_t + \varepsilon_t \qquad (t = 1, 2, \ldots)$$

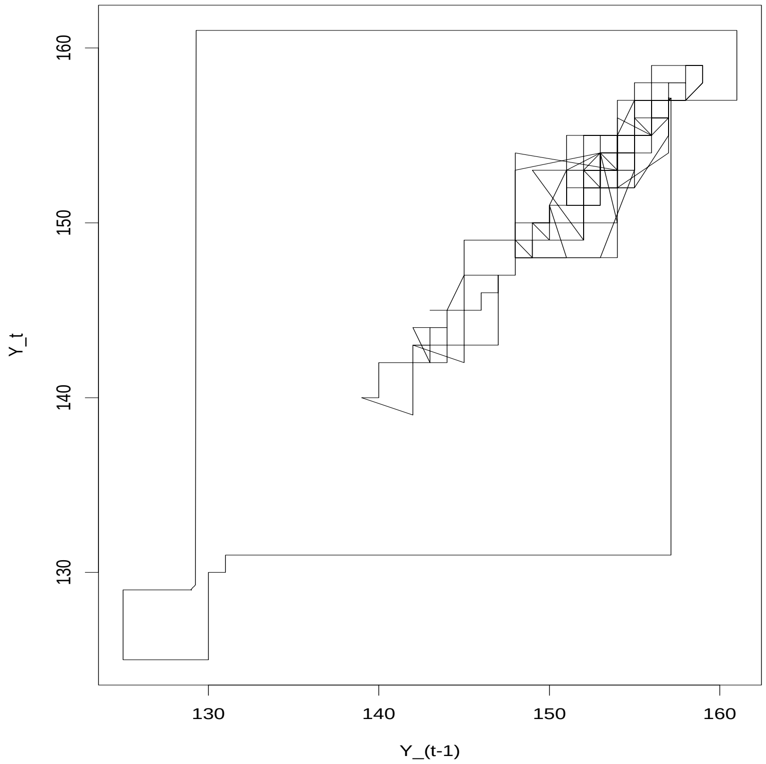



Figure 4.9: Cardio data from example 4.1. (for $t = 1, \ldots, 380$) plotted in a phase space.

for the observations $y_1, y_2, \ldots$ given the states $\beta_1, \beta_2, \ldots$, which is supplemented by a linear transition equation

$$\begin{aligned} \beta_t &= F_t \beta_{t-1} + v_t \qquad (t = 1, 2, \ldots) \\ \beta_0 &= a_0 + v_0 \end{aligned}$$

with Gaussian errors $\varepsilon_t$ and $v_t$, nonrandom vectors $z_1, z_2, \ldots$ and transition matrices $F_1, F_2, \ldots$ . This model can be solved with Kalman filters. Daumer & Falk (1998) define such a state space model for each possible location of a jump. The resulting family of models, called a multi-process model, is examined with Bayesian methods and jumps are detected by choosing the most likely model. For detecting outliers, a second multi-process model has to be introduced. Daumer & Falk (1998) apply their method to the data shown in Figure 6(a) and find changepoints at $t = 120$, 285, 506, 752 and 821. In contrast, the local ridging method with $c = 120$ uncovers the changepoints 120, 285, 378, 432, 512, 739 and 788. It appears that both methods uncover abrupt changes from a long term level but the local adaptive ridging method appears more flexible and gives alarm also at short term changes. Moreover both methods are equal in the speed of detection.

## 4.6 Conclusion

We showed that local smoothing methods can be used effectively for detecting jumps and bends in online monitoring. The algorithm combines the advantages of many other breakpoint detection methods: It can be used online, since only the data given until the examined time point are necessary for the estimations. Furthermore it is able to detect jumps or bends of flat and sloping trends. The method adapts to the variability of the data, which means that it will not give alarm for a small jump within highly fluctuating data, but will give alarm for the same jump for less variable data. Finally it is worth mentioning that only few computational effort is required, because the weights needed for the estimates have only to be calculated once and the variance calculations follow a simple update rule.

The only technical problem, namely the over-steering, can be solved quite satisfactory by adaptive ridging. However, we shouldn't suppress that it can't be avoided completely (see Figure 4.2 and 4.6). If one wants to exclude it totally, one has either to use only the data *after* a jump for the estimations of $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$ , which requires recalculating all weights after every jump, or to use methods like edge-preserving smoothing (see Chu, Glad, Godtliebsen & Marron, 1998). However, both ways require additional computational effort, so that it is questionable whether they convince in practice.

## 4.7 Appendix

**Technical Details**
*Derivation of (4.10)*
Note that

$$
\begin{aligned}
&\boldsymbol{d}_{t+1,h}^T \boldsymbol{D}_{t+1,h} \\
&= \left(y_{t+1} - \hat{\mu}_2(t+1), \boldsymbol{d}_{t,h-1}^T\right) \begin{pmatrix} \frac{y_{t+1}-\hat{\mu}_2(t+1)}{h+1} & \mathbf{0}_1 & \cdots & \mathbf{0}_{h_1} \\ \frac{\boldsymbol{d}_{t,h-1}}{h+1} & \frac{y_{t+1}-\hat{\mu}_2(t+1)}{h} & \cdots & \frac{y_{t+1}-\hat{\mu}_2(t+1)}{h-h_1+1} \\ & \frac{\boldsymbol{d}_{t,h-2}}{h} & \cdots & \frac{\boldsymbol{d}_{t,h-h_1-1}}{h-h_1+1} \end{pmatrix} \\
&= \left( \frac{\{y_{t+1} - \hat{\mu}_2(t+1)\}^2}{h+1} + \frac{\boldsymbol{d}_{t,h-1}^T \boldsymbol{d}_{t,h-1}}{h+1}, \ldots \right. \\
&\qquad \left. , \frac{\{y_{t+1} - \hat{\mu}_2(t+1)\}\{y_{t-h_1+1} - \hat{\mu}_2(t-h_1+1)\}}{h-h_1+1} + \frac{\boldsymbol{d}_{t-h_1,h-h_1-1}^T \boldsymbol{d}_{t,h-h_1-1}}{h-h_1+1} \right).
\end{aligned}
$$

Making use of $\boldsymbol{d}_{t-d,h-d-1}^T \boldsymbol{d}_{t,h-d-1}/(h-d+1) \sim \boldsymbol{d}_{t-d,h-d}^T \boldsymbol{d}_{t,h-d}/(h-d+2)$ provides (4.10) for $h$ sufficiently large.

*Update of $\hat{\rho}_t$ in model (4.18)*
Note that

$$\begin{aligned}\left(\sum_{i=0}^{h-1} \hat{\varepsilon}_{t-i}^2\right)^{-1} &= \left(\hat{\varepsilon}_t^2 - \hat{\varepsilon}_{t-h}^2 + \sum_{i=1}^{h} \hat{\varepsilon}_{t-i}^2\right)^{-1} \\ &\approx \left(\sum_{i=1}^{h} \hat{\varepsilon}_{t-i}^2\right)^{-1} - \left(\sum_{i=1}^{h} \hat{\varepsilon}_{t-i}^2\right)^{-2} (\hat{\varepsilon}_t^2 - \hat{\varepsilon}_{t-h}^2)\end{aligned}$$

so that the inverse can approximated by recursive updating. Setting $R_{2,t}^{-1} = (\sum_{i=1}^{h} \hat{\varepsilon}_{t-i}^2)^{-1}$ one gets $R_{2,t+1}^{-1} \approx R_{2,t}^{-1} - R_{2,t}^{-2}(\hat{\varepsilon}_t^2 - \hat{\varepsilon}_{t-h}^2)$. The numerator $R_{1,t} = \sum_{i=1}^{h} \hat{\varepsilon}_{t-i}\hat{\varepsilon}_{t-i+1}$ can be updated by $R_{1,t+1} = R_{1,t} + \hat{\varepsilon}_t\hat{\varepsilon}_{t+1} - \hat{\varepsilon}_{t-h}\hat{\varepsilon}_{t-h+1}$.

## Final remarks

This article is joint work with Göran Kauermann, University of Bielefeld (Germany). The content of this chapter is to appear in *Journal of Statistical Computation and Simulation* (Einbeck & Kauermann, 2003). The article is reprinted with the kind permission of Taylor & Francis ltd., which holds the copyright. The authors are grateful to the Institute of Medical Statistics and Epidemiology at the Technical University Munich (TUM) for providing the data and especially to Dr. Daumer for explaining aims and problems of applied online monitoring. The data were recorded within the the project B2 of the SFB 386 in collaboration with Trium Analysis Online GmbH, Women hospital and policlinic of TUM (Prof. Dr. K. T. M. Schneider, PD Dr. J. Gnirs), Institute of anaesthesiology, TUM (Prof. Dr. E. Kochs, O. Möllenberg).

The software for the algorithms presented in this section is available at `www.stat.uni-muenchen.de/~einbeck/software.html` (implemented in S-Plus).

# Chapter 5

# Local Principal Curves

## 5.1 Introduction

The classical problem of how to find the best curve passing through some data points $(x_i, y_i), i = 1, \ldots, n$ can be handled in two fundamentally different ways. Let us regard the data points as realizations of i.i.d. random variables $(X_i, Y_i)$ drawn from a population $(X, Y)$. A common approach is to regard $X$ as an explanatory variable for the dependent variable $Y$. This concept is used in all methods where the focus is on regression or smoothing and is especially useful when the objective is prediction of the dependent variable from the observations $x_i$. Thereby $X$ and $Y$ have an asymmetric relationship and cannot be interchanged without changes of the results.

In contrast, $X$ and $Y$ may be regarded as symmetric, thus we do not assume that one variable can be made reponsible for the value of the other one. Rather they are generated simultaneously from a common underlying distribution. These approaches are useful when the focus is on dimension reduction or simply description of the data. Representants here are methods like the ACE algorithm, canonical correlation or principal component analysis. Linear Principal components, introduced by Pearson (1901), are a common tool in multivariate analysis, applied for example in feature extraction or dimension reduction. Jolliffe (2002) gave an overview on properties and applications of principal components. Nonlinear principal components have been developed by Schölkopf & Smola (1998) and successfully employed for pattern recognition.

A natural extension of principal components are principal curves, which are descriptively defined as one-dimensional smooth curves that pass through the "middle" of a $p-$dimensional data set. Though this concept is intuitively clear, there is much flexibility in how to define the "middle" of a distribution or data cloud. Hastie & Stuetzle (1989) (hereafter HS), who did the groundbreaking work on principal curves, use the concept of self-consistency (Tharpey & Flury, 1996), meaning that each point of the principal curve is the average of all points that project there. A variety of other definitions of principal curves have been given subsequently by Tibshirani (1992), Kégl, Krzyzak, Linder & Zeger (2000) (hereafter KKLZ), and more recently Delicado (2001), which differ essentially in how the "middle" of the distribution is found.

Apart from Delicado (2001), all concepts mentioned above work more or less as follows: They start with a straight line, which is mostly the first principal component of the data set, and try to dwell out this line or concatenate other lines to the initial line until the resulting curve passes satisfactorily through the middle of the data. This methodology leads to some technical problems. HS generally exclude intersecting curves from the definition of principal curves and are not able to handle closed curves. Banfield & Raftery (1992) (hereafter BR) provide a bias corrected version of the HS algorithm which solves the latter problem, but yields more wiggly results than HS. Chang & Ghosh (1998) combine the algorithms of HS and BR and show that this yields a smooth and unbiased principal curve, at least for simple data situations. Tibshirani's theoretically attractive approach seems to have the same problems as HS, though not explicitely stated, and further seems to have a lack of flexibility for strongly skewed data. These difficulties have been solved by Verbeek, Vlassis & Kröse (2001), but at the expense of an apparently wiggly principal curve, since polygonal lines are connected in a somehow unsmooth manner. KKLZ work also with polygonal lines and obtain with high computational effort a smooth and flexible principal curve, which only fails for very complicated data structures. None of these algorithms seems to be able to handle curves which consist of some multiple or disconnected parts. Recently, Kégl & Krzyzak (2002) provided a promising algorithm to obtain *principal graphs*, i.e. multiple connected piecewise linear curves, in the context of skeletonization of hand-written characters.

All these methods have to be regarded as global, since in every step of their algorithms, or at least in the initial step, all available data points are used. As alternative to the global methods, which lead to exploding computational costs for large data sets or high-dimensional data, it would be desirable to have a local method at hand, which only considers data which are close to the target point. Lately, Delicado (2001) proposed the first principal curve approach which can be called local. Assume a d-dimensional random vector $X$ and $n$ random samples $X_i \in \mathbb{R}^d, i = 1, \ldots, n$ from $X$, where $X_i = (X_{i1}, \ldots, X_{id})$. For each point $x$, Delicado considers the hyperplane $H(x, b)$ which contains $x$ and is orthogonal to a vector $b$. The set of vectors

$b^*(x)$ minimizing the total variance $\phi(x, b) = TV(X|X \in H(x, b))$ defines a function $\mu^*(x) = E(X|X \in H(x, b^*(x)))$. Principal oriented points (POPs) are introduced as fix points of the function $\mu^*(\cdot)$. For a suitable interval $I \in \mathbb{R}$, $\alpha$ is called a principal curve of oriented points (PCOP) if $\{\alpha(s)|s \in I\}$ is a subset of the fix point set of $\mu^*$. Delicado shows that POPs exist, and that in case $b^*(x)$ is unique (this implies that the principal curve is a function), to each POP exists a PCOP passing through it. Since the hyperplanes $H$ are sets of measure zero, it is necessary to employ a kind of smoothing for calculating the conditional expectation on the hyperplane. This is achieved by projecting all data points on $H(x, b)$, obtaining points $X_i^H$, and assigning weights

$$w_i = w(|(X_i - x)^T b|), \tag{5.1}$$

where $w$ is a decreasing positive function, e.g. $w(d) = K(d/h)$, with a kernel function $K$. Let $\tilde{\mu}(x, b)$ denote the weighted expectation of the $X_i^H$ with weights $w_i$. Now $\mu^*(x)$ is approximated by $\tilde{\mu}^*(x) = \tilde{\mu}(x, \tilde{b}^*(x))$, where $\tilde{b}^*(x)$ (and hence $H$) is constructed such that the variance of the projected sample, weighted with $w_i$, is minimized. Localization enters here twofold. Firstly, by applying (5.1), points near to the hyperplane are upweighted. Secondly, a cluster analysis is performed on the hyperplane, and only data in the local cluster are considered for averaging. How is the principal curve found in practice? The algorithm searches the fix point set of $\tilde{\mu}^*(x)$ as follows. Repeatedly, choose a point randomly from the sample $X_1, \ldots, X_n$ and call it $x_{(0)}$. Then iterate $x_{(\ell)} = \tilde{\mu}^*(x_{(\ell-1)})$ until convergence. In this manner a finite set of POPs is obtained. However, no fix point theorem guarantees convergence of this algorithm, although Delicado reports quick convergence for some real data sets. In order to obtain a PCOP from a set of POPs, Delicado proposes an idea which we will further exploit. Assume an POP $x_1$ calculated as explained. From the set of principal directions $\tilde{b}^*(x_1)$, choose one vector $b_1$. Now walk a step of length $\partial$ from $x_1$ in direction of $b_1$, i.e.

$$x_2^0 = x_1 + \partial b_1, \tag{5.2}$$

where $\partial$ is previously fixed. The point $x_2^0$ serves as a new starting point for a new iterating process, leading to a new point $x_2$ of the principal curve. This is repeated $k$ times until no points $X_i$ can be considered to be near the hyperplane $H(x_k^0, b_k)$. Then return to $(x_1, b_1)$ and complete the principal curve in direction of $-b_1$. Afterwards move on to another of the previously chosen POPs and continue analogously.

Delicado's concept is mathematically elegant and theoretically well elaborated. It works fine even for some complicated data structures as spirals etc. (Delicado & Huerta, 2003), but fails for branched data (Evers, 2003). One might consider it as a drawback that the concept is mathematically and computationally demanding and not intuitively clear. Parameter selection is accordingly cumbersome (Delicado & Huerta, 2003).

In this chapter, we introduce a concept similar to that of Delicado. However, we replace the fix points of $\tilde{\mu}^*$ by local centers of mass, and replace the principal direction $b_1$ by a local principal component. We call the resulting curve, which consists of a series of local centers of mass, *local principal curve*. We introduce the notion of *coverage*, which evaluates the performance of the principal curve approximation and is a helpful tool to compare the performance of different principal curve algorithms. We show that, using this concept of coverage, the parameters which are necessary for our algorithm can easily be selected in a data-adaptive way. The price paid for the easiness of the concept is that in contrast to Delicado's approach there is no statistical model and consequently it is hard to derive theoretical results. However, in Section 5.5 we give a theoretical justification for our method. The algorithm will be presented in the following section.

## 5.2 The algorithm

Assume a d-dimensional data cloud $X_i \in \mathbb{R}^d, i = 1, \ldots, n$, where $X_i = (X_{i1}, \ldots, X_{id})$. We try to find a smooth curve which passes through the "middle" of the data cloud. The curve will be calculated by means of a series of local centers of mass of the data, according to the following strategy:

1. Choose a suitable starting point $x_{(0)}$. Set $x = x_{(0)}$.
2. Calculate the local center of mass $\mu^x$ around $x$.
3. Perform a principal component analysis locally at $x$.
4. Find the new value $x$ by following the first local principal component $\gamma^x$ starting at $\mu^x$.
5. Repeat steps 2 to 4 until $\mu^x$ remains (approximately) constant.

The series of the $\mu^x$ make up the desired curve. In the sequel we will explain these steps in detail:

*1. Selection of the starting point*

In principle, every point $x_{(0)} \in \mathbb{R}^d$ which is in or close to the data cloud can be chosen as starting point. There are two ideas which suggest themselves:

- Based on a density estimate the point with the highest density $x_{(0)} = max_{x\in\mathbb{R}}\hat{f}(x)$ is chosen.
- A point $x_{(0)} = X_i$ is chosen at random from the set of observations.

The advantage of the density method is that one can be quite sure not to start in a blind alley, whereas a randomly chosen point could be an outlier far from the data cloud which stops the algorithm already in the first loop. However, this is not very likely, and the computational costs of the second approach are much lower. Moreover, for handling crossings a randomly chosen starting point is even superior to a high density point.

*2. Calculating the local center of mass*

Let $H$ be a bandwidth matrix and $K_H(\cdot)$ a $d-$ dimensional kernel function. Given that all components of $X$ are measured on the same scale, we set $H = \{h^2 \cdot I : h > 0\}$, with $I$ standing for the d-dimensional identity matrix. For a detailed description of multivariate kernels and bandwidth matrices see Wand & Jones (1995). For selection of $h$, see Section 5.7. The local center of mass around $x$ is given by

$$\mu(x) = \frac{\sum_{i=1}^{n} K_H(X_i - x)X_i}{\sum_{i=1}^{n} K_H(X_i - x)} \tag{5.3}$$

This estimator and its relation to the Nadaraya-Watson estimator have been analyzed in Comaniciu & Meer (2002). For ease of notation, we will abbreviate $\mu^x = \mu(x)$ in the following. We denote by $\mu_j^x$ the $j$-th component of $\mu(x)$.

*3. Calculating the local principal component*

Let $\Sigma^x = (\sigma_{jk}^x)$ denote the local covariance matrix of $x$, whoose $(j,k)$-th entry $(1 \leq j,k \leq d)$ is given by

$$\sigma_{jk}^x = \sum_{i=1}^{n} k_i(X_{ij} - \mu_j^x)(X_{ik} - \mu_k^x) \tag{5.4}$$

with weights $k_i = K_H(X_i - x)/\sum_{i=1}^{n} K_H(X_i - x)$, and $H$ as in 2. Let $\gamma^x$ be the first eigenvector of $\Sigma^x$. Then $\gamma^x$ is the first column of the loadings matrix $\Gamma^x$ from the principal components decomposition $(\Gamma^x)^T\Sigma^x\Gamma^x = \Lambda^x$, where $\Lambda^x = diag(\lambda_1^x, \ldots, \lambda_d^x)$ is a diagonal matrix containing the ordered eigenvalues of $\Sigma^x$, with $\lambda_1^x \geq \ldots \geq \lambda_d^x$.

Note that the denotation "local principal components" is not new, but has been previously used for linear principal components localized in clusters (Skarbek, 1996; Kambhatla & Leen, 1997) or based on contiguity relations (Aluja-Banet & Nonell-Torrent, 1991) rather than by kernel functions.

*4. Obtaining an updated value*

The local principal component line $v^x$ can now be parameterized as

$$v^x(t) = \mu^x + t\gamma^x \quad (t \in \mathbb{R}), \tag{5.5}$$

and we obtain an updated value of $x$ by setting

$$x := \mu^x + t_0\gamma^x, \tag{5.6}$$

in analogy to step (5.2) of Delicado's algorithm. A suitable value of $t_0$ thereby has to be chosen beforehand. We defer the task of how to select $t_0$ to Section 5.7.

*5. Stop when $\mu^x$ remains constant*

When the end of the data cloud is reached, the algorithm will naturally get stuck and produce approximately constant values of $\mu^x$. One might stop before this state occures, e.g. when the difference between previous and current center of mass falls below a certain threshold.

The mechanism is demonstrated in Figure 5.1. The starting point $x_{(0)}$ is denoted by 0. The radius of the circle is equal to the bandwidth $h = 0.2$. Calculating the local center of mass around 0 yields the nearby point $m$. Moving along the first principal component with $t_0 = 0.2$ leads to the new point $x$ denoted by "1", and so on. The series of m's is the local principal curve. Note that the algorithm is based on finding an equilibration between opposing tendencies: On the one hand, the local principal components are oversteering, i.e. tending "outside" to the concave side of the curvature of the data cloud. On the other hand, the calculation of the local center of mass is smoothing the data towards the interior and thus in the opposite direction. These two effects together ensure that the estimated principal curve is not systematically biased.

## 5.3 Technical details

In practice, some modifications of the above algorithm are useful, which we describe in the following.

### 5.3.1 Maintainig the direction

A principal component line always has two directions, thus the corresponding eigenvector $\gamma^x$ could be replaced by its negative value $-\gamma^x$. Depending on the orientation of the eigenvector, the constructed curve moves in opposite directions. If this direction changes from one step to another, the algorithm dangles between these two

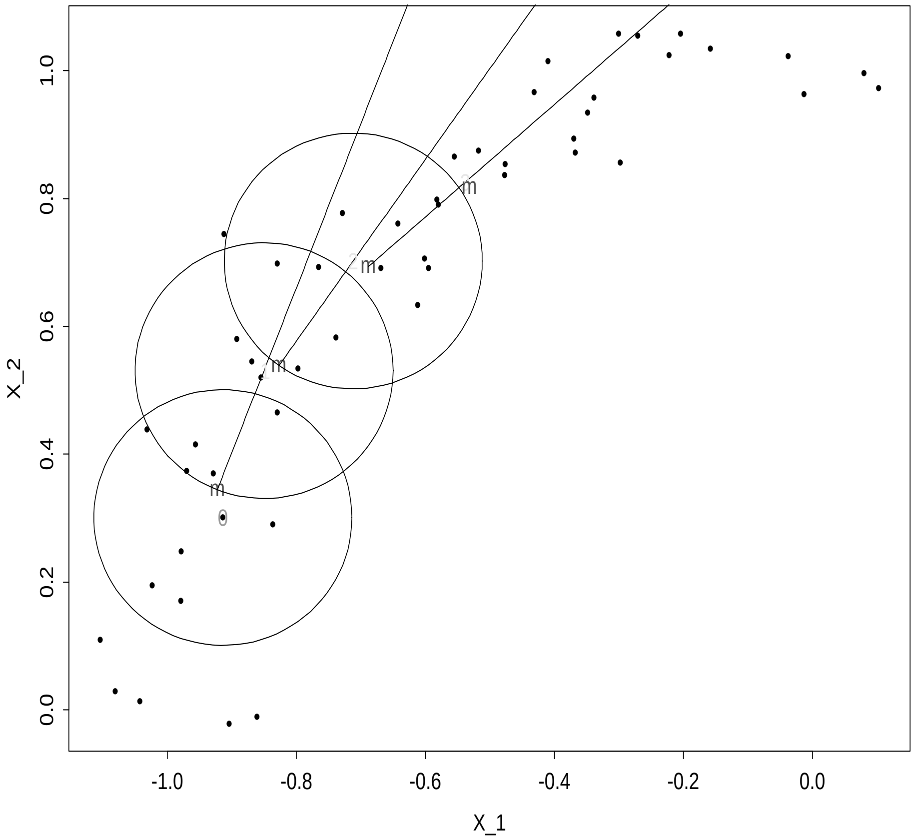



Figure 5.1: Demonstration of the local principal curve algorithm.

points and will never escape. Therefore one should check in every step that the local eigenvector has the same direction as in the previous step. This can be done by calculating the angle $\alpha^x_{(i)}$ between the eigenvectors $\gamma^x_{(i-1)}$ and $\gamma^x_{(i)}$ belonging to the $(i-1)$-th resp. $i$-th step , which is given by

$$\cos(\alpha^x_{(i)}) = \gamma^x_{(i-1)} \circ \gamma^x_{(i)},$$

where $\circ$ denotes the scalar product. If $\cos(\alpha^x_{(i)}) < 0$, set $\gamma^x_{(i)} := -\gamma^x_{(i)}$, and continue the algorithm as usual. This "signum flipping" has been applied in the step from "2" to "3" in Figure 5.1.

### 5.3.2 Running backwards from $x_{(0)}$

When one starts at a point $x_{(0)}$ and moves by means of local principal components to one "end“ of the cloud, one has omitted to consider the part between the starting point and the other end of the cloud, except if the data describe a closed curve, e.g. a circle or an ellipse. Therefore it is advisable to run from the starting point in both directions of the first principal component, what in practice means adding a 6th step to the algorithm:

6. For the starting direction $-\gamma^x_{(0)} := -\gamma^{x_{(0)}}$, perform steps 4 and 5.

### 5.3.3 Angle penalization

If the data cloud locally forms crossings, at each crossing the local principal curve has three possibilities where to move on. Often one desires that the curve goes straight on at each crossing, and does not turn arbitrarily to the left or right. In order to achieve this effect, we recommend to perform an angle penalization in addition to the signum flipping in each step of the algorithm. This might be done as follows:

Let $k$ be a positive number. For the angle $\alpha^x_{(i)}$, set

$$a^x_{(i)} := |\cos(\alpha^x_{(i)})|^k$$

and correct the eigenvectors according to

$$\gamma^x_{(i)} := a^x_{(i)} \cdot \gamma^x_{(i)} + (1 - a^x_{(i)}) \cdot \gamma^x_{(i-1)}$$

Thus, the higher the value of $k$, the more the curve is forced to move straight on. We recommend to set set $k = 1$ or 2. For higher values of $k$ the local principal curve looses too much flexibility.

### 5.3.4 Multiple initializations

Assume that the data cloud consists of several branches, which might or might not be connected. Then one single local principal curve will fail to describe the whole data set, but will only find one branch. This is a problem inherent to all global principal curve algorithms. In our approach this problem can be solved by doing multiple initializations, i.e. we choose subsequently a series of starting points, so that finally at least one starting point is situated on each branch, and perform the algorithm for each starting point. In this manner the whole data cloud will be covered by the local principal curve. The starting points can be imposed by hand on each of the branches, or, if this is not possible or too cumbersome, they might be chosen randomly. If one has for example two disconnected branches of the data cloud, which contain more or less the same amount of data, then the application of four randomly chosen starting points already effects that with 93.75% probability at least one starting point is on each cloud. For an arbitrary number of branches, Borel-Cantelli's Lemma tells us that with the number of starting points increasing to infinity, each branch is visited with probability 1. In practice this technique proves to work satisfactory, even for a high number of branches. To conclude, for a set of starting points $S_0$, we add a 7th step to the algorithm:

7. If $S_0 \neq \emptyset$, choose (without replacement) a new starting point $x_{(0)} \in S_0$ and start again with step 1.

It should be noted that our algorithm is deterministic given the starting points, but yields different principal curves for different starting points. However, since in each case the local centers of mass of the same data are calculated, differences of principal curves on the same branch are usually neglectable. In contrary, KKLZ's implementation of their algorithm is strongly indeterministic, and that even for equal starting conditions.

## 5.4 Examples

### 5.4.1 2-dimensional data

Firstly, we compare the results of our algorithm with some standard examples which were also examined by KKLZ (In this and the following examples, the curves from KKLZ and BR are obtained via the Principal Curves Java program from Balázs Kégl, available at `http://www.iro.umontreal.ca/~kegl/research/pcurves/`. The HS curves were obtained by Hastie's S-Plus function `http://lib.stat.cmu.edu/S/principal.curve`). We start with a circle with radius $r = 1$, which is contaminated with bivariate uncorrelated Gaussian noise with variance 0.04 in each component. The result is demonstrated in Figure 5.2.

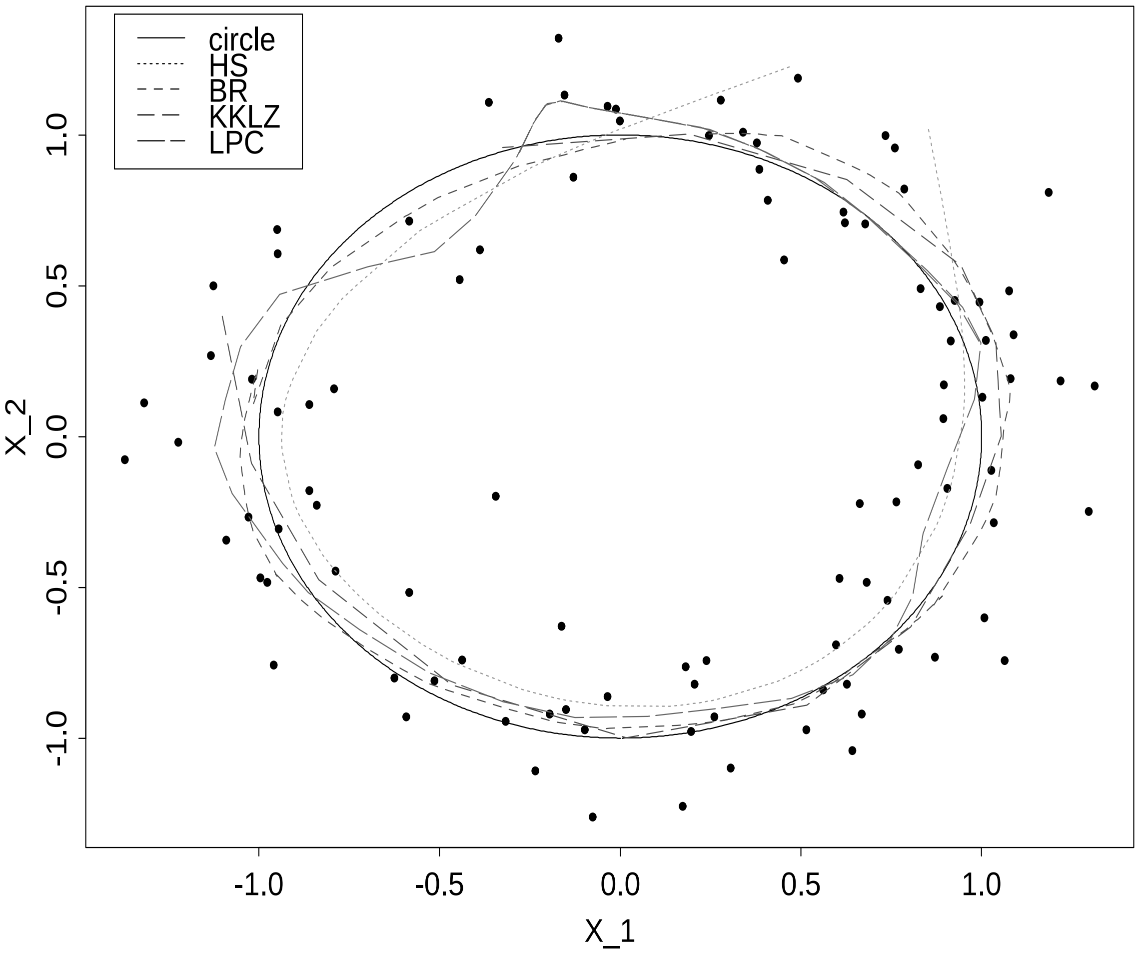



Figure 5.2: Local principal curve for an underlying circle in comparison to other principal curve algorithms.

We notice that only the BR and the proposed local principal curve (hereafter: LPC) algorithm produce a closed curve, whereas HS and KKLZ lead to an open curve. The LPC curve seems to be a bit wiggly in comparison to the other curves, but it should be noted that the LPC approach is fully nonparametric and is only steered by the data, but not by an initial line like the other approaches. This leads to more flexibility (looking at the data, the bump in the left top is not unlikely to be a real feature of the distribution) at the price of a higher variance.

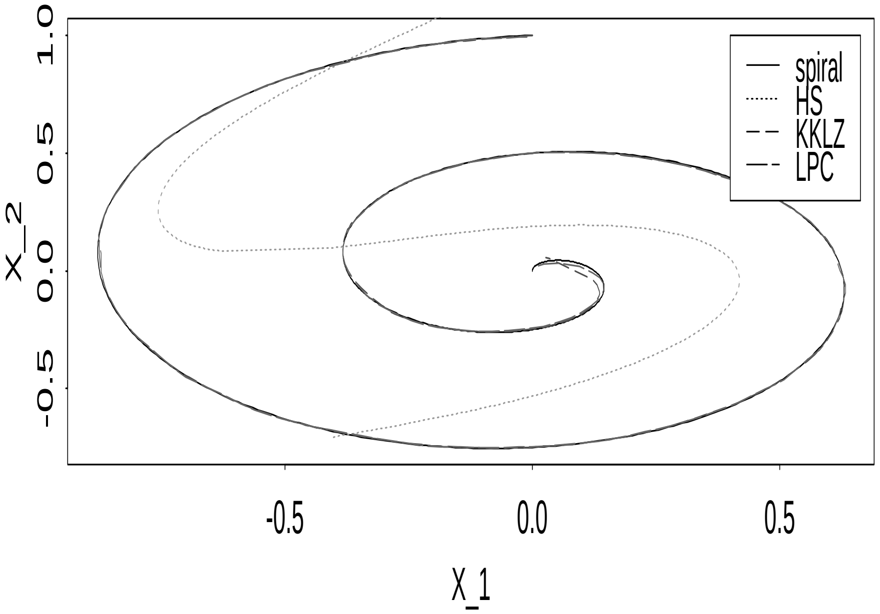
Secondly, we examine the spiral data from KKLZ, Figure 10, b) and c) (where the contaminated big spiral is newly simulated). The standard deviation of the noise is equal to 0.01 for both spirals, and in in each experiment 1000 data points were generated. The small spiral, see Figure 5.3, is found nearly perfectly by KKLZ and LPC, however the HS algorithm shows a fairly bad performance here. The big spiral is only found by LPC. KKLZ's polygonal line algorithm fails here and yields erratic results, which differ in each run of the algorithm. The result of HS is even worse (compare KKLZ, page 21, Figure 11.).

Finally, we consider real data recorded by the Office of Remote Sensing for Earth Resources, Pennsylvania State University, which show the location of floodplains in Beaver County, PA, USA, 1996 (Figure 5.4). For analyzing the data, we digitalized the map to a grid of $106 \times 70 = 7420$ digits. Figure 5.5 shows the result of a run of the LPC algorithm using the digitalized floodplain data. We used 50 initializations and a bandwidth $h = 1.5$ (The automatic selection routine from Section 5.7 suggests 2.5, but a smaller bandwidth seemed more appropriate in this case.). The principal curve uncovers nicely the principal courses of the floodplains. Taking a look at maps from Beaver county, we see that our principal curve reconstructs the underlying rivers resp. valleys in this district (The data as well as corresponding maps are available at PASDA - Pennsylvania Spatial Data Access, `www.pasda.psu.edu`. The best form to regard those maps is to open the ArcExplorerWeb at `http://www.esri.com/software/arcexplorer` and search in the opening menue for Pennsylvania Spatial Data Access, "PA Streams" or "PA Floodplains"). Note that a quite big cluster in the central bottom is not covered - this simply occurs because none of the randomly chosen starting points is situated there, and this isolated cluster cannot be reached by an external principal curve. More initializations would be necessary to solve this.

### 5.4.2 3-dimensional data

We now consider a data set included in the Splus software package, namely the "radial velocity of galaxy NGC7531". This data frame, recorded by Buta (1987), contains the radial velocity of 323 points of that spiral galaxy covering about 200 arc seconds in north-south and 135 arc seconds in east-west direction in the celestial sphere. All the measurements lie within seven slots crossing the origin. The x- and y-coordinate describe the east-west resp. north-south coordinate, and the z-coordinate is the radial velocity measured in km/sec. For simplicity, we only consider the first 61 data points of the data set (this corresponds to two slots crossing the origin).

Since the data are now situated on two (connected) branches, we need to inititialize more than once. We choose to initialize 4 starting points. We apply an angle penalization using $k = 2$, which serves to keep the curve on the correct slot at the crossing. The result is shown in Figure 5.6.



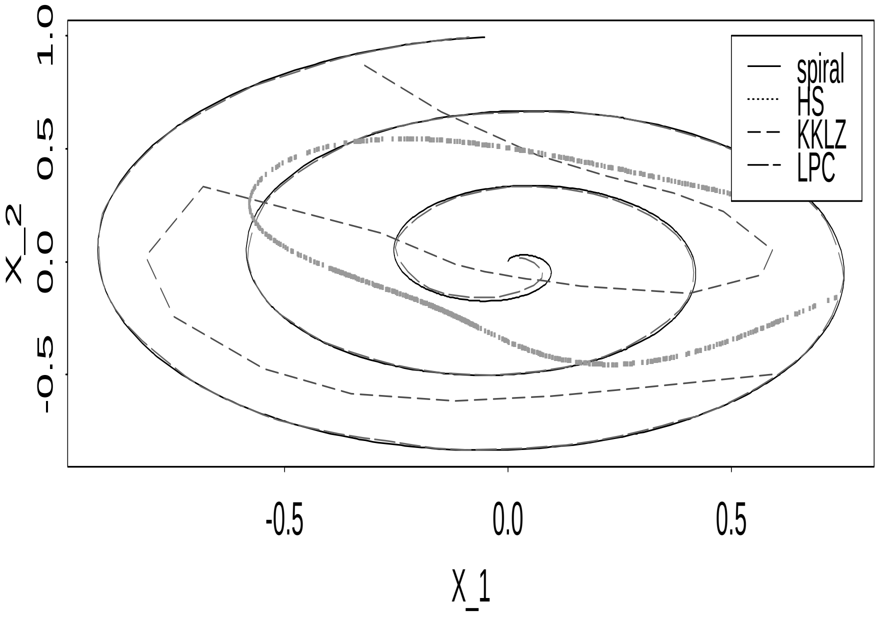




Figure 5.3: Local principal curve for underlying small and big spirals in comparison to other principal curve algorithms.

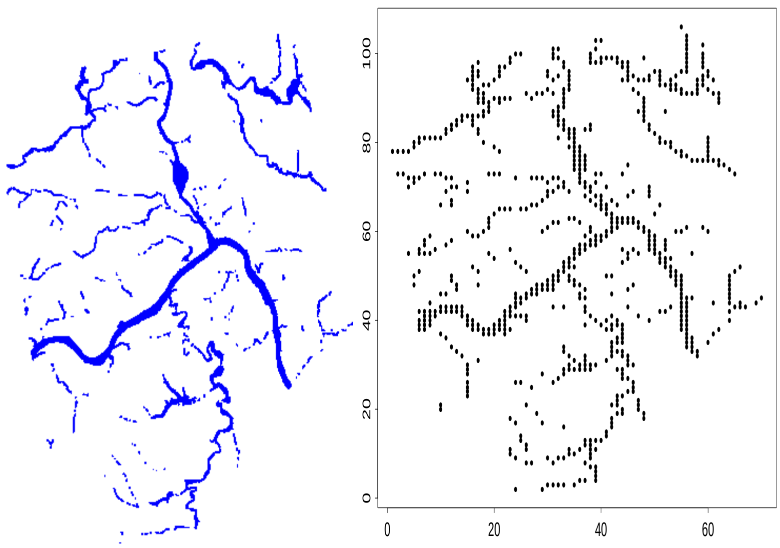



Figure 5.4: Floodplains in Beaver County, PA. (left: original, right: digitalized).

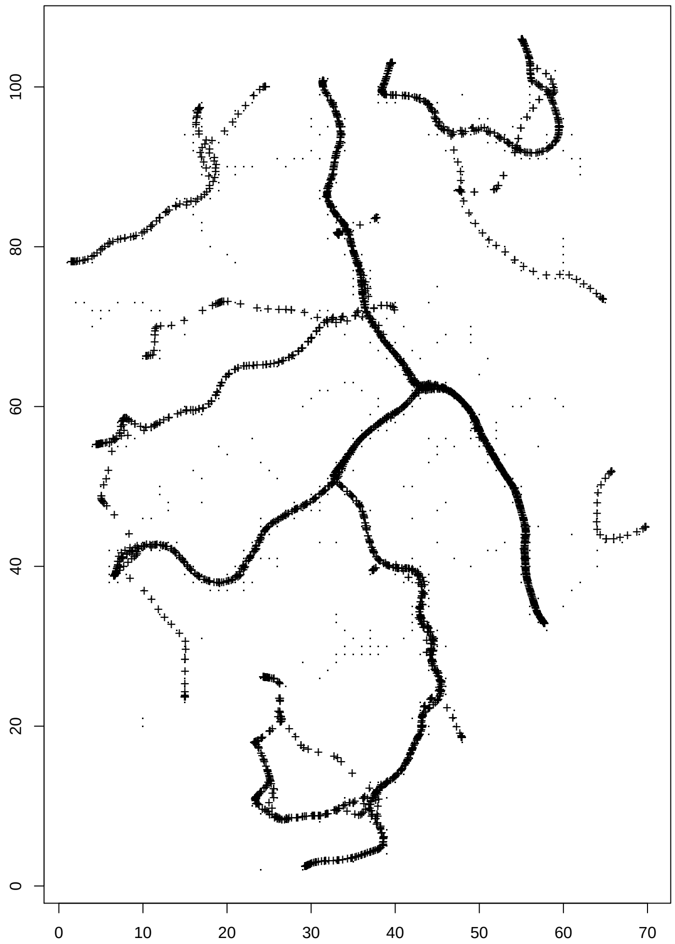



Figure 5.5: Floodplain data (.) with principal curves (+).

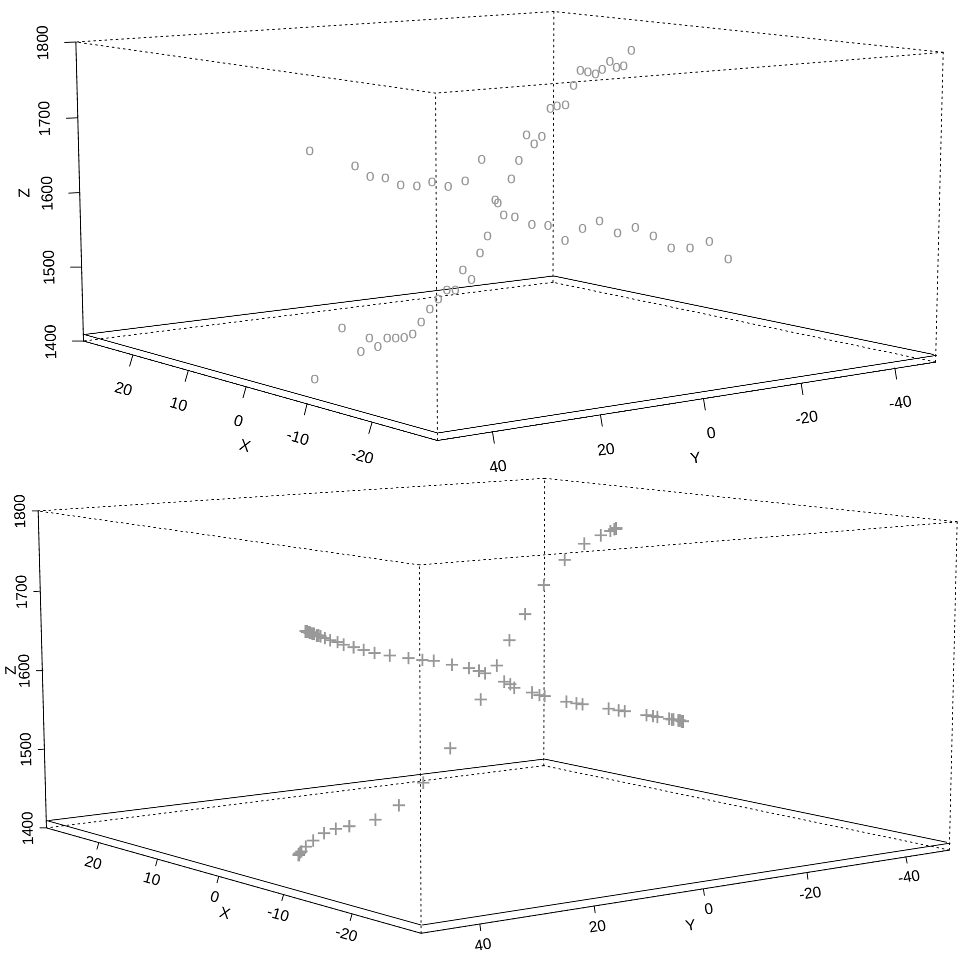



Figure 5.6: Galaxy data (o) with principal curves (+).

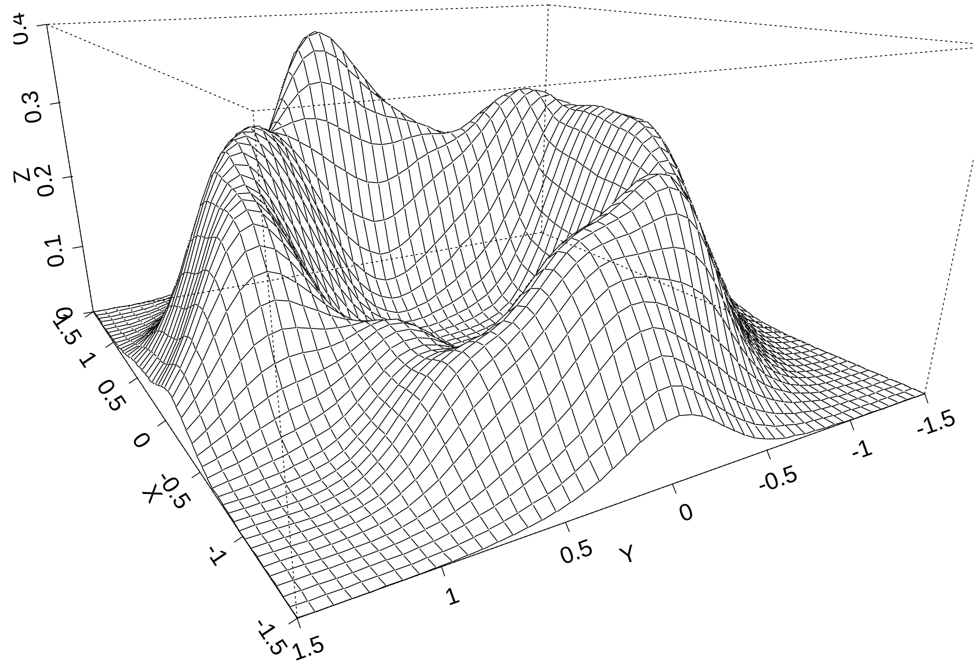



Figure 5.7: Kernel density estimation of simulated circle data.

## 5.5 Theoretical justification

This approach seems to be heuristic to some extent, since we have provided neither a model for the data nor a mathematical precise definition of a local principal curve. In this section we will give some idea which curve we are actually estimating via the LPC algorithm. When we started to work on principal curves, we were not primarily influenced by Delicado's work, but were guided by a simple and appealing idea. It is instructive to take a look at the circle data in Section 4.1. A kernel density estimation yields Figure 5.7.

Looking at this figure, the course of the principal curve is easy to imagine - one simply has to walk along the crest of the mountain. Unfortunately this crest line, which everybody is able to draw rapidly with a pencil, is mathematically intractable. To our knowledge, there exists no mathematical definition of a crest point. However, we will argue in the following that the principal curve we are estimating by means of our algorithm is approximating this crest line.

Comaniciu, Ramesh & Meer (2001) and Comaniciu & Meer (2002), among others, study the properties of the so-called *mean shift vector*

$$M(x) = \mu(x) - x, \tag{5.7}$$

with $\mu(x)$ being the local center of mass (5.3). They provide two results which are of interest for us:

For a Gaussian kernel $K$ and a bandwidth matrix $H = \{h^2 \cdot I : h > 0\}$,

(A) the mean shift vector (5.7) is proportional to the estimated gradient function $\hat{\nabla} f_K(x)$, where the estimated gradient is the gradient of the kernel density estimator

$$\hat{f}_K(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{5.8}$$

(B) the sequence

$$\begin{aligned} m^{(0)} &= x \\ m^{(k+1)} &= \mu(m^{(k)}) \end{aligned}$$

converges to a nearby point where the estimator (5.8) has zero gradient, i.e. to a mode candidate of the kernel density.

For other kernels these statements continue to hold under certain conditions, if in (5.8) $K$ is substituted by its *shadow*, see Comaniciu & Meer (2002).

Let us return to our algorithm now. For any point $x$, we can calulate a local center of mass $\mu(x)$ via function (5.3). It is easy to imagine, that the closer $x$ is to the middle of the distribution of the data, the smaller is the mean shift. According to (5.7) this mean shift is zero for the fix point set of function (5.3), i.e. the set

$$\{x|\mu(x) = x\}$$

(what shows another analogy of our concept to that of Delicado). Considering (A) we realize that these are the points where the estimated gradient function of the density is minimized, which is the case for modes of the estimated density. By applying (B), we thus have a tool for estimating the modes of the density of the data. This is however not our intention: An algorithm like this would get stuck at the modes and be unable to *connect* the modes in a proper way. Therefore, in each step of the algorithm, we employ only the first loop of the iterative process (B), which brings us near the crests, but not nesessarily in a mode point. Then, for not getting stuck in a mode, we move along a little step in direction of the local principal component (what means in practice: along a crest). If thereby, after one or more steps, a point $x_{(k)}$ is approached which is near to a new mode, then the local center of mass $\mu(x_{(k)})$ will tend to this mode, as the following (quite trivial) lemma shows. If no further modes exists, the algorithm will stop itself when the end of the data cloud is reached.

**Lemma 5.1.** *Let $X_i \in \mathbb{R}^d, i = 1, \ldots, n$ be a data cloud and $H$ be a bandwidth matrix. Let $\mu_0$ a fix point of (5.3) resp. $H$ and $x \longrightarrow \mu_0$. Then, applying the same bandwidth matrix $H$, we have convergence $\mu(x) \longrightarrow \mu_0$.*

**Proof**

$$|\mu(x) - \mu_0| = \left| \frac{\sum_{i=1}^n K_H(X_i - x)X_i}{\sum_{i=1}^n K_H(X_i - x)} - \frac{\sum_{i=1}^n K_H(X_i - \mu_0)X_i}{\sum_{i=1}^n K_H(X_i - \mu_0)} \right| \longrightarrow 0$$

for continuous non-zero kernel-functions.

## 5.6 Coverage

There is need for some criterion to evaluate the performance of a principal curve. This is usually done by means of a quantitative measure as the expected squared distance

$$\triangle(m) = E\left(\inf_t ||X - m(t)||^2\right) \tag{5.9}$$

between data $X$ and the curve $m$. Principal curves according to HS are critical points of (5.9), whereas principal curves by KKLZ are minimizing (5.9) over a class of curves with bounded length. Another quantitative measure is the generalized total variance (Delicado, 2001). However, definitions of this type are connected to an underlying stochastic model for the data, which is not used in our case. Therefore we propose a model-independent criterion to assess the quality of a principal curve. We define the *coverage* of a principal curve $m$ by the fraction of all data points which are situated in a certain neighborhood of the principal curve. More precisely, let a principal curve algorithm select a principal curve $m$ consisting of a set $P_m$ of points. Then

$$C_m(\tau) = \#\{x \in X | \exists p \in P_m \text{ with } ||x - p|| \leq \tau\}/n$$

is the coverage of curve $m$ with parameter $\tau$. Obviously the coverage is a monotone increasing function of $\tau$ and will reach the value 1 for $\tau$ tending to infinity. Note that the coverage can be interpreted as the empirical distribution function of the "residuals", i.e. the shortest distance between data and principal curve. For evaluating the quality of a principal curve fit it is necessary to take a look at the whole coverage curve $C_m(\tau)$. In Figure 5.8 we provide the coverage plots for the spiral data (Figure 4), each for the HS, KKLZ and LPC algorithms and for principal component analysis. For the small spiral, the coverage of the LPC and the polygonal line algorithm from KKLZ are comparable, whereas HS is falling back significantly and is performing only slightly better than the principal component approach. For the big spiral, the LPC algorithm clearly outperforms all other algorithms.

Certainly a concave coverage curve is desirable, i.e. it is "best" when rising rapidly for small $\tau$. The better the principal curve, the smaller is the area in the left top above the coverage curve, i.e. the area between $C_m(\tau)$, the line $\tau = 0$ and

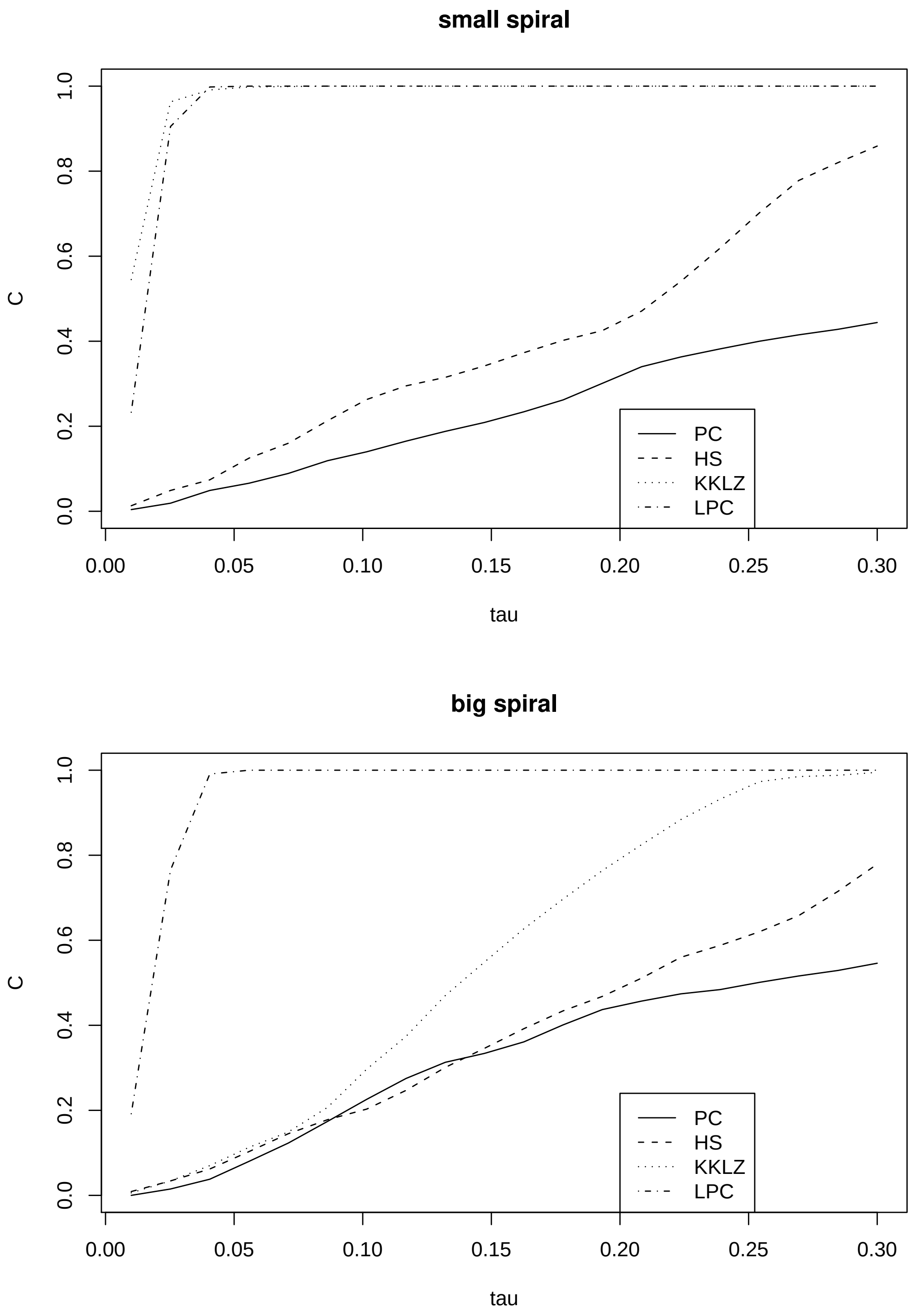



Figure 5.8: Coverage for small (top) and big (bottom) spiral data.

the constant function $c(\tau) = 1$. This area corresponds to the mean length of the observed residuals. To obtain a quantitative measure for the performance of a principal curve, we set this area in relation to the corresponding area obtained by standard principal component analysis. The smaller this quotient, the smaller is the relative mean length of the observed residuals and the better is the principal curve compared to principal components. The following table provides this quotient $A_C$ for HS, KKLZ and LPC, where the latter one is calculated applying the optimal bandwidths according to Section 5.7.

| algorithm | small spiral $A_C$ | big spiral $A_C$ |
|---|---|---|
| HS | 0.79 | 0.92 |
| KKLZ | 0.03 | 0.66 |
| LPC | 0.06 | 0.08 |

Table 5.1: Area-quotient $A_C$ for some principal curve algorithms.

For the HS algorithm, the quotient $A_C$ takes values near 1, which means a quite bad performance. KKLZ yields an excellent value for the small spiral and a rather unsatisfactory value for the big spiral. LPC performs fine in both cases.

## 5.7 Selection of parameters

The algorithm is based on two parameters which have to be selected beforehand: The bandwidth $h$ for the radius of the local center of mass and the value $t_0$ which determines the step length. Assume a center of mass $\mu^x_{(i)}$ at step $i$, using the data within a radius $h$ around a nearby value $x_{(i)}$. Starting from $\mu^x_{(i)}$, it seems sensible to walk along the first principal component $\gamma^x_{(i)}$ until the border of the circle around $x_{(i)}$ is reached, what leads roughly to the update rule

$$x_{(i+1)} = \mu^x_{(i)} + h\gamma^x_{(i)}.$$

This means that we employ

$$t_0 = h.$$

This rule works fine in practice and was applied in all examples in this paper. It now remains to select the value $h$, which plays the role of a classical smoothing parameter, thus the smaller the value of $h$, the more details are unveiled by the local principal curve and the more wiggly it is. To select $h$, we make use of the concept of coverage introduced in the previous section. The idea is the following: If a certain bandwidth $h$ is supposed to serve satisactory for calculating the local center of mass around $x$, we assume implicitely that this value $h$ covers more or less the width of the data cloud around $x$. Thus, as a criterion for the adequacy of a principal curve $m(h)$ calculated with a certain bandwidth $h$ we can apply its proper coverage $C_{m(h)}(h)$. We will will refer to this coverage as *self-coverage* hereafter. This curve has a typical behaviour: It starts with small values, then increases rapidly until a local maximum is reached, where the best fit is achieved. Afterwards the self-coverage curve is *falling* again or shows erratic behaviour, but finally rises up to 1 since for large bandwidths the coverage naturally takes the value 1. Note that the fact that $C_{m(h)}(h)$ is falling is not in contradiction with the property of monotoneness mentioned in the previous section: In contrast to $C_{m(h)}(h)$, the coverage $C_m(\tau)$ is calculated for the *same* principal curve $m$ for all $\tau$! Our parameter selection rule is the following:

Choose the *lowest* parameter $h$ for which

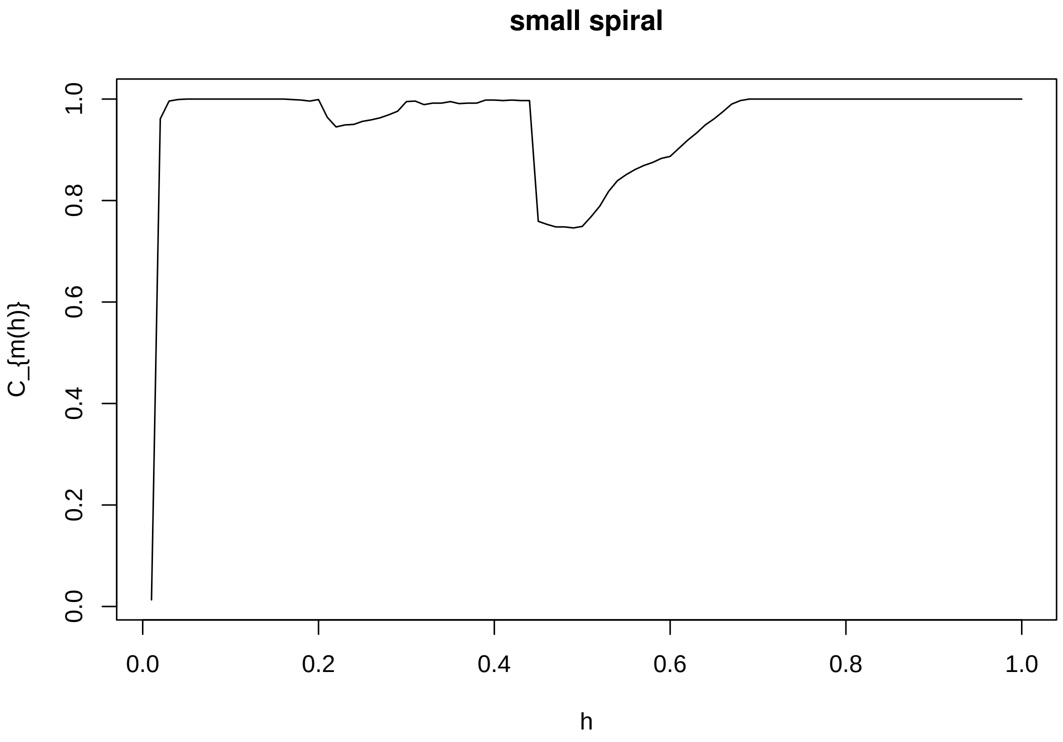
- the function $C_{m(h)}(h)$ achieves its first local maximum,
- or, if no local maximum exists, the function $C_{m(h)}(h)$ achieves the value 1.

We want to illustrate this methodology by means of the spiral data shown in Figure 5.3. For the small and the big spiral, we calculate the self-coverage over a grid from $h = 0.01$ up to $h = 1.0$ in steps of 0.01. The results are presented in Figure 5.9. Since the maxima are partially very flat, we provide in addition the numeric values (for the crucial range of $h$) in Table 5.2.

| $h$ | small spiral $C_{m(h)}(h)$ | big spiral $C_{m(h)}(h)$ |
|---|---|---|
| 0.01 | 0.013 | 0.177 |
| 0.02 | 0.961 | 0.589 |
| 0.03 | 0.996 | 0.990 |
| 0.04 | 0.999 | 0.997 |
| **0.05** | **1.00** | 0.998 |
| **0.06** | 1.00 | **1.00** |
| 0.07 | 1.00 | 1.00 |
| 0.08 | 1.00 | 1.00 |
| 0.09 | 1.00 | 1.00 |
| 0.1 | 1.00 | 0.998 |

Table 5.2: Self-coverage for spiral data.

For the small spiral, the first local maximum is achieved at value $h = 0.05$ with $C_{m(0.05)}(0.05) = 1$. Thus we choose $h = 0.05$. For the whole span from $h = 0.05$ to $h = 0.16$ we would however obtain an ideal principal curve (see the flat maximum). Afterwards the self-coverage is unstable and partially deteriorating. For large values



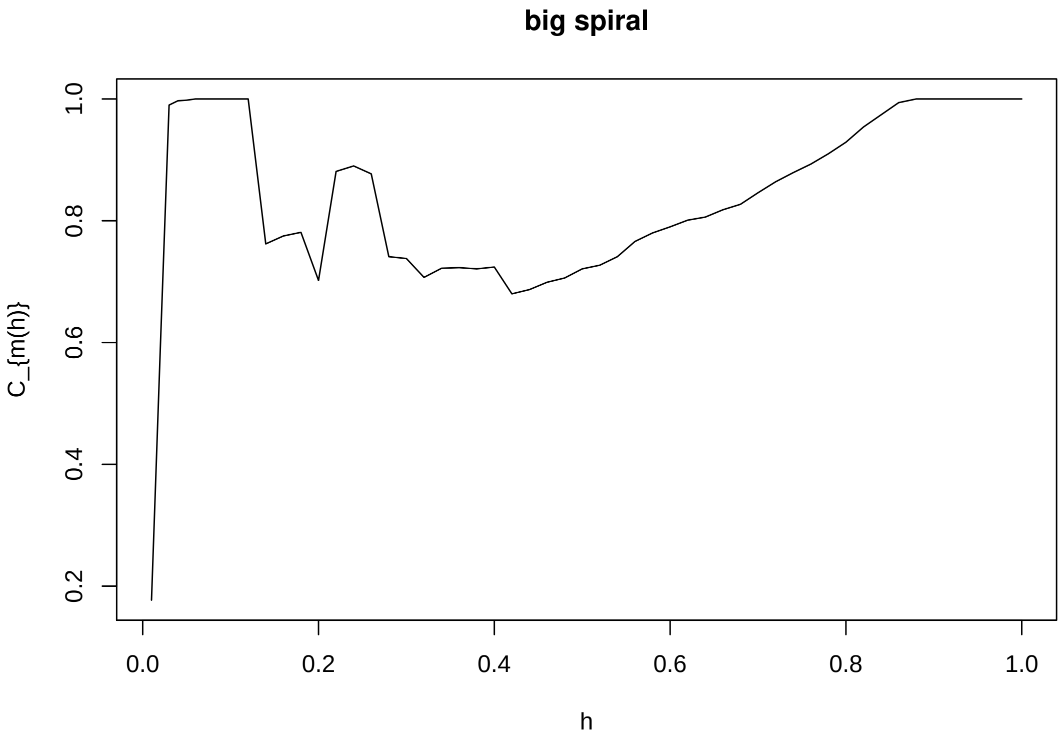




Figure 5.9: Self-coverage for spiral data.

of $h$ the self-coverage tends to 1. The big spiral data lead to a first local maximum starting at $h = 0.06$. Afterwards the curve shows erratic behaviour and approaches slowly to the constant value 1, which is reached at $h = 0.88$. In these calculations, we worked with one fixed starting point (more starting points should not be necessary, since the data cloud is connected and consists of only one branch).

## 5.8 Discussion

We demonstrated that the concept of applying local principal components in connection with the mean shift is a simple and useful tool for calculating principal curves, which shows mostly superior performance in simulated data sets compared to other principal curve algorithms. We showed that the algorithm works in simulated and real data sets even for highly complicated data structures. This includes data situations which yet could only be handled unsatisfactorily, as data with multiple or disconnected branches. Especially for noisy spatial data as the floodplain data the approach has a high potential to detect the underlying structure. We further provided a tool to select the necessary parameters in a data-adaptive way.

There is still need for further research concerning the theoretical background of the method. Though working fine, we still don't have a theoretical justification why we use *local principal components* to connect the modes of the density. This choice is sensible but in no way unique, and there seem to be many alternatives, such as the extrapolation of the already estimated part of the curve. Due to the nice properties of the mean shift, it might even work to use a line in an *arbitrary* direction, as long it is not orthogonal to the principal curve in the observed point. Important is simply that a movement is made - the mean shift will afterwards adjust the principal curve in direction of the "middle" of the data cloud. However, by applying local principal components the algorithm is fastest, most stable, and the results are as intuitively expected. We consider the first local principal component to be a (biased) approximation of the tangent to the crest line: One can easily derive from its definition that the first local principal component around $x$ is the line through $\mu^x$ which minimizes the weighted distance between the $X_i$ and the line, using the weights $k_i$ as in (5.4). The first local principal component is therefore that line that locally gives the best fit. For a more extensive treatment of this problem see Evers (2003).

Furthermore, it will be interesting to investigate if the proposed algorithm can be extended to obtain local prinicpal surfaces or even local principal manifolds of higher dimensions. This might be a quite difficult job, since yet easy techniques as the signum flipping or the mean shift will probably not be transferable to higher dimensional curves without cumbersome extra work.

## Final remarks

The content of this chapter is joint work with Gerhard Tutz and Ludger Evers from the University of Munich, Germany.

# Chapter 6

# Perspectives

In this chapter the most interesting open questions arising from this thesis are recalled, with the intention of providing a brief overview of possible starting points for further research.

In Chapter 2 remain the following open questions:

- Is it possible to derive an (asymptotic) solution for the optimal bandwidths used in the pre-fit algorithm, which takes into account that the basis in the second fit is itself a random variable?
- Do other basis functions exist, which allow the derivation of another extension of Taylor's formula (in other words: another special case of Widder's expansion), and thus enable interpretation of the parameters and computation of the bias of the local approximation?
- Moreover, another point of view might be interesting: We started with local polynomials and tried to improve the fit by replacing the basis. The other side of the coin is to start with a commonly used basis, e.g. B-Splines, and investigate the effect of localization. Finally one should obtain the same results.

Concerning Chapter 3, we are interested in

- how different bandwidth selection routines are influenced by horizontal outliers?
- how (local) smoothers with weights $\alpha = f, \alpha = 1, \alpha = 1/f$ behave for rising sample sizes $n$ ? For small $n$, $\alpha = f$ should be preferred, while for $n \longrightarrow \infty$ the setting $\alpha = 1/f$ should be superior.

Both these points are cases for simulation studies and thus suitable topics for diploma theses. Moreover, there remain some conceptual (if not: philosophical) questions:

- To what extent does "robustification" mean "manipulation"? Is the proposed method possibly a useful tool for hiding any undesired effects?
- Does the found analogy between local smoothing and sample theory reflect a general statistical principle? Does the theory favor the "weaker", "less probable", or "needy" data, while practice shows that it is better to rely on the majority, which provides the less risky information?

Although Chapter 4 does not pose any mentionable questions, there are interesting extensions that are worth consideration:

- In principle, the technique is easily extended to the multivariate case. For a large number of simultaneously observed time series, how can regard be paid to the dependence structure without high computational costs? Is it possible to perform an online dimension reduction? First approaches to answer these questions are given by Gather, Fried & Lanius (2003).

The most perspectives for further research undoubtedly stem from Chapter 5. Regarding the concept presented there, we left unanswered

- how our empirical principal curve could be embedded into a suitable statistical model,
- and how can be better reasoned why we use local principal components to walk along the crest of the density mountain?

For further conceptual research, the following is of particular interest:

- How can the principal curve be used in practice, e.g. can it be employed to reduce the dimension in the predictor space of a high dimensional regression problem?
- How might the second, third, ... local principal component be exploited? It is an obvious idea to use this information to detect the bifurcation points.
- How can the method be extended to find principal surfaces, etc. ?

A variety of ideas and approaches to the above mentioned topics are presented in Evers (2003). There remains one final point to mention:

- Is the proposed concept a useful starting point to construct a framework for non-functional regression, i.e. regression for situations where for a given value of $x$ more than one value of $y$ exists? This is an upcoming, intriguing area of research, and should receive rising attention in the future.

Regarding this list, one might have the impression that this thesis ended up with more open problems than answered questions. This is true to some extent, and is probably true for any scientific work. We can be quite happy about that, since otherwise Schiaparelli (1866) would already have made everything that happened in the last 137 years and 129 pages of this thesis superfluous...

# References


Aluja-Banet, T. and Nonell-Torrent, R. (1991). Local principal component analysis. *Qüestioó* **3**, 267–278.

Banfield, J. D. and Raftery, A. E. (1992). Ice flow identification in stallite images using mathematical morphology and clustering about principal curves. *J. Amer. Statist. Assoc.* **87**, 7–16.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes, Theory and Application*. Englewood Cliffs, New Jersey, USA: Prentice Hall.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion). In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Reinhart and Winston.

Bauer, M., Gather, U., and Imhoff, M. (1999). Analysis of high dimensional data from intensive care medicine. In R. Payne & P. Green (Eds.), *Proceedings in Computational Statistics*, pp. 185–190. Springer-Verlag, Berlin.

Bellmann, R. E. (1961). *Adaptive Control Processes*. Princeton: Princeton University Press.

Bernstein, S. (1914). Sur la définition et les propriétés dos fonctions d'une variable réelle. *Mathematische Annalen* **75**, 449–468.

Blumenthal, L. M. (1926). Concerning the remainder term in Taylor's formula. *American Mathematican Monthly* **33**, 424–426.

Boas, R. P. (1970). Generalized Taylor series, quadrature formulas, and a formula by Kronecker. *SIAM Review* **12**, 116–119.

Braga, A. L. F., Saldiva, P. H. N., Pereira, L. A. A., Menezes, J. J. C., Conceição, G. M. S., Lin, C. A., Zanobetti, A., Schwartz, J., and Dockery, D. W. (2001). Health effects of air pollution exposure on children and adolescents in São Paulo, Brazil. *Pediatr Pulmonol* **31**, 106–113.

Brewer, K. (2002). *Combined Survey Sampling Inference*. London: Arnold.

Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer-Verlag, Berlin, New York.

Buta, R. (1987). The structure and dynamics of ringed galaxies, III: Surface photometry and kinematics of the ringed nonbarred spiral NGC 7531. *The*

*Astrophysical Journal Supplement Series* **64**, 1–137.

Carroll, R. J. (2003). Longitudinal and clustered data and non/semiparametric regression. Talk on 14th of May 2003 at the Department of Statistics, University of Munich.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.

Cauchy, A.-L. (1821). *Cours d'analyse de l'École Polytechnique.* Paris: L'Imprimerie Royale.

Chang, K. and Ghosh, J. (1998). Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III* **3307**, 120–129.

Choi, E. and Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika* **85**, 333–345.

Chu, C. K., Glad, I. K., Godtliebsen, F., and Marron, J. (1998). Edge-preserving smoothers for image processing (with discussion). *J. Amer. Statist. Assoc.* **93**, 526–541.

Chu, C. K. and Marron, J. (1991). Choosing a kernel regression estimator (with discussion). *Statistica Sinica* **6**, 404–436.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83**, 596–610.

Cleveland, W. S. and Grosse, E. (1991). Computational methods for local regression. *Statist. Comput.* **1**, 47–62.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Machine Intell.* **24**, 603–619.

Comaniciu, D., Ramesh, V., and Meer, P. (2001). The variable bandwidth mean shift and data-driven scale seclection. In *Proceedings 8th International Conference on Computer Vision*, Vancouver, BC, Canada, pp. 438–445.

Conceição, G. M. S., Miraglia, S. G. E. K., Kishi, H. S., Saldiva, P. H. N., and Singer, J. M. (2001). Air pollution and children mortality: a time series study in São Paulo, Brazil. *Environ Health Perspect* **109**, 347–350.

Daumer, M. (1997). Online monitoring of change points. *Biomedizinische Technik* **42**, 95–96.

Daumer, M. (1999). *Verfahren und Vorrichtung zur Erkennung von Driften, Sprüngen und/oder Ausreißern von Meßwerten.* Patent PCT/DE 99/01820.

Daumer, M. and Falk, M. (1998). On-line change-point detection (for state space models) using multi-process kalman filters. *Linear Algebra and its Applica-*

*tions* **284**, 125–135.

De Forest, E. L. (1873). On some methods of interpolation applicable to the graduation of irregular series, such as tables of mortality. Annual Report of the Board of Regents of the Smithsonian Institution for 1873.

Delicado, P. (2001). Another look at principal curves and surfaces. *Journal of Multivariate Analysis* **77**, 84–116.

Delicado, P. and Huerta, M. (2003). Principal curves of oriented points: Theoretical and computational improvements. *Computational Statistics* **18**, 293–315.

Ding, X. (1998). On random Taylor series. *Wuhan Univ. Journal of Natural Science* **3**, 257–260.

Doksum, K., Petersen, D., and Samarov, A. (2000). On variable bandwidth selection in local polynomial regression. *Journal of the Royal Statistical Society, Series B* **62**, 431–448.

Einbeck, J. (2003). Multivariate local fitting with general basis functions. *Computational Statistics* **18**, 185–203.

Einbeck, J. and Kauermann, G. (2003). Online monitoring with local smoothing methods and adaptive ridging. *J. Statist. Comput. Simul.* **73**, 913–929.

Evers, L. (2003). *Partitionierungsverfahren zur Dimensionsreduktion.* Diploma thesis, Universität München.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate statistische Verfahren.* Berlin / New York: de Gruyter.

Fahrmeir, L. and Künstler, R. (1999). Penalized likelihood smoothing in robust state space models. *Metrika* **49**, 173–191.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer Verlag.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *Journal of the Royal Statistical Society, Series B* **57**, 371–395.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman and Hall.

Fan, J., Gijbels, I., Hu, T.-C., and Huang, L.-S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica* **6**, 113–127.

Fan, J., Hu, T.-C., and Truong, Y. K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics* **21**, 433–446.

Fan, J. and Marron, J. (1993). Comments on "Local regression: Automatic kernel carpentry". *Statistical Science* **8**, 129–134.

Fan, J. and Marron, J. (1994). Fast implementations of nonparametric curve estimates. *Journal of Computational and Graphical Statistics* **3**, 35–56.

Firey, W. J. (1960). Remainder formulae in Taylor's theorem. *American Mathematican Monthly* **67**, 903–905.

Forster, O. (1999). *Analysis 1*. Braunschweig/Wiesbaden: Vieweg.

Fowlkes, E. B. (1986). Some diagnostics for binary logistic regression via smoothing (with discussion). *Proceedings of the Statistical Computing Section, American Statistical Association* **1**, 54–56.

Gasser, T., Müller, H.-G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B* **47**, 238–252.

Gather, U., Bauer, M., Imhoff, M., and Löhlein, D. (1998). Statistical pattern detection in univariate time series of intensive care on-line monitoring data. *Intensive Care Medicine* **24**, 1305–1314.

Gather, U., Fried, R., and Lanius, V. (2003). Online monitoring of high dimensional physiological time series - a case study. *Estadistica*, (to appear).

Grabert, H. (1982). *Projection Operator Techniques in Nonequilibrium Statistical Mechanics*. Berlin: Springer.

Hall, P. and Jones, M. (1990). Adaptive M-estimation in nonparametric regression. *Ann. Statist.* **18**, 1712–1728.

Hall, P. and Titterington, D. M. (1992). Edge-preserving and peak-preserving smoothing. *Technometrics* **34**, 429–440.

Härdle, W. and Gasser, T. (1984). Robust nonparametric function fitting. *Journal of the Royal Statistical Society, Series B* **46**, 42–51.

Härdle, W., Hall, P., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameter from their optimum? *J. Amer. Statist. Assoc.* **83**, 86–95.

Härdle, W. and Scott, D. W. (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics* **7**, 97–128.

Hart, J. D. and Lee, C.-L. (2002). Robustness of one-sided cross-validation to autocorrelation. Submitted to: Journal of Multivariate Analysis.

Hart, J. D. and Yi, S. (1998). One-sided cross-validation. *J. Amer. Statist. Assoc.* **93**, 620–631.

Hastie, T. and Loader, C. (1993a). Local regression: Automatic kernel carpentry. *Statistical Science* **8**, 120–129.

Hastie, T. and Loader, C. (1993b). Rejoinder to: "Local regression: Automatic kernel carpentry". *Statistical Science* **8**, 139–143.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84**, 502–516.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* New York: Springer.

Henderson, R. (1916). Note on graduation by adjusted average. *Transcations of the Actual Society of America* **25**, 43–48.

Honda, T. (2000). Nonparametric estimation of a conditional quantile for $\alpha$-mixing processes. *Ann. Inst. Statist. Math.* **52**, 459–470.

Horvitz, D. G. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663–685.

Hummel, P. M. and Seebeck, C. L. (1949). A generalization of Taylor's expansion. *American Mathematical Monthly* **56**, 243–247.

Hurvich, C., Simonoff, J., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271–293.

Imhoff, M. and Bauer, M. (1996). Time series analysis in critical care monitoring. *New Horizons* **4**, 519–531.

Jolliffe, I. T. (2002). *Principal Component Analysis, Second Edition.* New York: Springer.

Jones, P. (1951). Brook Taylor and the mathematical theory of linear perspective. *Amer. Math. Monthly* **58**, 597–606.

Kambhatla, N. and Leen, T. K. (1997). Dimension reduction by local PCA. *Neural Computation* **9**, 1493–1516.

Katkovnik, V. Y. (1979). Linear and nonlinear methods of nonparametric regression analysis. *Soviet Automatic Control* **5**, 25–34.

Kauermann, G. (2001). Edge preserving smoothing by local mixture modeling. SFB 386 Discussion Paper 255, Institut für Statistik, Universität München.

Kégl, B. and Krzyzak, A. (2002). Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 59–74.

Kégl, B., Krzyzak, A., Linder, T., and Zeger, K. (2000). Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 281–297.

Königsberger, K. (2000). *Analysis 2.* Berlin/Heidelberg: Springer.

Lagrange, J. L. (1797). *Théorie des fonctions analytique.* Paris: Imprimerie de la

Republique.

Lay, S. R. (1990). *Analysis with an Introduction to Proof.* New Jersey: Prentice Hall.

Lhuilier, S. (1786). *Exposition elementaire des principes des calculs superieurs.* Berlin.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor in measured with/without error. *J. Amer. Statist. Assoc.* **95**, 520–534.

Loader, C. R. (1999a). Bandwidth selection: Classical or plug-in? *Ann. Statist.* **27**, 415–438.

Loader, C. R. (1999b). *Local Regression and Likelihood.* New York: Springer.

Macaulay, F. R. (1931). *Smoothing of Time Series.* National Bureau of Economic Research.

Maclaurin, C. (1742). *Treatise of fluxions.* Edinburgh.

Massay, W. A. and Whitt, W. (1993). A probabilistic generalization of Taylor's theorem. *Statist. Probab. Letters* **16**, 51–54.

McDonald, J. A. and Owen, A. B. (1986). Smoothing with split linear fits. *Technometrics* **28**, 195–208.

Müller, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82**, 231–238.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika 78*(3), 521–530.

Müller, H.-G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27**, 299–337.

Nadaraya, E. A. (1964). On estimating regression. *Theory Prob. Appl.* **9**, 141–142.

O'Connor, J. J. and Robertsen, E. F. (2000). Brook Taylor. *http://www-history.mcs.st-andrews.ac.uk/history/Mathematicians/Taylor.html.*

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85**, 66–72.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**, 559–572.

Raina, R. K. and Koul, C. L. (1979). On Weyl fractional calculus. *Proceedings of the American Mathematical Society* **73**, 188–192.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis.* New York: Springer.

Raßer, G. (2003). *Clustering Partition Models for Discrete Structures with Applications in Geographical Epidemiology.* Ph. D. thesis, Universität München.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257–1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

Schiaparelli, G. V. (1866). Sul modo di ricavare la vera espressione delle leggi delta natura dalle curve empiricae. *Effemeridi Astronimiche di Milano per l'Arno* **857**, 3–56.

Schmidt, G., Mattern, R., and Schüler, F. (1981). Biochemical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column and without protective helmet under effects of impact. Technical Report Project 65 EEC Research Program on Biomechanics of Impacts, Universität Heidelberg, Germany.

Schölkopf, B. and Smola, A. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299–1319.

Schwartz, J. (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Canadian Journal of Statistics* **22**, 471–487.

Seifert, B. and Gasser, T. (1996). Finite-sample analysis of local polynomials: Analysis and solutions. *J. Amer. Statist. Assoc.* **91**, 267–275.

Seifert, B. and Gasser, T. (2000). Data adaptive ridging in local polynomial regression. *Journal of Comput. and Graph. Statistics* **9**, 338–360.

Shiskin, J., Young, A., and Musgrave, J. (1967). The X-11 variant of the census method II seasonal adjustment program. Technical Report 15, Department of Statistics, Stanford University.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* London: Chapman & Hall.

Singer, J. M., André, C. D. S., Lima, P. L., and Conceição, G. M. S. (2002). Association between atmospheric pollution and mortality in São Paulo, Brazil: Regression models and analysis strategy. In Y. Dodge (Ed.), *Statistical Data Analysis based on the L1 norm and related methods*, pp. 439–450. Birkhäuser, Berlin.

Skarbek, W. (1996). Local principal components analysis for transform coding. In *Int. Symposium on Nonlinear Theory and its Applications - NOLTA'96, Proceedings*, Katsurahara-so, Japan, pp. 413–416.

Spencer, J. (1904). On the graduation of rates of sickness and mortality. *Journal of the Institute of Actuaries* **38**, 334–343.

Staniswalis, J. G., Messer, K., and Finston, D. R. (1993). Kernel estimators for multivariate regression. *Nonparametric Statistics* **3**, 103–121.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–

645.

Taylor, B. (1715). *Methodus incrementorum directa et inversa.* London.

Tharpey, T. and Flury, B. (1996). Self-consistency: A fundamental concept in statistics. *Statistical Science* **11**, 229–243.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing* **2**, 183–190.

Trujillo, J. J., Rivero, M., and Bonilla, B. (1999). On a Riemann-Liouville generalized Taylor's formula. *Journal of Mathematical Analysis and Applications* **231**, 255–265.

Truong, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.* **17**, 606–617.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local approximation approach. *Problems of Information Transmission* **22**, 133–146.

Verbeek, Vlassis, N., and Kröse, B. (2001). A soft k-segments algorithm for principal curves. In *Proceedings International Conference on Artificial Neural Networks*, Vienna, Austria, pp. 450–456.

Wahba, G. (1984). Cross-validated spline methods for the estimation of multivariate functions from data on functionals. In H. A. David & H. T. David (Eds.), *Statistics: An Appraisal*, pp. 205–235. Iowa State University Press.

Walsh, J. L. (1929). Note on the expansion of analytic functions in series of polynomials and in series of other analytic functions. *Transactions of the American Mathematical Society* **31**, 53–57.

Wand, M. P. (1992). Error analysis for general multivariate kernel estimators. *Nonparametric Statistics* **2**, 1–15.

Wand, M. P. (2001). Smoothing and mixed models. Talk on 3th of November 2001 at the Euroworkshop on Statistical Modelling in Höhenried, Germany.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–250.

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88**, 520–528.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman & Hall.

Wang, F. T. and Scott, D. W. (1994). The $L_1$ method for robust nonparametric regression. *J. Amer. Statist. Assoc.* **89**, 65–76.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā, Series A,* **26**, 359–372.

Widder, D. V. (1928). A generalization of Taylor's series. *Transactions of the American Mathematical Society* **31**, 43–52.

Widder, D. V. (1929). On the expansion of analytic functions of the complex variable in generalized Taylor series. *Transactions of the American Mathematical Society* **31**, 43–52.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society, Series B* **65**, 95–114.

Yang, L. and Tschering, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B* **61**, 793–815.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.* **93**, 228–237.

Zhao, L. H. (1999). Improved estimators in nonparametric regression problems. *J. Amer. Statist. Assoc.* **94**, 164–173.

**Lebenslauf**

| | |
|---|---|
| *Name* | Jochen Einbeck |
| *Familienstand* | verheiratet mit Flávia Alexandre Einbeck, 1 Kind |
| *Staatsangehörigkeit* | deutsch |
| *Geburtsdatum/-ort* | 13.07.1973 in Hagen (Westfalen) |

*Ausbildung*

| | |
|---|---|
| 08.1980-07.1984 | Grundschule Borstel-Hohenraden (Schleswig-Holstein) |
| 09.1984-07.1993 | Gymnasium Miesbach (Oberbayern) |
| 10.1993-06.1999 | Studium Mathematik/Physik (Lehramt Gymnasium) an der LMU München, Abschluss mit dem ersten Staatsexamen |
| Seit 10.1999 | Promotion in Statistik bei Prof. Dr. Gerhard Tutz (LMU München) |

*Berufliche Tätigkeiten*

- in der Scheckkartenproduktion bei Electronic Payment Cards (Gmund a. T.) im Herbst 1996 und Frühjahr 1997
- Korrektur von Übungsblättern und Klausuren in Statistik und Stochastik am Math. Institut der LMU im WiSe 1997/98 und SoSe 1998
- Seit Oktober 1999 wissenschaftlicher Mitarbeiter am Institut für Statistik der LMU München, 4 Stunden Lehrverpflichtung, seit September 2000 Socrates/ERASMUS Programmbeauftragter

*Auslandsaufenthalte*

- August/September 1999: Englisch-Sprachkurs in Toronto, Kanada
- März 2001: Portugiesisch-Sprachkurs in São Paulo, Brasilien
- März bis Juli 2002: Forschungsaufenthalt an der Universidade de São Paulo, Brasilien