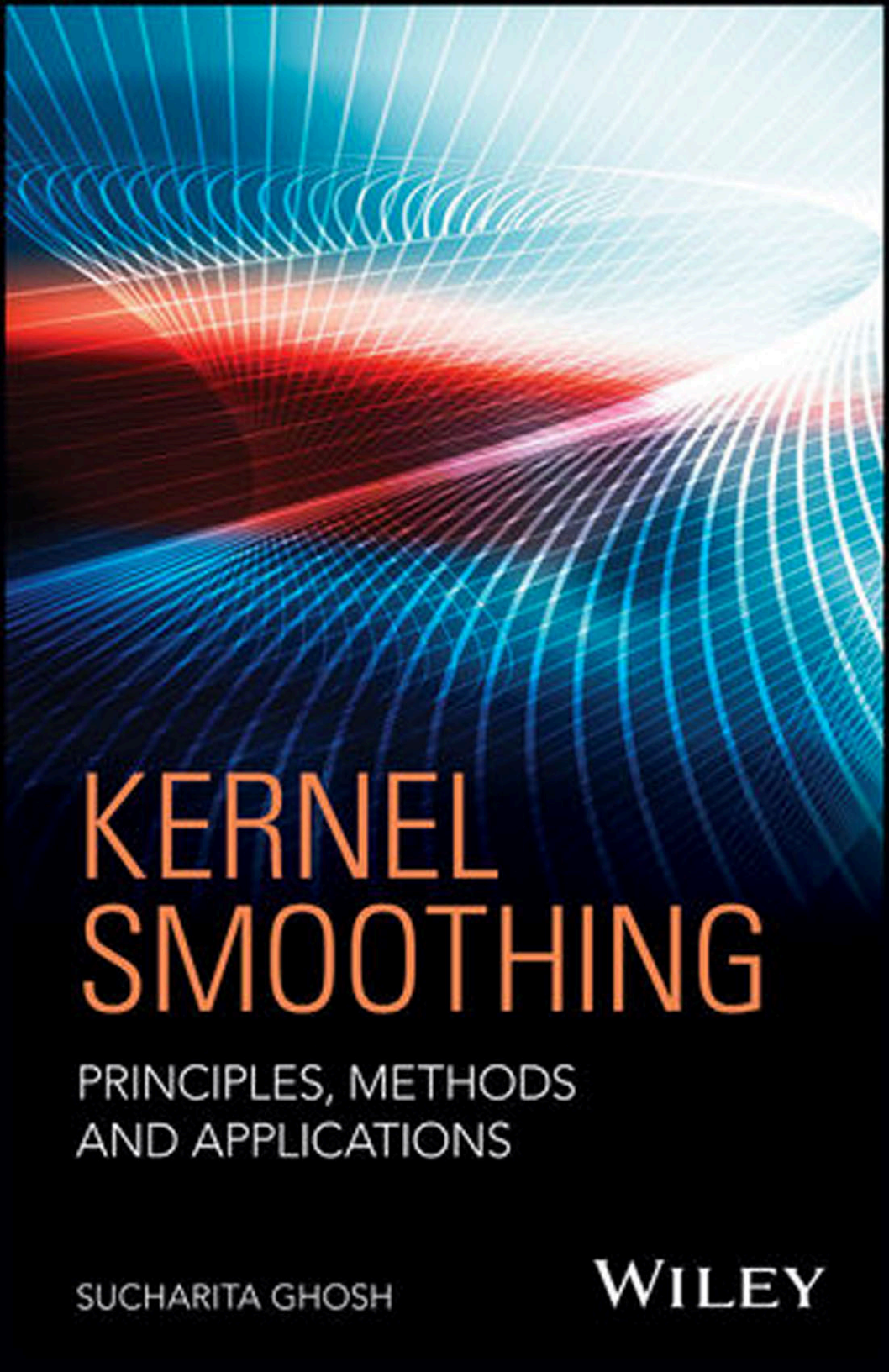
KERNEL
SMOOTHING
PRINCIPLES, METHODS
AND APPLICATIONS
SUCHARITA GHOSH
WILEY

Kernel Smoothing

Kernel Smoothing

Principles, Methods and Applications

Sucharita Ghosh
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland

WILEY

This edition first published 2018
© 2018 by John Wiley & Sons Ltd.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this titleis available athttp://www.wiley.com/go/permissions.

The right of Sucharita Ghosh to be identified as the author of this work has been asserted in accordance with law.

Registered Office(s)
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty
While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Ghosh, S. (Sucharita), author.
Title: Kernel smoothing : principles, methods and applications / by Sucharita Ghosh.
Description: First edition. | Hoboken, NJ : John Wiley & Sons, 2018. | Includes bibliographical references and index. |
Identifiers: LCCN 2017039516 (print) | LCCN 2017046749 (ebook) | ISBN 9781118890509 (pdf) | ISBN 9781118890516 (epub) | ISBN 9781118456057
Subjects: LCSH: Smoothing (Statistics) | Kernel functions.
Classification: LCC QA278 (ebook) | LCC QA278 .G534 2018 (print) | DDC 511/.42–dc23
LC record available at https://lccn.loc.gov/2017039516

Cover Design: Wiley
Cover Image: © PASIEKA/SPL/Gettyimages

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Typically, patterns in real data, which we may call curves or surfaces, will not follow simple rules. However, there may be a sufficiently good description in terms of a finite number of interpretable parameters. When this is not the case, or if the parametric description is too complex, a nonparametric approach is an option. In developing nonparametric curve estimation methods, however, sometimes we may take advantage of the vast array of available parametric statistical methods and adapt these to the nonparametric setting. While assessing properties of the nonparametric curve estimators, we will use asymptotic arguments.

This book grew out of a set of lecture notes for a course on smoothing given to the graduate students of Seminar für Statistik (Department of Mathematics, ETH, Zürich). To understand the material presented here, knowledge of linear algebra, calculus, and a background in statistical inference, in particular the theory of estimation, testing, and linear models should suffice. The textbooks *Statistical Inference (Chapman & Hall)* by Samuel David Silvey, *Regression Analysis, Theory, Methods and Applications (Springer-Verlag)* by Ashis Sen and Muni Srivastava, *Linear Statistical Inference, second edition (John Wiley)* by Calyampudi Radhakrishna Rao, and Robert Serfling's book *Approximation Theorems of Mathematical Statistics (John Wiley)* are excellent sources for background material. For nonparametric curve estimation, there are several good books and in particular the classic *Density Estimation (Chapman & Hall)* by Bernard Silverman is a must-have for anyone venturing into this topic. The present text also includes some discussions on nonparametric curve estimation with time series and spatial data, in particular

with different correlation types such as long-memory. A nice monograph on long-range dependence is *Statistics for Long-Memory Processes (Chapman & Hall)* by Jan Beran. Additional references to this topic as well as an incomplete list of textbooks on smoothing methods are included in the list of references.

Our discussion on nonparametric curve estimation starts with density estimation (Chapter 1) for continuous random variables, followed by a chapter on nonparametric regression (Chapter 2). Inspired by applications of nonparametric curve estimation techniques to dependent data, several chapters are dedicated to a selection of problems in nonparametric regression, specifically trend estimation (Chapter 3) and semiparametric regression (Chapter 4), with time series data and surface estimation with spatial observations (Chapter 5). While, for such data sets, types of dependence structures can be vast, we mainly focus on (slow) hyperbolic decays (long memory), as these types of data occur often in many important fields of applications in science as well as in business. Results for short-memory and anti-persistence are also presented in some cases. Of additional interest are spatial or temporal observations that are not necessarily Gaussian, but are unknown transformations of latent Gaussian processes. Moreover, their marginal probability distributions may be time (or spatial location) dependent and assume arbitrary (non-Gaussian) shapes. These types of model assumptions provide flexible yet parsimonious alternatives to stronger distributional assumptions such as Gaussianity or stationarity. An overview of the relevant literature on this topic is in *Long Memory Processes – Probabilistic Properties and Statistical Models (Springer-Verlag)* by Beran et al. (2013). This is advantageous for analyzing large-scale and long-term spatial and temporal data sets occurring, for instance, in the geosciences, forestry, climate research, medicine, finance, and others. The literature on nonparametric curve estimation is vast. There are other important methods that have not been covered here, such as wavelets – see Percival and Walden (2000), splines (a very brief discussion is included here in Chapter 2 of this book); see in particular Wahba (1990) and Eubank (1988), as well as other approaches. This book looks at kernel smoothing methods and even for kernel based approaches, admittedly, not all topics are presented here, and the focus is merely on a selection.

The book also includes a few data examples, outlines of proofs are included in several cases, and otherwise references to relevant sources are provided. The data examples are based on calculations done using the S-plus statistical package (TIBCO Software, TIBCO Spotfire) and the R-package for statistical computing (The R Foundation for Statistical Computing).

Various people have been instrumental in seeing through this project. First and foremost, I am very grateful to my students at ETH, Zürich, for giving me the motivation to write this book and for pointing out many typos in earlier versions of the lecture notes. A big thank you goes to Debbie Jupe, Heather Kay, Richard Davies, and Liz Wingett, at John Wiley & Sons in Chichester, West Sussex, Alison Oliver at Oxford and to the editors at Wiley, India, for their support from the start of the project and for making it possible. I am grateful to the Swiss National Science Foundation for funding PhD students, the IT unit of the WSL for infallible support and for maintaining an extremely comfortable and state-of-the-art computing infrastructure, and the Forest Resources and Management Unit, WSL for generous funding and collaboration. Special thanks go to Jan Beran (Konstanz, Germany) for many helpful remarks on earlier versions of the manuscript and long-term collaboration on several papers on this and related topics. I also wish to thank Yuanhua Feng (Paderborn, Germany), Philipp Sibbertsen (Hannover, Germany), Rafal Kulik (Ottawa, Canada), Hans Künsch (Zurich, Switzerland), and my graduate students Dana Draghicescu, Patricia Menéndez, Hesam Montazeri, Gabrielle Moser, Carlos Ricardo Ochoa Pereira, and Fan Wu, for close collaboration, as well as Bimal Roy and various other colleagues at the Indian Statistical Institute, Kolkata and Liudas Giraitis at Queen Mary, University of London, for fruitful discussions and warm hospitality during recent academic trips. I want to thank the following for sharing data and subject specific knowledge, which have been used in related research elsewhere or in this book: Christoph Frei at MeteoSwiss and ETH, Zürich, various colleagues at the University of Bern, in particular, Willy Tinner at the Oeschger Centre for Climate Change Research, Brigitta Ammann at the Institute of Plant Sciences and Jakob Schwander at the Department of Physics, as well as Matthias Plattner at Hintermann & Weber, AG, Switzerland and various colleagues

from the Swiss Federal Research Institute WSL, Birmensdorf, in particulear Urs-Beat Brändli, Fabrizio Cioldi and Andreas Schwyzer, all at the Forest Resources and Management unit. Data obtained from the MeteoSwiss, the Swiss National Forest Inventory, the Federal Office of the Environment (FOEN) in Switzerland, and various public domain data sets made available through the web platforms of the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the Meteorological Office, UK (Met Office) used in related research elsewhere or used in this book for methodological illustrations are gratefully acknowledged.

My deepest gratitude goes to my family and friends. I want to thank my family Céline and Jan for being with me every step of the way, making sure that I finish this book at last, my family in India for their unfailing support, our colleagues Suju and Yuanhua for their hospitality on many occasions, Maria, Gunnar, Shila, and Goutam for holding the fort during conferences and other long trips, Wolfgang for his sense of humor, and last but not the least, Sir Hastings, our lovely Coton de Tuléar, for keeping us all on track with his incredible wit and judgment.

Sucharita Ghosh
Birmensdorf

1

Density Estimation

1.1 Introduction

Use of sampled observations to approximate distributions has a long history. An important milestone was Pearson (1895, 1902a, 1902b), who noted that the limiting case of the hypergeometric series can be written as in the equation below and who introduced the Pearsonian system of probability densities. This is a broad class given as a solution to the differential equation

$$\frac{df}{dx} = \frac{(x-a)f}{b_0 + b_1 x + b_2 x^2} \tag{1.1}$$

The different families of densities (Type I–VI) are found by solving this differential equation under varying conditions on the constants. It turns out that the constants are then expressible in terms of the first four moments of the probability density function (pdf) f, so that they can be estimated given a set of observations using the method of moments; see Kendall and Stuart (1963).

If the unknown pdf f is known to belong to a known parametric family of density functions satisfying suitable regularity conditions, then the maximum likelihood (MLE; Fisher 1912, 1997) can be used to estimate the parameters of the density, thereby estimating the density itself. This method has very powerful statistical properties, and continues to be perhaps the most popular method of estimation in statistics. Often, the MLE is the solution to an *estimating equation*, as is also the case for the *method of least squares*. These procedures then come under the general

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

framework of *M-estimation*. Two other related approaches that use ranks of the observations are the so-called *L-estimation* and *R-estimation*, where the statistics are respectively linear combinations of the order statistics or of their ranks. These estimation methods are covered in many standard textbooks. Some examples are Rao (1973, chapters 4 and 5), Serfling (1986, chapters 7, 8, and 9), and Sen and Srivastava (1990).

1.1.1 Orthogonal polynomials

Yet another approach worth mentioning here is the use of Orthogonal polynomials (see Szegő 2003). In this method, the unknown density is approximated by a sum of weighted linear combinations of a set of basis functions. Čencov (1962) provides a general description whereas other reviews are in Wegman (1972) and Silverman (1980). Additional background information and further references can be found in Beran et al. (2013, Chapter 3) and Kendall and Stuart (1963, Chapter 6). The essential idea behind the use of Orthogonal polynomials is as follows (see Rosenblatt 1971):

Suppose that the pdf

$$f : \mathbb{R} \to \mathbb{R} \tag{1.2}$$

belongs to the space $\mathbb{L}^2\{\mathbb{R}, G\}$ of all square integrable functions with respect to the weight function G, i.e.,

$$\int_{-\infty}^{\infty} f^2(x)G(x)\,dx < \infty \tag{1.3}$$

holds, where $\mathbb{R} = (-\infty, \infty)$ denotes the real line. Also, let $\{G_l(x)\}$ be a complete and orthonormal sequence of functions in $\mathbb{L}^2\{\mathbb{R}, G\}$. Then f admits an expansion

$$f(x) = \sum_l a_l G_l(x) \tag{1.4}$$

which converges to f in $\mathbb{L}^2\{\mathbb{R}, G\}$, where a_l is defined as

$$a_l = \int_{-\infty}^{\infty} f(x)G_l(x)G(x)\,dx. \tag{1.5}$$

This formula immediately suggests an unbiased estimator of the coefficient a_l using sampled observations, followed by a substitution in the expansion for f.

As an example, we take a brief look at the Gram–Charlier series representation followed by a further extension due to Schwartz (1967). The Gram–Charlier series of Type A is based on Hermite polynomials H_l and the standard normal pdf ϕ. Note that, for Edgeworth expansion based methods, one would consider the Fourier transform of the product $H_l(x)\phi(x)$ and move on to an expansion that uses the cumulant generating function (see Kendall and Stuart 1963).

First of all consider the pdf f such that it can be expressed as

$$f(x) = \phi(x) \sum_{l=0}^{\infty} c_l H_l(x). \tag{1.6}$$

For conditions under which this is valid, see two theorems due to Cramér quoted in Kendall and Stuart (1963, pp. 161–162) as well as some historical notes in Cramér (1972).

Here ϕ is the standard normal pdf, i.e.,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, x \in \mathbb{R} \tag{1.7}$$

and H_l is the Hermite polynomial of degree l, i.e.,

$$H_l(x) = (-1)^l \frac{1}{\phi(x)} \frac{d^l}{dx^l} \phi(x). \tag{1.8}$$

Using the orthogonality property of the Hermite polynomials, i.e.,

$$\frac{1}{l!} \int_{-\infty}^{\infty} \phi(x) H_l(x) H_m(x)\, dx = 0, \text{if } l \neq m \tag{1.9}$$

$$= 1, \text{if } l = m, \tag{1.10}$$

we have

$$\int_{-\infty}^{\infty} f(x) H_l(x)\, dx = \sum_{j=0}^{\infty} c_j \int_{-\infty}^{\infty} \phi(x) H_j(x) H_l(x)\, dx \tag{1.11}$$

$$= c_l \int_{-\infty}^{\infty} \phi(x) H_l^2(x)\, dx = l! c_l. \tag{1.12}$$

In other words, the coefficients c_l are

$$c_l = \frac{1}{l!} \int_{-\infty}^{\infty} f(x) H_l(x)\, dx = \frac{1}{l!} \mathbb{E}(H_l(X)). \tag{1.13}$$

Due to previous detailed work by Chebyshev, the Hermite polynomials are also known as the Chebyshev–Hermite polynomials. In fact, contributions of Laplace are also known. See Sansone (2004) and Szegő (2003) for additional information.

The above formula for c_l implies that these coefficients may be estimated from a given set of observations $X_1, \ldots, X_n$ from f as sample means of Hermite polynomials, i.e.,

$$\widehat{c}_l = \frac{1}{l!}\frac{1}{n}\sum_{j=1}^{n} H_l(X_j). \tag{1.14}$$

Since with increasing l, estimation of higher order moments are involved, this method however is not optimal. From a statistical view–point, one option is to consider a finite sum.

To this end, Schwartz (1967) considers a pdf f that is square integrable (or simply bounded) and seeks to give an approximation of the form

$$\tilde{f}(x) = \sum_{l=0}^{M_n} d_{l,n} G_l(x) \tag{1.15}$$

where M_n is a sequence of integers depending on the sample size n, $d_{l,n}$ are estimated from observed data, and G_l are Hermite functions

$$G_l(x) = (2^l l! \sqrt{\pi})^{-1/2} e^{-x^2/2} 2^{l/2} H_l(\sqrt{2}x). \tag{1.16}$$

The Hermite functions $G_l(x)$ form a complete orthonormal set over the real line. Examples of Hermite polynomials and Hermite functions are in Figure 1.1 and Figure 1.2. Moreover, due to a theorem of Cramér (see Schwartz 1967), $|G_l(x)|$ is bounded above by a constant that does not depend on x or l. Since f is square integrable, f can be expanded (orthogonal series expansion) as

$$f(x) = \sum_{l=0}^{\infty} d_l G_l(x) \tag{1.17}$$

where

$$d_l = \int_{-\infty}^{\infty} f(x) G_l(x)\, dx = \mathbb{E}(G_l(X)). \tag{1.18}$$

Schwartz (1967) proposes the estimator

$$\widehat{f}(x) = \sum_{l=0}^{M_n} \widehat{d}_{l,n} G_l(x) \tag{1.19}$$

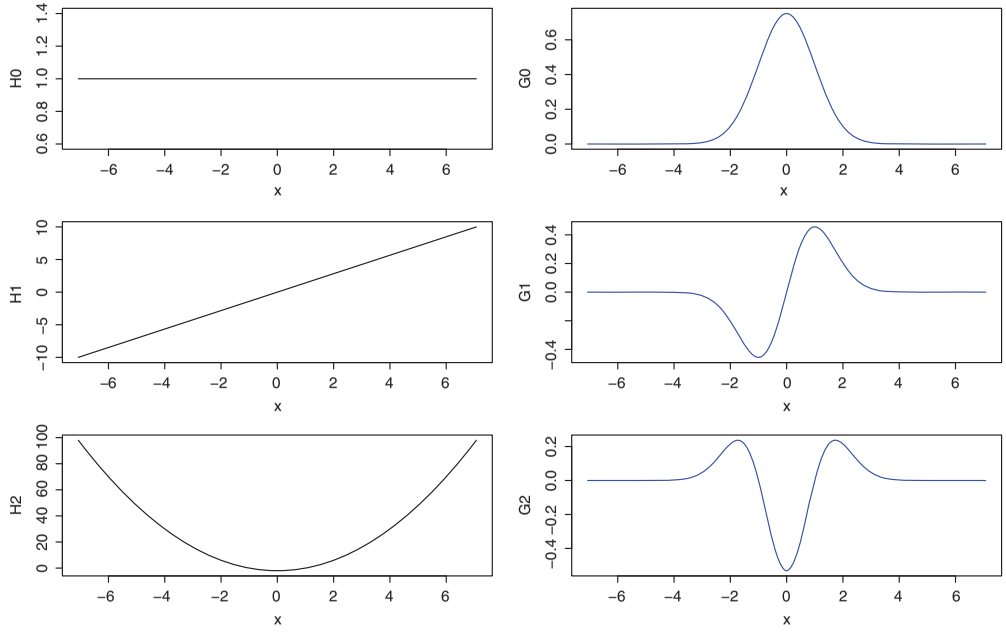

Figure 1.1 Rescaled Hermite polynomials $H_l^{(re)}(x)$ of degree l for $l = 0, 1, 2$ and the corresponding Hermite functions (right) $G_l(x)$. These functions are related via the relation $G_l(x) = (2^l l! \sqrt{\pi})^{-1/2} e^{-x^2/2} H_l^{(re)}(x)$, where $H_l^{(re)}(x) = (-1)^l exp(x^2)\, d^l/dx^l\{exp(-x^2)\} = 2^{l/2} H_l(\sqrt{2}x)$, where H_l is the Hermite polynomial of degree l.

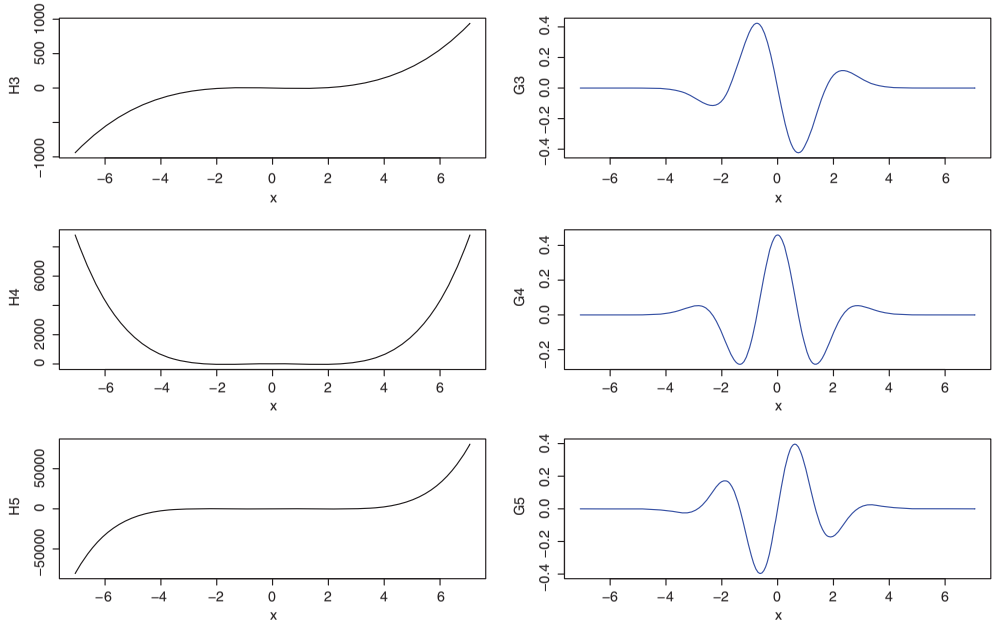

Figure 1.2 Rescaled Hermite polynomials $H_l^{(re)}(x)$ of degree l for $l = 3, 4, 5$ and the corresponding Hermite functions (right) $G_l(x)$. These functions are related via the relation $G_l(x) = (2^l l! \sqrt{\pi})^{-1/2} e^{-x^2/2} H_l^{(re)}(x)$, where $H_l^{(re)}(x) = (-1)^l exp(x^2)\, d^l/dx^l\{exp(-x^2)\} = 2^{l/2} H_l(\sqrt{2}x)$, where H_l is the Hermite polynomial of degree l.

where $M_n \to \infty$ and $M_n = o(n)$ as $n \to \infty$ and the coefficients $\widehat{d}_{l,n}$ are estimators based on the sample means of Hermite functions

$$\widehat{d}_l = \frac{1}{n} \sum_{j=1}^{n} G_l(X_j). \tag{1.20}$$

Under some conditions on the rth derivative ($r \geq 2$) of $f(x)\phi(x)$, Schwartz (1967) derives asymptotic properties of his estimator including the rate of convergence to zero of the mean integrated squared error (MISE).

There are various textbooks and review papers that give excellent overvews of nonparametric density estimation techniques. Basic developments and related information can, for instance, be found in Watson and Leadbetter (1964a, 1964b), Shapiro (1969), Rosenblatt (1971), Bickel and Rosenblatt (1973), Rice and Rosenblatt (1976), Tapia and Thompson (1978), Wegman (1982), Silverman (1986), Hart (1990), Jones and Sheather (1991), Müller and Wang (1994), Devroye (1987), Müller (1997), Loader (1999), and Heidenreich et al. (2013). Various textbooks have addressed applied aspects and included various theoretical results on general kernel smoothing methods. Some examples are Bowman and Azzalini (1997), Wand and Jones (1995), Simonoff (1996), Scott (1992), Thompson and Tapia (1987), and others.

In this chapter, we focus on a selection of ideas for density estimation with independently and identically distributed (iid) observations, restricting ourselves to continuous random variables. We start with the univariate case and the multivariate case is mentioned in the sequel.

Let $X, X_1, X_2, \ldots, X_n$ be iid real-valued univariate continuous random variables with an absolutely continuous cumulative distribution function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du, x \in \mathbb{R} \tag{1.21}$$

where $f(x)$ denotes the probability density function (pdf). The pdf f will be assumed to be a three times continuously differentiable function with finite derivatives. Further conditions will be added in the sequel.

The problem is a nonparametric estimation of $f(x), x \in \mathbb{R}$, using $X_1, X_2, \ldots, X_n$.

1.2 Histograms

The most widely used nonparametric density estimation method is the histogram, especially for univariate random variables. The idea has a long history and the name "histogram" seems to have been used for the first time by Karl Pearson 1895). Basic information on the use of the histogram as a graphical tool to display frequency distributions can be found in any elementary statistical textbook.

Construction of a histogram proceeds as follows. We consider the univariate case. Let

$$\mathcal{A} = \{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_j, \dots\} \tag{1.22}$$

be a partition of the real line into disjoint intervals $\mathbb{A}_j$, the jth interval having the width b_j. Let f_j be the frequency or the number of observations falling in $\mathbb{A}_j$ such that

$$\sum_j f_j = n \tag{1.23}$$

Then the histogram estimate of $f(x)$ at $x \in \mathbb{A}_j$ is given by

$$\widehat{f}_{hist}(x) = \frac{f_j}{nb_j}, x \in \mathbb{A}_j. \tag{1.24}$$

In practice, one starts with k *bins* (or *cells*), $\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_k$, where k is an arbitrary positive integer. In the case of univariate data, the bins are non-overlapping intervals on the real line. If, for instance, the bin widths are equal, blocks with heights proportional to f_j placed on $\mathbb{A}_j, j = 1, , \dots, k$, produce a histogram. The block heights may be scaled so that the sum of all block sizes is equal to 1. Typically, b and k will depend on the sample size n. Thus let $k = k_n$ and $b = b_n$ be sequences such that, with increasing sample size, i.e., as $n \to \infty$,

$$b \to 0, k \to \infty, kb \to \infty, nb \to \infty. \tag{1.25}$$

For instance, the bins may be defined as

$$\mathbb{A}_j = (t_j - b/2, t_j + b/2] \tag{1.26}$$

$$\text{where } t_j = (j - 1/2)b, j = 1, 2, \dots, k \tag{1.27}$$

and $b > 0$ is the *bin width*. Thus

$$t_1 = b/2, t_2 = 3b/2, t_3 = 5b/2 \dots, t_k = (2k - 1)b/2 \tag{1.28}$$

so that the bins are

$$\mathbb{A}_1 = (0, b], \mathbb{A}_2 = (b, 2b], \ldots \mathbb{A}_k = ((k-1)b, kb]. \quad (1.29)$$

This will produce a histogram whose left end-point is set at zero. More generally, let the starting point be at t_0 and consider $j \in \mathbb{Z}$, where $\mathbb{Z}$ denotes the set of all integers. Then, using the bin width b as above, the jth bin $\mathbb{A}_j$ can be defined as the interval $(t_0 + jb, t_0 + (j+1)b]$.

The frequency for $\mathbb{A}_j$ is

$$f_j = \#\{i | X_i \in \mathbb{A}_j, i = 1, 2, \ldots, n\} \quad (1.30)$$

and the histogram estimator of $f(x)$ for a fixed $x \in \mathbb{R}$ is given by

$$\widehat{f}_{hist}(x) = \frac{f_j}{nb}, x \in \mathbb{A}_j. \quad (1.31)$$

Figure 1.3 is a histogram of duration of eruptions in minutes (number of observations $n = 272$) for the Old Faithful geyser in the Yellowstone National Park, Wyoming, USA (source: R). Four different bin widths ($b = 0.01, b = 0.1, b = 0.5$, $b = 1$) are used for illustration. A parametric formulation of the distribution of the data could, for instance, involve a mixture of two normals. We make a note that the discussion on density estimation in this section is focused on iid data. However, the Old Faithful data may in fact be treated as time series observations, so that the expressions for the asymptotic variance calculations do not apply to these data and would need to be modified by incorporating serial correlations. See Azzalini and Bowman (1990) for an interesting analysis of the physical processes behind these data. Some summary statistics for this data set are given below using the summary( ) function in R.

```
> summary(faithful[,1])
   Min. 1st Qu. Median  Mean 3rd Qu.  Max.
  1.600   2.163  4.000 3.488   4.454 5.100
```

1.2.1 Properties of the histogram

Consider the case of equal bin width s. The specific choice of the bin width b directly affects the resulting shape and properties of the histogram. Large b will result in a larger bias whereas a smaller b will increase the variance. How this happens can be

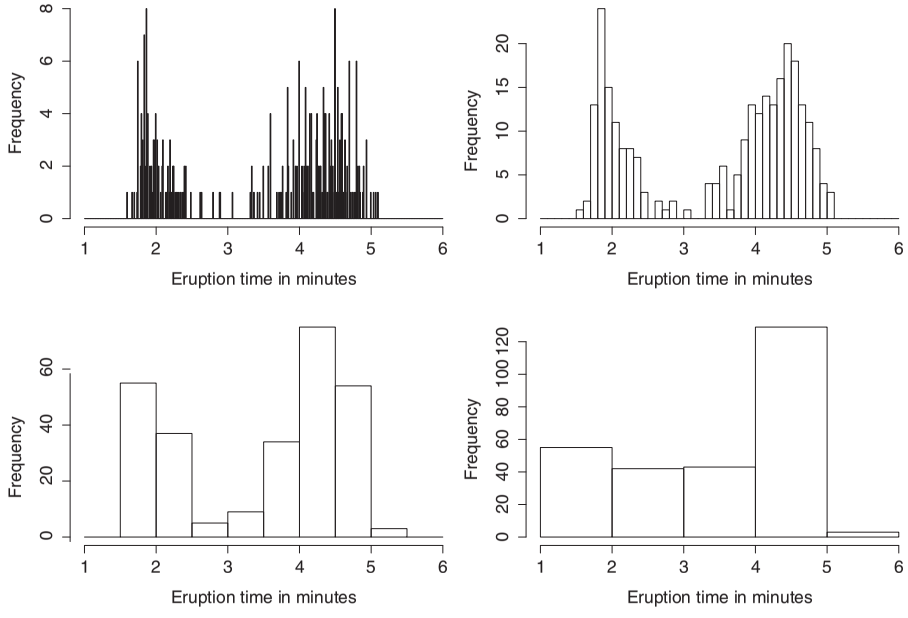

Figure 1.3 Histograms using four different bin widths for the duration (minutes) of eruptions ($n = 272$) for the Old Faithful geyser in the Yellowstone National Park, Wyoming, USA (*Source:* Data from Old Faithful Geyser Data in R). The plots indicate a bimodal distribution, a well-known feature of this data set.

seen from an asymptotic analysis of the histogram. First of all, for each j, the frequency f_j is a sum of zero–one random variables of the type

$$f_j = \sum_{i=1}^{n} I(X_i, \mathbb{A}_j) \tag{1.32}$$

where

$$\begin{aligned} I(X_i, \mathbb{A}_j) &= 1, \text{if } X_i \in \mathbb{A}_j \\ &= 0 \text{ if } X_i \notin \mathbb{A}_j. \end{aligned} \tag{1.33}$$

In particular, the histogram estimator of $f(x), x \in \mathbb{A}_j$, is a sample mean

$$\widehat{f}_{hist}(x) = \frac{1}{nb} \sum_{i=1}^{n} I(X_i, \mathbb{A}_j), x \in \mathbb{A}_j. \tag{1.34}$$

The bias and the variance of the histogram estimator $\widehat{f}_{hist}(x)$ can be derived by noting that due to the iid assumption, the bin frequencies f_j are binomial random variables, i.e.,

$$f_j \sim Binomial(n, p_j) \tag{1.35}$$

where

$$p_j = P(X_i \in \mathbb{A}_j) = \int_{\mathbb{A}_j} f(u)\, du. \tag{1.36}$$

This means

$$\mathbb{E}(\widehat{f}_{hist}(x)) = np_j/(nb) = p_j/b. \tag{1.37}$$

Due to the mean value theorem,

$$p_j = bf(\zeta_j) \tag{1.38}$$

for some $\zeta_j \in \mathbb{A}_j$ so that

$$\mathbb{E}(\widehat{f}_{hist}(x)) = f(\zeta_j). \tag{1.39}$$

Asymptotic unbiasedness is obvious, for instance, if we assume Lipschitz continuity, i.e., suppose that there exists a constant $\delta_j > 0$ such that for all $\zeta_1, \zeta_2 \in \mathbb{A}_j$,

$$|f(\zeta_1) - f(\zeta_2)| < \delta_j |\zeta_1 - \zeta_2|. \tag{1.40}$$

Since $x, \zeta_j \in \mathbb{A}_j$ and the width of the bin $\mathbb{A}_j$ is b, the absolute bias in $\widehat{f}_{hist}(x))$ for $x \in \mathbb{A}_j$ is

$$|\mathbb{E}(\widehat{f}_{hist}(x)) - f(x)| = |f(\zeta_j) - f(x)| < \delta_j |\zeta_j - x| \leq \delta_j b. \quad (1.41)$$

Due to the previous assumption on the bin width, as $n \to \infty$, $b \to 0$, so that

$$|\mathbb{E}(\widehat{f}_{hist}(x)) - f(x)| \to 0 \quad (1.42)$$

i.e., $\mathbb{E}(\widehat{f}_{hist}(x))$ is asymptotically unbiased.

As for the variance, again, since f_j is a binomial random variable,

$$\mathbb{V}ar\,(\widehat{f}_{hist}(x)) = np_j(1-p_j)/(n^2b^2) = p_j(1-p_j)/(nb^2). \quad (1.43)$$

Applying the mean value theorem,

$$\begin{aligned} \mathbb{V}ar\,(\widehat{f}_{hist}(x)) &= nbf(\zeta_j)(1 - bf(\zeta_j))/(n^2b^2) \\ &= f(\zeta_j)(1 - bf(\zeta_j))/(nb) = f(\zeta_j)/(nb) - f^2(\zeta_j)/n. \end{aligned} \quad (1.44)$$

Since as $n \to \infty$, $nb \to \infty$, $\mathbb{V}ar\,(\widehat{f}_{hist}(x))$ converges to zero asymptotically. This result, together with the asymptotic unbiasedness of the histogram, proves pointwise weak consistency of $\widehat{f}_{hist}(x)$ at x.

An asymptotic expression for an upper bound for the mean squared error (mse) can now be given. If we take the leading term of the variance of the histogram, combining it with its bias, we can write down an upper bound for the asymptotic mean squared error (AMSE). Thus, for $x \in \mathbb{A}_j$,

$$AMSE(\widehat{f}_{hist}(x)) \leq f(\zeta_j)/(nb) + \delta_j^2 b^2. \quad (1.45)$$

Differentiating the expression on the right-hand side and equating to zero we get that

$$b_{opt} = \left\{ f(\zeta_j) / \left(2n\delta_j^2 \right) \right\}^{1/3} \quad (1.46)$$

minimizes the above upper bound, i.e., the optimal bandwidth converges to zero at the rate $O(n^{-1/3})$. Substituting this value in the upper bound for the AMSE, we get

$$f(\zeta_j)/(nb_{opt}) + \delta^2 b_{opt}^2 \propto n^{-2/3}. \quad (1.47)$$

In other words, the upper bound at $b = b_{opt}$ of the AMSE converges to zero at the rate $O(n^{-2/3})$. As we shall see later in this chapter, a better convergence rate, namely $O(n^{-4/5})$, is possible to achieve, for instance using an appropriate kernel. For further detailed results on the histogram estimator, see Freedman and Diaconis (1981) and Scott (1979, 1992).

To obtain an approximation to the bias, not just an upper bound, let j be fixed and consider $x, u \in \mathbb{A}_j$, where $\mathbb{A}_j = (t_0 + jb, t_0 + (j+1)b]$, for some point t_0. We derive the asymptotic expressions for the bias and the variance of $\widehat{f}_{hist(x)}$. First of all, since both x and u are in $\mathbb{A}_j$,

$$|u - x| \leq b \tag{1.48}$$

where b is the width of $\mathbb{A}_j$. By Taylor series expansion,

$$f(u) = f(x) + (u - x)f^{(1)}(x) + O(b^2), \tag{1.49}$$

so that

$$\begin{aligned} p_j &= \int_{t_0+jb}^{t_0+(j+1)b} f(u)\, du \\ &= bf(x) + f^{(1)}(x)\{b^2(1+2j) - 2b(x - t_0)\}/2 + O(b^3). \end{aligned} \tag{1.50}$$

This implies that the bias is equal to

$$\begin{aligned} \mathbb{E}(\widehat{f}_{hist}(x)) - f(x) &= p_j/b - f(x) \\ &= f^{(1)}(x)\{b(1+2j)/2 - (x - t_0)\} + O(b^2). \end{aligned} \tag{1.51}$$

In the above expression for the bias, the coefficient of $f^{(1)}(x)$ is the distance between x and the mid-point $\zeta_j^{(mid)}$ of the interval $\mathbb{A}_j$.

Since $|x - t_0| \leq b$ and $b \to 0$ as $n \to \infty$ and j is fixed, the bias of $\widehat{f}_{hist}(x)$ for $x \in (t_0 + jb, t_0 + (j+1)b]$ converges to zero with increasing sample size at the rate $O(b)$ unless $b(1+2j)/2 - (x - t_0) = 0$, or equivalently unless $x = \zeta_j^{(mid)} = (t_0 + jb) + b/2$. In the case $x = \zeta_j^{(mid)}$, the bias of $\widehat{f}_{hist}(x)$ converges to zero at the rate $O(b^2)$.

Similarly, an asymptotic expression for the variance of the histogram estimator $\widehat{f}_{hist}(x)$ can be found. First of all, as observed

earlier in this section,

$$\mathbb{V}ar\{\widehat{f}_{hist}(x)\} = p_j(1-p_j)/(nb^2), \text{where } x \in \mathbb{A}_j. \tag{1.52}$$

Due to the mean value theorem, we may write $p_j = bf(\zeta_j)$ where $\zeta_j \in \mathbb{A}_j$, so that

$$\mathbb{V}ar\{\widehat{f}_{hist}(x)\} = bf(\zeta_j)\{1-bf(\zeta_j)\}/(nb^2). \tag{1.53}$$

Since $x, \zeta_j \in \mathbb{A}_j$, $|x-\zeta_j| \leq b \to 0$ as $n \to \infty$, by Taylor series expansion

$$f(\zeta_j) = f(x) + (\zeta_j - x)f^{(1)}(x) + O(b^2) \tag{1.54}$$

so that substitution yields

$$\mathbb{V}ar\{\widehat{f}_{hist}(x)\} = f(x)/(nb) + O(1/n). \tag{1.55}$$

In other words, the variance of the histogram $\widehat{f}_{hist}(x)$ converges to zero with increasing sample size at the rate $O(1/(nb))$.

There is a simple remedy for avoiding the bias problem in histograms. This is considered in *frequency polygons*.

1.2.2 Frequency polygons

The histogram estimators at the centers of the bins can be used to construct *frequency polygons*, which we denote by $\widehat{f}_{poly}$. In this method, frequency data are displayed by joining the $\widehat{f}_{hist}(\zeta_j^{(mid)})$ values using straight lines, where $\zeta_j^{(mid)}$ is the mid-point of bin j. Thus if $x = \zeta_j^{(mid)}$ then

$$\widehat{f}_{poly}\left(\zeta_j^{(mid)}\right) = \widehat{f}_{hist}\left(\zeta_j^{(mid)}\right). \tag{1.56}$$

For all other x's that are not the mid-points of the various bins, say, $\zeta_j^{(mid)} \leq x \leq \zeta_{j+1}^{(mid)}$, $\widehat{f}_{poly}(x)$ is obtained by joining $(\zeta_j^{(mid)}, \widehat{f}_{hist}(\zeta_j^{(mid)}))$ and $(\zeta_{j+1}^{(mid)}, \widehat{f}_{hist}(\zeta_{j+1}^{(mid)}))$ by a straight line. This means, if $\widehat{f}_{hist}(\zeta_j^{(mid)}) \geq \widehat{f}_{hist}(\zeta_{j+1}^{(mid)})$ and $\zeta_j^{(mid)} \leq x \leq \zeta_{j+1}^{(mid)}$ then

$$\mathbb{E}(\widehat{f}_{poly}(x) - f(x)) \leq \mathbb{E}\left(\widehat{f}_{hist}(\zeta_j^{(mid)}) - f(x)\right) = O(b^2) \tag{1.57}$$

Similarly, if $\widehat{f}_{hist}(\zeta_j^{(mid)}) \leq \widehat{f}_{hist}(\zeta_{j+1}^{(mid)})$ and $\zeta_j^{(mid)} \leq x \leq \zeta_{j+1}^{(mid)}$ then

$$\mathbb{E}(\widehat{f}_{poly}(x) - f(x)) \leq \mathbb{E}\left(\widehat{f}_{hist}(\zeta_{j+1}^{(mid)}) - f(x)\right) = O(b^2). \tag{1.58}$$

Therefore, the bias of $\widehat{f}_{poly}(x)$ is of the order $O(b^2)$.

1.2.3 Histogram bin widths

There are several propositions for selecting the widths of the histogram.

1.2.3.1 Sturges' rule

As Scott (1992) explains, Sturges' rule (Sturges 1926) is a rule for number of bins with constant bin width. The idea is based on the convergence of the binomial distribution to the normal. Consider the ideal case of a histogram for an appropriatey scaled normal data, where n values fall into k bins centered at $0, 1, \ldots, k-1$ according to the formula $f_j = \binom{k-1}{j}$, f_j being the frequency for the bin centered at j. This leads to

$$n = \sum_{j=0}^{k-1} f_j = 2^{k-1}. \tag{1.59}$$

Taking the logarithm, one gets Sturges' rule, namely,

$$k \approx 1 + log_2(n). \tag{1.60}$$

In particular, as n becomes large, the relative frequency histogram (or the binomial distribution with parameters $k-1$ and $p = 1/2$) converges to a normal pdf with mean $(k-1)/2$ and variance $(k-1)/4$. When the data are not normal, Doane (1976) proposes to extend Sturges' rule by including the standardized skewness coefficient $\sqrt{b_1}/s.d.(\sqrt{b_1})$, where

$$\sqrt{b_1} = \sum_{j=1}^{n} (X_j - \bar{X})^3 / \left(\sum_{i=j}^{n} (X_j - \bar{X})^2\right)^{3/2} \tag{1.61}$$

is the sample skewness, for normal data its approximate variance being, as $n \to \infty$ (Pearson 1936),

$$\mathbb{V}ar\,(\sqrt{b_1}) = 6/n + o(1/n), \tag{1.62}$$

so that for skewed data, a larger number of bins are obtained. Doane's proposed formula for the extra number of bins is

$$k_{extra} \approx log_2\left(1 + \sqrt{nb_1/6}\right) \tag{1.63}$$

which converges to zero when the skewness coefficient approaches zero. See Doane (1976) for details. The ideal bin width may then be taken as $b_{opt} = (X_{(n)} - X_{(1)})/k$, where $X_{(j)}$ is the jth order statistic so that $X_{(n)} - X_{(1)}$ is simply the range of the observations.

1.2.3.2 Other rules: integrated squared density derivatives

The approaches that we mention here are concerned with estimation of integrals of squares of derivatives of the unknown pdf f. Consider the problem of finding the optimum constant bin width that minimizes the leading term of the asymptotic integrated mean squared error. First of all, assume that $\int_{-\infty}^{\infty}(f^{(1)}(x))^2dx$ is finite and positive, and consider bins of constant width b. To determine b, consider for simplicity the bin $(0, b]$ and let $x \in (0, b]$.

Then the asymptotic expression for the variance of the histogram estimator is

$$\mathbb{V}ar\{\widehat{f}_{hist}(x)\} = f(x)/(nb) + O(1/n), \tag{1.64}$$

so that integrating out $x \in \mathbb{R}$,

$$\text{Total integrated variance} = 1/(nb) + O(1/n). \tag{1.65}$$

Similarly, the asymptotic expression for the bias is (substitute $j = 0, t_0 = 0$)

$$\begin{aligned}\mathbb{E}(\widehat{f}_{hist}(x)) - f(x) &= f^{(1)}(x)\{b(1 + 2j)/2 - (x - t_0)\} + O(b^2)\\ &= f^{(1)}(x)\{b/2 - x\} + O(b^2).\end{aligned} \tag{1.66}$$

This gives the leading term of the mean integrated squared bias as

$$\int_0^b (b/2 - x)^2\{f^{(1)}(x)\}^2dx = b^3/12\{f^{(1)}(\zeta)\}^2 \tag{1.67}$$

for some $\zeta \in (0, b]$. Doing smilar calculations for all (infinitely many) bins, we get (see Scott 1979, 1992 and Freedman and Diaconis 1981)

$$\text{Total integrated squared bias} = b^2/12 \sum_{j, \zeta_j \in \mathbb{A}_j} \{f^{(1)}(\zeta_j)\}^2 b$$
$$\approx b^2/12 \int_{-\infty}^{\infty} \{f^{(1)}(x)\}^2 dx \quad (1.68)$$

Combining, the leading term in the integrated mean squared error is

$$\text{AMISE} = 1/(nb) + b^2/12 \int_{-\infty}^{\infty} \{f^{(1)}(x)\}^2 dx. \quad (1.69)$$

Differentiating with respect to b and equating to zero, one gets the rule (see Scott 1979)

$$b_{opt} = n^{-1/3} \left(\frac{6}{R(f^{(1)})} \right)^{1/3}. \quad (1.70)$$

As can be seen above, for real data analysis, the main problem in implementing the asymptotic formula for b_{opt} is the presence of the unknown quantity $R(f^{(1)})$. This is an interesting general problem, namely estimation of the integrated squared density derivative $R(f^{(p)}), p \geq 0$, and this has been addressed by various authors. Some of the main ideas involve a direct plug-in approach or the use of cross-validation. Some relevant references are Kronmal and Tarter (1968), Woodroofe (1970), Rudemo (1982), Bowman (1984), Stone (1984), Silverman (1986), Hall and Marron (1987), Bickel and Ritov (1988), Jones and Sheather (1991), and Scott and Terrell (1987), as well as Wand (1997) and Jones et al. (1996). For histograms, we briefly describe the *reference to a standard distribution* approach due to Scott (1979, 1992) and the *method of finite differences* due to Scott and Terrell (1987). Kernel based methods are taken up in the context of kernel estimation of the density f.

Reference to a known distribution

This rule is due to Scott (1979) with further suggestions by Freedman and Diaconis (1981). In this method, one uses a known parametric family to estimate the quantity $R(f^{(1)})$ in the

formula for the optimum bin width. In particular, the zero mean normal distribution $N(0, \sigma^2)$ is often used as the reference distribution due to the importance of scale in the bin width selection problem (see Silverman 1986 and Jones et al. 1996). The idea is to consider f that is close to the normal pdf with zero mean and variance σ^2, which is also the population variance of X_i.

Using the idea of *Gaussian reference* due to Tukey (1977, p. 623) (also see Deheuvels 1977), Scott (1979) proposes to replace the quantity $(6/R(f^{(1)}))^{1/3}$ by $3.49s$, where s is an estimate of the standard deviation and $R(g)$ denotes the integral of the square of a function g. Using the $N(0, \sigma^2)$ pdf, the integrated squared first derivative becomes

$$\begin{aligned} R(f^{(1)}) &= \int_{-\infty}^{\infty} (f^{(1)}(x))^2 dx \\ &= \int_{-\infty}^{\infty} \left(\frac{-x}{\sigma^2}\right)^2 \cdot \left(\frac{1}{\sqrt{2\pi}\sigma} exp\{-x^2/(2\sigma^2)\}\right)^2 dx \\ &= 1/(4\sigma^3\sqrt{\pi}). \end{aligned} \tag{1.71}$$

Substituting this formula in b_{opt}, we get

$$b_{opt} = (24\sqrt{\pi})^{1/3}\sigma n^{-1/3}, \tag{1.72}$$

and using an estimator $\hat{\sigma}$ for σ, the data-driven formula

$$b_{opt} = 3.5\hat{\sigma}n^{-1/3} \tag{1.73}$$

is obtained. As for the choice of $\hat{\sigma}$, one option is to use the sample standard devition $s = \{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\}^{1/2}$. Another suggestion has been to use (see Silverman 1986)

$$IQR_{sample}/IQR_{\phi} = IQR_{sample}/1.349, \tag{1.74}$$

where IQR_{sample} is the sample interquartile range and IQR_{ϕ} is the interquartile range of the standard normal distribution, or rather

$$\hat{\sigma} = min(s, IQR_{sample}/1.349), \tag{1.75}$$

which seems to work well with unimodal as well as moderately bimodal densities (see Silverman 1986, p. 47).

Method using finite differences

In this method (see Scott and Terrell 1987 and Scott 1992, p. 75), the quantity $R(f^{(1)})$ is estimated by using finite differences of the histogram. The estimator is

$$\widehat{R(f^{(1)})} = b \sum_j d_j^2 - 2/(nb^3) \tag{1.76}$$

where $d_j = (f_{j+1}/(nb) - f_j/(nb))/b$, $b \to 0$ and $k \to \infty$ as $n \to \infty$, f_j being the frequency for bin j.

1.2.4 Average shifted histogram

An idea related to the histograms is the so-called *ASH*. See Scott (1985, 1992) for greater details. The average shifted histogram (ASH) is constructed by averaging over several choices of the starting point or the bin origin. For increasing number of choices of the starting point, the idea is to asymptotically reduce the effect of the starting point used to define the bins. Thus, suppose that the lth histogram estimator is constructed using the bins

$$\mathbb{A}_{j,l} = (t_l + jb, t_l + (j+1)b]. \tag{1.77}$$

The corresponding histogram estimator is then

$$\widehat{f}_{hist,l}(x) = \frac{1}{nb} \sum_{i=1}^{n} I(X_i, \mathbb{A}_{jl}), x \in \mathbb{A}_{j,l} \tag{1.78}$$

where $x \in \mathbb{A}_{j,l}$ and $I(X_i, \mathbb{A}_{jl}) = 1$ if $X_i \in \mathbb{A}_{j,l}$ and $I(X_i, \mathbb{A}_{jl}) = 0$ otherwise. At the next step, considering L different values for the starting point t_l, $l = 1, 2, \ldots, L$, one has the ASH estimator of $f(x)$ as

$$\widehat{f}_{ASH}(x) = \frac{1}{L} \sum_{l=1}^{L} \widehat{f}_{hist,l}(x) = \frac{1}{nbL} \sum_{l=1}^{L} \sum_{i=1}^{n} I(X_i, \mathbb{A}_{jl}) I(x, \mathbb{A}_{jl}) \tag{1.79}$$

1.3 Kernel density estimation

While the histogram or the Naive estimator gives a nonparametric estimator of the probability density function, a smoother density estimator may be desirable. For the Old Faithful data

example, a parametric formulation of the distribution of the data could, for instance, involve a mixture of well-chosen (known) density functions, such as the normal for this particular data set. See Johnson and Kotz (1994) for a detailed account of parametric families of distributions. Here we focus on nonparametric curve estimation methods and in particular kernel density estimation, a method that leads to smoother curves than the histogram or the Naive estimator, but of course, depending on the choice of the kernel. The uniform kernel, for instance, will give rise to less smooth estimators (Naive estimator), as opposed to another kernel such as the Gaussian.

Walter and Blum (1979) note that many density estimators may be expressed as a generalized kernel estimator. Terrell and Scott (1992) state that a density estimator that is continuous and Gateaux differentiable on the empirical distribution function (edf) may be written as a generalized kernel estimator, using a generalized kernel K_G; also see Scott (1992) for further explanations. Thus if $\widehat{f}$ is a density estimator, then

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_G(X_i, x, F_n). \tag{1.80}$$

Other than histogram type estimators, the basic idea of kernel density estimation seems to have arisen in the context of smoothing periodograms (Einstein 1914, Yaglom 1987, and Daniell 1946), and from the Naive estimator due to Fix and Hodges (1951, unpublished; see Silverman and Jones (1989) for a reprinted version). Rosenblatt (1956) discusses consisteny of the Naive estimator as well as kernel density estimation with a general kernel, whereas Parzen (1962) provides further insights, and the estimator is often referred to as the Parzen–Rosenblatt density estimator; also see Akaike (1954), Whittle (1958), Bartlett (1963), and Farrell (1967), among others.

A related problem is spectral density estimation, for which one may refer to standard time series books, such as Priestley (1989); also see Beran et al. (2013) for a review of recent developments, in particular for time series with hyperbolically decaying correlations. For early research on this topic, see in particular Bartlett and Medhi (1955), Whittle (1957), and Parzen (1957, 1961, 1962). Interesting historical notes are, for instance, in Brillinger (1993).

1.3.1 Naive density estimator

Another approach to displaying relative frequencies while avoiding the problem of choosing a starting point for the bins is the method of *difference quotient* using the empirical distribution function (edf). This estimator due to Fix and Hodges (1951), further discussed by Rosenblatt (1956), is also called the *Naive estimator*. First of all, considering the density $f(x)$ as the derivative of the cumulative distribution function $F(x) = P(X \leq x)$,

$$f(x) = \lim_{b \to 0} \frac{1}{2b} P\{x - b < X \leq x + b\} = \lim_{b \to 0} \frac{1}{2b} \{F(x + b) - F(x - b)\}. \tag{1.81}$$

Then a natural estimator for $f(x)$ is

$$\widehat{f}_{naive}(x) = \frac{1}{2b} \{F_n(x + b) - F_n(x - b)\} \tag{1.82}$$

where F_n is the edf, i.e.,

$$F_n(x) = \#\{i | X_i \leq x\}/n. \tag{1.83}$$

In terms of relative frequencies in the interval $(x - b, x + b]$,

$$\widehat{f}_{naive}(x) = \frac{1}{2nb} \sum_{i=1}^{n} I(X_i, (x - b, x + b]) \tag{1.84}$$

where $(x - b, x + b]$ is an interval of length $2b$ and $b > 0$ is a bandwidth that converges to zero with increasing sample size. Specifically, as $n \to \infty$, $b \to 0$ and we also let $nb \to \infty$. The indicator function $I(X_i, (x - b, x + b])$ equals 1 if $\{x - b < X_i \leq x + b\}$ and assumes the value 0 otherwise. Asymptotic properties of the Naive estimator can be derived by noting the fact that since $X_1, X_2, \ldots, X_n$ are iid, $\sum_{i=1}^{n} I(X_i, (x - b, x + b])$ is a binomial random variables with mean

$$\mathbb{E} \left\{ \sum_{i=1}^{n} I(X_i, (x - b, x + b]) \right\} = n(F(x + b) - F(x - b)) \tag{1.85}$$

and variance

$$\begin{aligned} &\mathbb{V}ar \left\{ \sum_{i=1}^{n} I(X_i, (x - b, x + b]) \right\} \\ &\quad = n(F(x + b) - F(x - b))(1 - F(x + b) + F(x - b)). \end{aligned} \tag{1.86}$$

Since $b \to 0$ with increasing sample size, using Taylor series expansion,

$$F(x+b) - F(x-b) = 2bf(x) + O(b^3). \tag{1.87}$$

Substitution yields

$$\mathbb{E}\{\widehat{f}_{naive}(x)\} = \{2bf(x) + O(b^3)\}/(2b) = f(x) + O(b^2) \tag{1.88}$$

and

$$\begin{aligned}\mathbb{V}ar\{\widehat{f}_{naive}(x)\} &= \{2bf(x) + O(b^3)\}\{1 - 2bf(x) + O(b^3)\}/(4nb^2) \\ &= f(x)/(2nb) + o(1/(nb)).\end{aligned} \tag{1.89}$$

The Naive estimator can also be written up as a kernel estimator by taking the uniform (or the rectangular) kernel: $K_{uniform}(u) = 1/2, -1 < u \leq 1$ and $K_{uniform}(u) = 0$ otherwise. Then

$$\widehat{f}_{naive}(x) = \frac{1}{nb}\sum_{i=1}^{n} K_{uniform}\left(\frac{X_i - x}{b}\right). \tag{1.90}$$

Note that

$$\int_{-1}^{1} K^2_{uniform}(u)\, du = 1/2 \tag{1.91}$$

so that the variance of the Naive estimator is

$$\mathbb{V}ar\{\widehat{f}_{naive}(x)\} = f(x)R(K_{uniform})/(nb) + o(1/(nb)) \tag{1.92}$$

where, for a square integrable function g, we use the notation

$$R(g) = \int g^2(u)du. \tag{1.93}$$

Thus the bias of the Naive estimator converges to zero at the rate $O(b^2)$ whereas its variance converges to zero at the rate $O(1/(nb))$. These are also the rates that are typical for kernel density estimators with iid data. Aymptotic expression for the mean squared error can now be minimized with respect to b, to obtain a formula for the optimal bandwidth. This optimal bandwidth can be shown to converge to zero at the rate $O(n^{-1/5})$ so that the best rate for the asymptotic mean squared error for $\widehat{f}_{naive}(x)$ will then be $O(n^{-4/5})$. This is a clear improvement over the histogram estimator. However, the Naive estimator $\widehat{f}_{naive}(x)$ is not continuous; it has jumps at the points $x = X_i \pm b$ and zero derivatives at

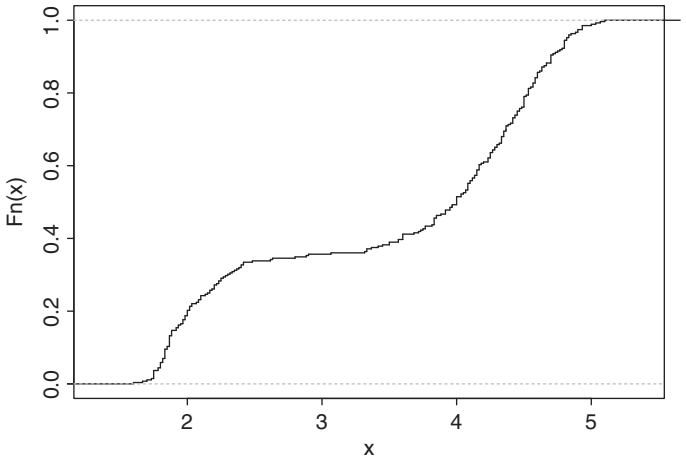

Figure 1.4 The Old Faithful duration of eruptions data from the Old Faithful geyser in the Yellowstone National Park, Wyoming, USA (*Source:* Data from Old Faithful Geyser Data in R). The R-function *ecdf* is used to produce the plot.

all $x \in (X_i - b, X_i + b)$. In particular, with an appropriate choice of the kernel, smoother density estimators can be achieved. This will be taken up in the more general context of kernel density estimation, where the kernel will belong to a broader class.

Figure 1.4 shows the edf for the Old Faithful eruption data and Figure 1.5 illustrates the Naive density estimator for the same data set using four different choices of the bandwidth, namely, $b = 0.01, b = 0.1, b = 0.5, b = 1$. A larger bandwidth leads to smoother estimators but with higher bias. In particular, the bimodal characteristic is no longer obvious when $b = 1$ is chosen whereas the density estimator is "following" the data (the relative frequencies), so to speak, when the bandwidth is very small ($b = 0.1$ in our example).

Intuitively, a small bandwidth will contain fewer observations, so that, being a sample mean, the variance of the density estimator will rise. In contrast, having a small variance, a large bandwidth will lead to a smoother density estimator, but the resulting estimator will tend to have larger bias as it will miss the local features. There is thus a *trade-off* regarding the choice of the bandwidth, which will be discussed in the sequel.

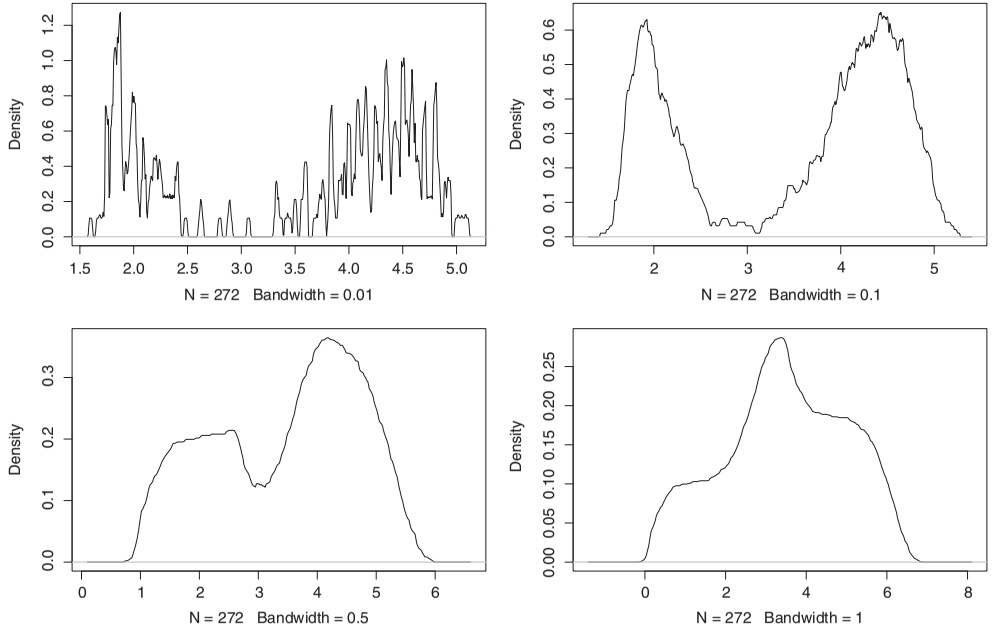

Figure 1.5 Kernel density estimators using the uniform (or the rectangular) kernel for the duration (minutes) of eruptions ($n = 272$) for the Old Faithful geyser in the Yellowstone National Park, Wyoming, USA (*Source:* Data from Old Faithful Geyser Data in R). Four different bandwidths ($b = 0.01, b = 0.1, b = 0.5, b = 1$) are used for illustration.

As regards the choice of the kernel, generally speaking much of its properties, such as continuity, differentiability, etc., will be inherited by the density estimator. This also implies that depending on the situation, some further fine-tuning may be required. One example is density estimation for bounded data, e.g., when the observations are necessarily non-negative. The kernel density estimator as defined above may lead to positive values for the density estimator when the support is below zero. One idea is to use asymmetric kernels. Another option is to recognize that density estimation can be viewed as a regression problem so that the issue of boundary bias may be handled via local polynomials; see Cheng et al. (1997).

1.3.2 Parzen–Rosenblatt kernel density estimator

The Parzen–Rosenblatt kernel density estimator uses more general kernels and is given by

$$\widehat{f}_{PR}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right), \tag{1.94}$$

where K is a kernel and $b = b_n$ is a sequence of bandwidths that converge to zero with increasing value of n, though not too fast. The exact nature of the rate of convergence of b will be specified in the sequel.

To see how the method works, note that the kernel density estimator can be seen as a sum of terms like

$$w_i(x)/n = \frac{1}{nb} K\left(\frac{x - X_i}{b}\right), \tag{1.95}$$

where the symmetric kernel K with bandwidth b has the center of its support at X_i. Figure 1.6 shows the kernel density estimate for a random sample (simple sampling without replacement) of size 25 from the Old Faithful eruption data using the Gaussian kernel $K(u) = (\sqrt{2\pi})^{-1/2} exp(-u^2/2)$ and the bandwidth $b = 0.5761$. The R-function *density* is used along with the *default* bandwidth option. Superimposed on the plot are $w_i(x)/n$ functions centered at 10 randomly chosen X_i values.

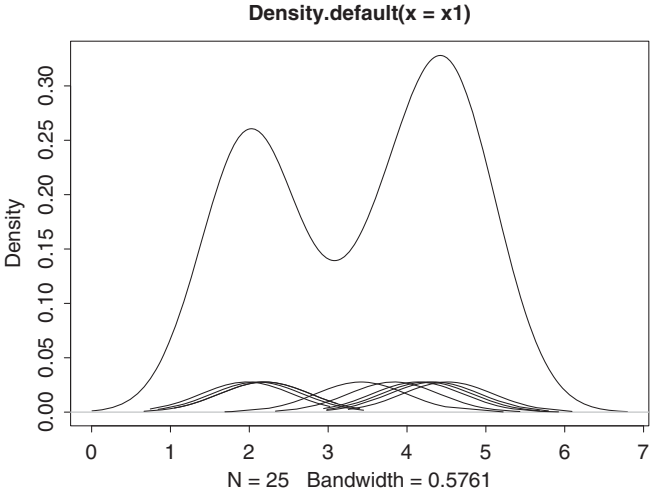

Figure 1.6 Kernel density estimation for a random sample of 25 observations from the Old Faithful eruptions data from the Old Faithful geyser in the Yellowstone National Park, Wyoming, USA (*Source:* Data from Old Faithful Geyser Data in R). The Gaussian kernel is used along with the R-function "density", with the default bandwidth $b = 0.5761$.

Some summary statistics for Figure 1.6 are given below.

```
> x <- faithful[,1]
> set.seed(56699934)
> x1 <- sample(x, size=25)
> summary(x1)
   Min. 1st Qu. Median  Mean 3rd Qu.  Max.
  1.733   2.100  3.733 3.405   4.500 5.067
> density(x1)
Call:
density.default(x = x1)
Data: x1 (25 obs.); Bandwidth 'bw' = 0.5761
      x                y
Min.   :0.0048  Min.   :0.000502
1st Qu.:1.7024  1st Qu.:0.032486
Median :3.4000  Median :0.152728
Mean   :3.4000  Mean   :0.147084
3rd Qu.:5.0976  3rd Qu.:0.239504
Max.   :6.7952  Max.   :0.328021
```

Figure 1.7 displays the Parzen–Rosenblatt kernel density estimator for the Old Faithful eruption data (Source: R) using the Gaussian (normal) kernel. Thus recalling that the pdf of the normal distribution with mean μ and variance $\sigma^2 > 0$, $(N(\mu, \sigma^2))$ evaluated at $x \in \mathbb{R}$ is $\frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\}$, so that, substituting $\mu = X_i$ and $\sigma = b$, the density estimator is

$$\widehat{f}_{PR}(x) = \frac{1}{n}\sum_{i=1}^{n} w_i(x) = \frac{1}{nb}\frac{1}{\sqrt{2\pi}}\sum_{i=1}^{n} exp\left\{-\frac{1}{2}\left(\frac{x - X_i}{b}\right)^2\right\}. \tag{1.96}$$

1.3.2.1 Kernel density estimator as a pdf

Finite sample properties of $\widehat{f}_{PR}(x)$ can be derived easily. In general the regularity properties of K will be inherited by $\widehat{f}_{PR}$; see discussions in Silverman (1986, Chapter 3). First of all, since K is a pdf,

$$\begin{aligned} &\widehat{f}_{PR}(x) \geq 0, \text{for all } x \in \mathbb{R}, \text{and} \\ &\int_{-\infty}^{\infty} \widehat{f}_{PR}(x)\, dx = 1 \end{aligned} \tag{1.97}$$

i.e., a global (constant) bandwidth can ensure that $\widehat{f}_{PR}(x)$ is a pdf as well. The kernel density estimator $\widehat{f}_{PR}(x)$ is a convolution (Shapiro 1969) of the sample (empirical) distribution and the smooth kernel K, i.e., the kernel density estimator is obtained by smoothing $dF_n(x)$ where $F_n(x)$ is the edf, a non-smooth step function. Thus the kernel density estimator is obtained by "by linear smoothing of the observed density" (Whittle 1958). The degree of the smoothness can be controlled by choosing K and b appropriately. As mentioned above, taking K to be a rectangular kernel, one arrives at the Naive estimator.

We take a brief look at some generating functions and moments of $\widehat{f}_{PR}$. First of all, the empirical characteristic function (ecf) is

$$\phi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}, t \in \mathbb{R} \tag{1.98}$$

and let the (population) characteristic function for f and K be

$$\begin{aligned} \phi_f(t) &= \int_{-\infty}^{\infty} e^{itx} f(x) dx, t \in \mathbb{R} \\ \phi_K(t) &= \int_{-\infty}^{\infty} e^{itx} K(x) dx, t \in \mathbb{R} \end{aligned} \tag{1.99}$$

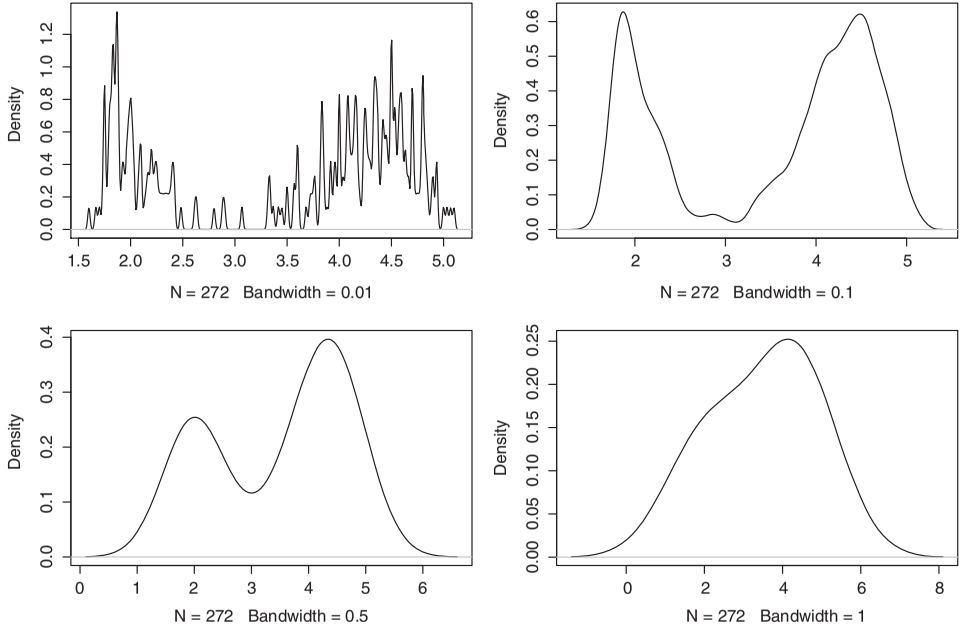

Figure 1.7 Kernel density estimators using the Gaussian kernel for the duration (minutes) of eruptions ($n = 272$) for the Old Faithful geyser in the Yellowstone National Park, Wyoming, USA (*Source:* Data from Old Faithful Geyser Data in R). Four different bandwidths ($b = 0.01, b = 0.1, b = 0.5, b = 1$) are used for illustration.

Then the characteristic function of the pdf $\widehat{f}_{PR}(x)$ is

$$\begin{aligned}\phi_{\widehat{f}_{PR}}(t) &= \int_{-\infty}^{\infty} e^{itx}\widehat{f}_{PR}(x)\,dx \\ &= \frac{1}{nb}\sum_{j=1}^{n}\int_{-\infty}^{\infty} e^{itx}K\left(\frac{X_j - x}{b}\right)dx \\ &= \phi_n(t)\phi_K(-bt) = \phi_n(t)\mathcal{R}e(\phi_K(-bt)) \end{aligned} \tag{1.100}$$

since K is assumed to be symmetric around zero, where

$$\mathcal{R}e(\phi_K(t)) = \int_{-\infty}^{\infty} cos(tx)K(x)dx. \tag{1.101}$$

Taking expectation,

$$\mathbb{E}\{\phi_{\widehat{f}_{PR}}(t)\} = \phi_f(t)\phi_K(-bt) = \phi_f(t)\mathcal{R}e(\phi_K(-bt)). \tag{1.102}$$

In particular, asymptotic properties may be derived. See Csörgő (1981a, 1981b) and Feuerverger and Mureika (1977).

Similarly, if the moment generating function (*Laplace transform*) (mgf) of f and K exist, let these be denoted by

$$\begin{aligned}\mu_f(t) &= \int_{-\infty}^{\infty} e^{tx}f(x)dx \\ \mu_K(t) &= \int_{-\infty}^{\infty} e^{tx}K(x)dx \end{aligned} \tag{1.103}$$

and let the empirical moment generating function (emgf) be

$$\mu_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{tX_j}, \tag{1.104}$$

for $t \in \mathbb{R}$. Then the mgf of $\widehat{f}_{PR}(x)$ can be written as the product

$$\mu_{\widehat{f}_{PR}}(t) = \mu_n(t)\mu_K(-bt). \tag{1.105}$$

Taking expectation,

$$\mathbb{E}\{\mu_{\widehat{f}_{PR}}(t)\} = \mu_f(t)\mu_K(-bt), \tag{1.106}$$

so that, as $n \to \infty$, since $b \to 0$, for each t,

$$\mathbb{E}\{\mu_{\widehat{f}_{PR}}(t)\} \to \mu_f(t), \tag{1.107}$$

since K integrates to 1. In other words, $\mu_{\hat{f}_{PR}}(t)$ is an asymptotically unbiased estimator of $\mu_f(t)$. Being a convolution, the characteristic function and the moment generating function (*Laplace transform*) of $\hat{f}_{PR}(x)$ are respectively products of the empirical characteristic function and the characteristic function for K and the empirical moment generating function and the moment generating function (*Laplace transform*) for K when they exist. Note that when there are no easy expressions for the density function, such as for stable distributions, the characteristic function for $\hat{f}$ may be worth investigating for studying asymptotic properties of $\hat{f}_{PR}$ or to develop further optimality criteria. Moreover, Taylor series-of-fit tests may be developed. The empirical transforms $\mu_n(t)$ and $\phi_n(t)$ have been studies extensively in the literature for Taylor series-of-fit tests. See Csörgő (1981a, 1981b), Feuerverger and McDunnough (1981a, 1981b), Feuerverger and Mureika (1977), Ghosh (1996, 2003), Ghosh and Beran (2006), Ghosh and Ruymgaart (1992), Ghosh and Beran (2000, 2006), Feuerverger and Ghosh (1988), Ghosh (2013), Koutrevelis (1980), Koutrevelis and Meintanis (1999), and others; for related ideas see Cao and Lugosi (2005) for Taylor series-of-fit tests using kernel density estimators.

The moments of the pdf $\hat{f}_{PR}(x)$ can also be derived under appropriate moment conditions on the kernel K. In particular, the mth moment ($m \in \mathbb{N}$) can be given as

$$\int_{-\infty}^{\infty} x^m \hat{f}_{PR}(x)\,dx = \sum_{r=0}^{m} \frac{m!}{r!(m-r)!}(-b)^{m-r}\mu_r^{(1)}\mu_{m-r}^{(K)} \qquad (1.108)$$

where $\mu_r^{(1)} = \sum_{j=1}^{n} X_j^r/n$ is the rth sample (raw) moment and $\mu_r^{(K)} = \int_{-\infty}^{\infty} u^r K(u)\,du$, etc. Thus, $\mu_1^{(1)} = \sum_{j=1}^{n} X_j/n = \bar{X}$, $\mu_2^{(1)} = \sum_{j=1}^{n} X_j^2/n$, $\mu_0^{(K)} = 1$, $\mu_1^{(K)} = 0$, etc. For instance, the first two moments are

$$\begin{aligned}
&\int_{-\infty}^{\infty} x\hat{f}_{PR}(x)\,dx = \bar{X}\\
&\int_{-\infty}^{\infty} x^2\hat{f}_{PR}(x)\,dx = \mu_2^{(1)} + b^2\mu_2^{(K)}\\
&\int_{-\infty}^{\infty} (x-\bar{X})^2\hat{f}_{PR}(x)\,dx = (n-1)s^2/n + b^2\mu_2^{(K)}. \qquad (1.109)
\end{aligned}$$

where $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ is the sample variance.

1.3.2.2 Mean integrated squared error (MISE)

This quantity is

$$\mathbb{E}\left\{\int_{-\infty}^{\infty}(\widehat{f}_{PR}(x)-f(x))^2dx\right\}$$
$$=\mathbb{E}\left\{\int_{-\infty}^{\infty}\widehat{f}^2_{PR}(x)\,dx-2\int_{-\infty}^{\infty}\widehat{f}_{PR}(x)f(x)\,dx\right\}+\int_{-\infty}^{\infty}f^2(x)\,dx. \tag{1.110}$$

Using the facts that (a) the $X_1,\ldots X_n$ are iid random variables and (b) K is a symmetric kernel, the above terms can be written up in terms of convolutions and their integrals.

First of all, writing

$$K_b(u)=\frac{1}{b}K(u/b) \tag{1.111}$$

and convolution of two functions f and g as

$$(f\otimes g)(x)=\int_{-\infty}^{\infty}f(u)g(x-u)\,du \tag{1.112}$$

we have, for the first term in the MISE,

$$\mathbb{E}\left\{\int_{-\infty}^{\infty}\widehat{f}^2_{PR}(x)\,dx\right\}$$
$$=\frac{1}{n^2b^2}\mathbb{E}\left\{\int_{-\infty}^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}K\left(\frac{X_i-x}{b}\right)K\left(\frac{X_j-x}{b}\right)dx\right\}$$
$$=\frac{1}{n^2b^2}\mathbb{E}\left\{\int_{-\infty}^{\infty}\sum_{i=1}^{n}K^2\left(\frac{X_i-x}{b}\right)dx\right\}$$
$$+\frac{1}{n^2b^2}\mathbb{E}\left\{\int_{-\infty}^{\infty}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}K\left(\frac{X_i-x}{b}\right)K\left(\frac{X_j-x}{b}\right)dx\right\}$$
$$=\frac{1}{nb}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}K^2(u)f(x-bu)\,dudx$$
$$+\frac{n-1}{nb}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}K(u)f(x-bu)\,du\right)^2dx$$
$$=\frac{1}{n}\int_{-\infty}^{\infty}\left(K_b^2\otimes f\right)(x)\,dx+\left(1-\frac{1}{n}\right)\int_{-\infty}^{\infty}\left\{(K_b\otimes f)(x)\right\}^2dx. \tag{1.113}$$

Similarly for the second term,

$$\begin{aligned}\mathbb{E}\left\{\int_{-\infty}^{\infty}\widehat{f}_{PR}(x)f(x)\,dx\right\} &= \mathbb{E}\left\{\int_{-\infty}^{\infty}\frac{1}{nb}\sum_{i=1}^{n}K\left(\frac{X_i-x}{b}\right)f(x)\,dx\right\}\\ &= \mathbb{E}\left\{\int_{-\infty}^{\infty}\frac{1}{b}K\left(\frac{X_i-x}{b}\right)f(x)\,dx\right\}\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}K(u)f(x-bu)\,du\right)f(x)\,dx\\ &= \int_{-\infty}^{\infty}(K_b\otimes f)(x)f(x)\,dx. \qquad (1.114)\end{aligned}$$

Note that if K and f belong to the same family of densities that are closed under convolution, such as the Gaussian, then simpler expressions can be derived for specific distributions (also see Deheuvels 1977, Fryer 1976, and Marron and Wand 1992). An interesting case is the family of normal mixtures, i.e., the density f is of the form

$$f(x) = \sum_{r=1}^{m} p_r\phi\left(x|\mu_r,\sigma_r^2\right) \qquad (1.115)$$

where $0 \le p_r \le 1$ and $p_1 + \cdots p_m = 1$ whereas $\phi(x|\mu_r,\sigma_r^2)$ is the pdf of the normal distribution with mean μ_r and variance σ_r^2 (see Marron and Wand 1992 for details).

It turns out that optimum selection of the bandwidth b is of major relevance for estimating f. However, as we shall see next, the asymptotic bias and variance expressions indicate that large values of b tend to increase bias whereas small values of b increase the variance of the estimator. One option to obtain the optimum bandwidth is to minimize the mean squared error (mse) of $\widehat{f}_{PR}(x)$. Derivation of the expressions for bias and variance and hence of the mean squared error requires imposing appropriate conditions on the density f.

1.3.2.3 Asymptotic unbiasedness

To start with we consider the issue of bias. It is well known that nonparametric curve estimates are not unbiased in finite samples. However, asymptotic unbiasedness can be achieved under suitable regularity conditions. Parzen (1962) established asymptotic unbiasedness of the kernel density estimator under

relatively weak conditions. An outline of his proof is given here. Also see Bochner (1955, Theorem 1.1.1). Consider the formula for the density estimator $\widehat{f}_{PR}$ where the kernel

$$K : \mathbb{R} \to \mathbb{R}$$

is a continuous function such that

$$\sup_{u \in \mathbb{R}} |K(u)| < \infty \tag{1.116}$$

$$\lim_{u \to \mathbb{R}} |uK(u)| = 0 \tag{1.117}$$

$$\int_{u \in \mathbb{R}} |K(u)| du = 1 \tag{1.118}$$

whereas the bandwidth b converges to zero. Also recall that, $X_1, X_2, \ldots, X_n \sim f$. The aim is to estimate the bounded continuous pdf $f(x)$. For fixed x, the bias of $\widehat{f}_{PR}(x)$ is

$$\begin{aligned}
\mathbb{E}\{\widehat{f}_{PR}\} - f(x) &= \frac{1}{b}\mathbb{E}\left\{K\left(\frac{X_i - x}{b}\right)\right\} - f(x) \\
&= \frac{1}{b}\int_{-\infty}^{\infty} K\left(\frac{y-x}{b}\right) f(y)\, dy - f(x)\int_{-\infty}^{\infty} K(y) dy \\
&= \int_{-\infty}^{\infty} \{f(x+y) - f(x)\}\frac{1}{b}K\left(\frac{y}{b}\right) dy
\end{aligned} \tag{1.119}$$

Now let $\delta > 0$. Then the bias is

$$\begin{aligned}
\mathbb{E}\{\widehat{f}_{PR}\} - f(x) = &\int_{-\delta}^{\delta} \{f(x+y) - f(x)\}\frac{1}{b}K\left(\frac{y}{b}\right) dy \\
&+ \int_{|y|>\delta} \{f(x+y) - f(x)\}\frac{1}{b}K\left(\frac{y}{b}\right) dy.
\end{aligned} \tag{1.120}$$

Taking the absolute value,

$$|\mathbb{E}\{\widehat{f}_{PR}\} - f(x)| \leq A_n + B_n \tag{1.121}$$

where

$$\begin{aligned}
A_n &= \sup_{|y| \leq \delta} |f(x+y) - f(x)| \times \int_{-\delta}^{\delta} \frac{1}{b}\left|K\left(\frac{y}{b}\right)\right| dy \\
&\leq \sup_{|y| \leq \delta} |f(x+y) - f(x)| \times \int_{-\infty}^{\infty} \frac{1}{b}\left|K\left(\frac{y}{b}\right)\right| dy \\
&= \sup_{|y| \leq \delta} |f(x+y) - f(x)|.
\end{aligned} \tag{1.122}$$

For B_n, we have

$$B_n = \int_{|y|>\delta} f(x+y)\frac{1}{b}\left|K\left(\frac{y}{b}\right)\right| dy + \int_{|y|>\delta} f(x)\frac{1}{b}\left|K\left(\frac{y}{b}\right)\right| dy \tag{1.123}$$

Now, the first term on the right-hand side of B_n is

$$\begin{aligned} C_{1n} &= \int_{|y|>\delta} \frac{f(x+y)}{y}\frac{y}{b}\left|K\left(\frac{y}{b}\right)\right| dy \\ &\leq \frac{1}{\delta}\sup_{|u|>\delta/b}|uK(u)| \times \int_{|y|>\delta} f(x+y)\,dy \\ &\leq \frac{1}{\delta}\sup_{|u|>\delta/b}|uK(u)| \int_{-\infty}^{\infty} f(y)\,dy. \end{aligned} \tag{1.124}$$

The second term is

$$C_{2n} = f(x) \times \int_{|u|>\delta/b} |K(u)|du.$$

The above derivation shows that if we let b tend to zero as n tends to infinity, and then let δ go to zero, both C_{1n} and C_{2n} and hence B_n, as well as A_n will converge to zero, thus proving asymptotic unbiasedness of the estimator $\widehat{f}_{PR}$.

1.3.2.4 Leading terms: bias, variance, mean squared error

In the discussion above, the kernel need not have assumed non-negative values only. In what follows, unless otherwise specified, we will let the kernel $K(u)$, $u \in \mathbb{R}$, be non-negative, specifically a continuous probability density function (pdf). Being a pdf, such a kernel ensures that $\widehat{f}_{PR}$ itself is a pdf. In addition, the kernel may be assumed to have bounded derivatives up to a certain order, be symmetric about zero, and satisfy some moment conditions such as having a non-zero second moment and finite fourth moment. Kernels that are not pdfs, e.g., kernels assuming both negative and positive values (Parzen 1962, Bartlett 196), asymmetric kernels (Chen 2000, 2002), boundary kernels (Müller 1993), as well as kernels satisfying other moment conditions such as several vanishing moments (higher-order kernels; see Parzen 1962, Bartlett 1963, and Gasser and Müller 1984) are also used. Further conditions may be added depending on the problem at hand.

Following Parzen (1962) and Rosenblatt (1956), we will consider mean square consistency of the curve estimator. Pointwise weak consistency can be shown by establishing convergence of the mean squared error (mse) to zero, via Chebyshev's inequality. Various authors have also considered strong consistency and uniform consistency. Some references are Parzen (1962), Nadaraya (1965), Schuster (1969), Van Ryzin (1969), and Silverman (1978), among others.

As for uniform consistency, Parzen (1962) imposes a condition on the characteristic function (*Fourier transform*) of K. This leads to a simple proof and this is outlined further below. Watson and Leadbetter (1963) derive optimal kernels that the minimize mean integrated squared error (MISE). However, their solution depends on the unknown density. Cline (1988) defines admissible kernels in terms of the MISE and establishes characteristic function based conditions for admissibility. For exact mse and MISE calculations, see Fryer (1976), Deheuvels (1977), and Marron and Wand (1992). Among other authors, Epanechnikov (1969) considers non-negative kernels for twice differentiable densities; also see Sacks and Ylvisaker (1981) and Farrell (1967).

For the discussion here, we let K be a symmetric continuous probability density function, specifically,

$$\begin{aligned} &(i)\ \ K(u) \geq 0, (ii) \int_{-\infty}^{\infty} K(u)\, du = 1, (iii)\, K(u) = K(-u), \\ &(iv)\ \ \int_{-\infty}^{\infty} u^2 K(u)\, du \neq 0, (v) \int_{-\infty}^{\infty} |u|^3 K(u)\, du < \infty, \\ &(vi)\ \ \sup_{u \in \mathbb{R}} K(u) < \infty. \end{aligned} \tag{1.125}$$

Note in particular that (*ii*) and (*vi*) imply that the kernel is square integrable, i.e.,

$$\int_{-\infty}^{\infty} K^2(u)\, du < \infty. \tag{1.126}$$

On the other hand, (*iii*) imples that all odd order moments of K when they exist vanish and that the characteristic function of K is real.

The bandwidth b is such that as $n \to \infty$,

$$(i)\ b \to 0, (ii)\ nb \to \infty. \tag{1.127}$$

Other conditions on b will be mentioned in the sequel.

As for the probability density function f, we assume that its first and second derivatives are continuous, and the third derivative is bounded. In the context of deriving the mean integrated squared error, we will also assume square integrability of f and its derivatives as required.

To assess the convergence of the density estimator to the unknown density f, we will look at the mean squares error (mse) of the density estimator and its integral (MISE); see Parzen (1962), Rosenblatt (1956, 1991), Farrell (1972), Silverman (1986), Nadraya (1989), Prakasa Rao (1983), as well as Birgé and Massart (1995), among others, for various early developments and basic results. For exact calculations, see Marron and Wand (1992) and references therein. However, other distances have also been considered. For L_1-norm (mean absolute deviation) based results and references see Devroye (1987).

For iid data, derivation of the asymptotic expressions for bias and variance of $\widehat{f}_{PR}(x)$ is relatively simple, where one uses Taylor series expansion of appropriate quantities. First of all, let x be fixed.

Since $X_1, X_2, \ldots, X_n$ are iid random variables with pdf f,

$$\mathbb{E}(\widehat{f}_{PR}(x)) = \frac{1}{b}\int K((u-x)/b)f(u)\,du. \tag{1.128}$$

Substituting $y = (u-x)/b$, $u = x + by$ and $du = bdy$ we have

$$\mathbb{E}(\widehat{f}_{PR}(x)) = \int K(y)f(x+by)\,dy. \tag{1.129}$$

By Taylor series expansion, $f(x+by) = f(x) + byf^{(1)}(x) + b^2y^2f^{(2)}(x)/2! + O(|by|^3)$ so that by substitution and applying the properties of the kernel K, we have

$$\mathbb{E}(\widehat{f}_{PR}(x)) = f(x) + b^2f^{(2)}(x)\int y^2K(y)\,dy/2 + o(b^2). \tag{1.130}$$

Thus one may summarize:

Bias of $\widehat{f}_{PR}(x)$:

$$\begin{aligned}\mathbb{B}ias(\widehat{f}_{PR}(x)) &= \mathbb{E}(\widehat{f}_{PR}(x)) - f(x)\\ &= \frac{b^2}{2}f^{(2)}(x)\int_{-\infty}^{\infty} u^2K(u)\,du + o(b^2), \text{as } n \to \infty.\end{aligned} \tag{1.131}$$

Therefore, the nonparametric density estimator is not unbiased in finite samples. However, due to the assumption on the bandwidth b, it is asymptotically unbiased and the leading term in the asymptotic expression for bias is given as above. As for the variance, due to the iid assumption about the observations, being a sample mean,

$$\mathbb{V}ar\,(\widehat{f}_{PR}(x)) = \frac{1}{nb^2}\mathbb{V}ar\left\{K\left(\frac{X_j - x}{b}\right)\right\}. \tag{1.132}$$

However,

$$\mathbb{V}ar\left\{K\left(\frac{X_j - x}{b}\right)\right\} = \int_{-\infty}^{\infty} K^2\left(\frac{u-x}{b}\right) f(u)\,du - \left\{\int_{-\infty}^{\infty} K\left(\frac{u-x}{b}\right) f(u)\,du\right\}^2. \tag{1.133}$$

Now using Taylor series expansion and arguing as above, the result follows by applying the properties of b when $n \to \infty$ and K; we have:

Variance of $\widehat{f}_{PR}(x)$:

$$\mathbb{V}ar\,(\widehat{f}_{PR}(x)) = \frac{1}{nb} f(x) \int_{-\infty}^{\infty} K^2(u)\,du + o\left(\frac{1}{nb}\right) \text{ as } n \to \infty \tag{1.134}$$

Combining the above, the asymptotic expression for the mse is

mse of $\widehat{f}_{PR}(x)$:

$$\begin{aligned} mse(\widehat{f}_{PR}(x)) &= \{Bias\,(\widehat{f}_{PR}(x))\}^2 + Var(\widehat{f}(x)) \\ &= \left\{\frac{b^2}{2} f^{(2)}(x) \int_{-\infty}^{\infty} u^2 K(u)\,du + o(b^2)\right\}^2 \\ &\quad + \frac{1}{nb} f(x) \int_{-\infty}^{\infty} K^2(u)\,du + o\left(\frac{1}{nb}\right), \end{aligned} \tag{1.135}$$

so that as $n \to \infty$, the *leading term* in the mse is

$$\begin{aligned} AMSE(\widehat{f}_{PR}(x)) &= \frac{b^4}{4}\{f^{(2)}(x)\}^2 \left\{\int_{-\infty}^{\infty} u^2 K(u)\,du\right\}^2 \\ &\quad + \frac{1}{nb} f(x) \int_{-\infty}^{\infty} K^2(u)\,du \end{aligned} \tag{1.136}$$

Using the notations $R(g) = \int_{-\infty}^{\infty} g^2(u)\,du$ and $\mu_2(g) = \int_{-\infty}^{\infty} u^2 g(u)\,du$ for appropriately defined function g, we have

$$AMSE(\widehat{f}_{PR}(x)) = \frac{b^4}{4}\{f^{(2)}(x)\}^2\mu_2^2(K) + \frac{1}{nb}f(x)R(K). \tag{1.137}$$

Similarly, the leading term in the asymptotic integrated mean squared error is defined as

$$AMISE\,(\widehat{f}_{PR}) = \int_{-\infty}^{\infty} AMSE(\widehat{f}_{PR}(x))\,dx \tag{1.138}$$

$$= \frac{b^4}{4}R(f^{(2)})\mu_2^2(K) + \frac{1}{nb}R(K). \tag{1.139}$$

1.3.2.5 Central limit theorem

The Parzen–Rosenblatt density estimator can be viewed as a sample mean of terms like

$$w_{i,n}(x) = (1/b)K((X_i - x)/b) \tag{1.140}$$

where for every fixed $x \in \mathbb{R}$, $w_{i,n}(x)$ are iid with the same distribution as

$$w_n(x) = (1/b)K((X - x)/b), \tag{1.141}$$

$f(x)$ being the pdf of X, estimation of which is of interest in the current context.

When $X_1, \ldots, X_n$ are iid with pdf f, various versions of the central limit theorem are available that ensure the pointwise convergence of the rescaled and centered $\widehat{f}_{PR}(x)$ to the standard normal, where $x \in \mathbb{R}$. A necessary and sufficient condition (Loève 1960) is given in Parzen (1962), namely that, for every $\epsilon > 0$, as $n \to \infty$, and fixed $x \in \mathbb{R}$,

$$nP\left[n^{-1/2}|w_n(x) - \mathbb{E}(w_n(x))|/\sqrt{\mathbb{V}ar\,(w_n(x))} \geq \epsilon\right] \to 0, \tag{1.142}$$

a sufficient condition for which is, for some $\delta > 0$,

$$\mathbb{E}[n^{-\delta/2}|w_n(x) - \mathbb{E}(w_n(x))|^{2+\delta}]/(\mathbb{V}ar\,(w_n(x)))^{\delta} \to 0, \tag{1.143}$$

as $n \to \infty$. A sufficient condition for this is

$$\int_{-\infty}^{\infty} |K(u)|^{2+\delta} du < \infty. \tag{1.144}$$

Also a Berry–Esseen bound gives an appreciation of the error in normal approximation. It is, for a suitable constant $C > 0$,

$$\sup_z \left| P\left[\frac{\widehat{f}_{PR}(x) - \mathbb{E}(\widehat{f}_{PR}(x))}{\sqrt{\mathbb{V}ar\,(\widehat{f}_{PR}(x))}} \le z \right] - \Phi(z) \right| \le C \frac{n^{-1/2}\mathbb{E}|w_n(x)|^3}{\mathbb{V}ar\,(w_n(x))^{3/2}}$$

$$\sim \frac{1}{(nbf(x))^{1/2}} \left[\frac{\int_{-\infty}^{\infty} |K(y)|^3 dy}{\left(\int_{-\infty}^{\infty} K^2(y)dy\right)^{1/2}} \right]. \tag{1.145}$$

For additional details, see Parzen (1962).

One interesting fact concerns the covariance between density estimates at x_1 and x_2 where $x_1 \neq x_2$ are fixed. It turns out that as $n \to 0$, this covariance converges to zero, i.e., the density estimates have a local characteristic. This is easy to see, since

$$\begin{aligned}
\mathbb{C}ov[\widehat{f}_{PR}(x_1), \widehat{f}_{PR}(x_2)] &= \frac{1}{n^2 b^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}ov\left[K\left(\frac{X_i - x_1}{b}\right), K\left(\frac{X_j - x_2}{b}\right) \right] \\
&= \frac{1}{nb^2} \mathbb{C}ov\left[K\left(\frac{X_i - x_1}{b}\right), K\left(\frac{X_i - x_2}{b}\right) \right] \\
&= \frac{1}{nb^2} \left[\int_{-\infty}^{\infty} K\left(\frac{u - x_1}{b}\right) K\left(\frac{u - x_2}{b}\right) f(u)\, du \right] \\
&\quad - \frac{1}{nb^2} \left[\left(\int_{-\infty}^{\infty} K\left(\frac{u - x_1}{b}\right) f(u)\, du \right) \left(\int_{-\infty}^{\infty} K\left(\frac{u - x_2}{b}\right) f(u)\, du \right) \right] \\
&= \frac{1}{nb} \left[\int_{-\infty}^{\infty} K(y) K\left(y - \frac{x_2 - x_1}{b}\right) f(x_1 + by)\, dy - bf(x_1)f(x_2) + o(b^2) \right] \\
&= O\left(\frac{1}{nb}\right),
\end{aligned} \tag{1.146}$$

and the kernel K is such that $K(u) \to 0$ as $|u| \to \infty$. In other words,

$$\mathbb{C}ov[(nb)^{-1/2}\widehat{f}_{PR}(x_1), (nb)^{-1/2}\widehat{f}_{PR}(x_2)] \to 0 \tag{1.147}$$

as $n \to \infty$ where x_2 and x_1 are fixed and distinct real numbers.

The multivariate central limit theorem (Bradley 1983) can be used to prove the asymptotic multivariate normal distribution of the rescaled and centered density estimates computed at distinct and fixed real numbers $x_1, x_2, \ldots x_p$, where $p \ge 1$ is a finite integer. Specifically, consider the random vector

$$\mathbf{Z}_n = (Z_{1,n}, Z_{2,n}, \ldots, Z_{p,n})' \tag{1.148}$$

where, for $i = 1, 2, \ldots, p$,

$$Z_{i,n} = (nb)^{-0.5}[\widehat{f}_{PR}(x_i) - \mathbb{E}(\widehat{f}_{PR}(x_i))]. \tag{1.149}$$

Then as $n \to \infty$, $\mathbf{Z}_n$ converges in distribution to the multivariate normal distribution with zero mean and a $p \times p$ covariance matrix,

$$\mathbf{\Sigma}_f = \begin{pmatrix} \sigma_{1,1} & \sigma_{2,1} & \cdots & \sigma_{p,1} \\ \sigma_{1,2} & \sigma_{2,2} & \cdots & \sigma_{p,2} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1,p} & \sigma_{2,p} & \cdots & \sigma_{p,p} \end{pmatrix}, \tag{1.150}$$

where

$$\sigma_{i,i} = f(x_i) \int_{-\infty}^{\infty} K^2(u)\, du \tag{1.151}$$

and

$$\sigma_{i,j} = 0, \text{if } i \neq j. \tag{1.152}$$

1.3.2.6 Weak uniform consistency

For every fixed $x \in \mathbb{R}$, convergence of the bias and the variance of the estimated density ensures pointwise weak consistency. Some authors have considered uniform consistency, e.g., Parzen (1962). The aim is to prove uniform convergence in probability of the estimator to the unknown density function.

$$\lim_{n\to\infty} P\left\{ \sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - f(x)| > \epsilon \right\} = 0. \tag{1.153}$$

Due to Markov's inequality, a sufficient condition is

$$P\left\{ \sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - f(x)| > \epsilon \right\} < \frac{\mathbb{E}\left\{ \sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - f(x)| \right\}}{\epsilon} \tag{1.154}$$

However,

$$\begin{aligned} \mathbb{E}\left\{ \sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - f(x)| \right\} \leq \mathbb{E}\left\{ \sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - \mathbb{E}(\widehat{f}_{PR}(x))| \right\} \\ + \left\{ \sup_{x\in\mathbb{R}} |\mathbb{E}(\widehat{f}_{PR}(x)) - f(x)| \right\}. \end{aligned} \tag{1.155}$$

The convergence of the second term on the right to zero can, for instance, be established by requiring that f is three times continuously differentiable, its third derivative being bounded. This was the line of argument used for deriving the asymptotic expression for the bias. In other words, a sufficient condition for weak uniform consistency is

$$\mathbb{E}\left\{\sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - \mathbb{E}(\widehat{f}_{PR}(x))|\right\} \to 0 \tag{1.156}$$

as $n \to \infty$. To prove this, Parzen considers a kernel that has a characteristic function that is absolutely integrable on the real line. So let ψ be the characteristic function of K, i.e.,

$$\psi_K(t) = \int_{-\infty}^{\infty} exp(-\imath tx)K(x)\,dx \tag{1.157}$$

where $\imath = \sqrt{-1}$ and $t \in \mathbb{R}$. Then substitution yields

$$\widehat{f}_{PR}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\imath tx}\psi_K(bt)c_n(t)\,dt, \tag{1.158}$$

where c_n is the empirical characteristic function (ecf) for the data $X_1, X_2, \ldots, X_n$, i.e.,

$$c_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{\imath tX_j}. \tag{1.159}$$

For fixed $t \in \mathbb{R}$, $c_n(t)$ is a complex valued random variable. Let $C(t)$ be the characteristic function for the random variable X with pdf f, i.e.,

$$C(t) = \int_{-\infty}^{\infty} exp(-\imath tx)f(x)\,dx. \tag{1.160}$$

Then the ecf $c_n(t)$ is complex valued. Moreover, its real and the imaginary parts are

$$re(c_n(t)) = \frac{1}{n}\sum_{j=1}^{n} cos(tX_j), \tag{1.161}$$

$$im(c_n(t)) = \frac{1}{n}\sum_{j=1}^{n} sin(tX_j), \tag{1.162}$$

with expectations being equal to the real and the imaginary parts of $C(t)$ respectively. Considering

$$Y_{1,n}(t) = \sqrt{n}(re(c_n(t)) - re(C(t))) \tag{1.163}$$

$$Y_{2,n}(t) = \sqrt{n}(im(c_n(t)) - im(C(t))) \tag{1.164}$$

as real-valued stochastic processes in t (see, for example, Feuerverger and Mureika 1977), it is easy to check that for t_1, $t_2 \in \mathbb{R}$,

$$\begin{aligned}\mathbb{C}ov\,(Y_{1,n}(t_1), Y_{1,n}(t_2)) &= \frac{1}{2}[re(C(t_1 + t_2)) + re(C(t_1 - t_2))] \\ &\quad - re(C(t_1))re(C(t_2)) \\ \mathbb{C}ov\,(Y_{2,n}(t_1), Y_{2,n}(t_2)) &= \frac{1}{2}[-re(C(t_1 + t_2)) + re(C(t_1 - t_2))] \\ &\quad - im(C(t_1))im(C(t_2)).\end{aligned} \tag{1.165}$$

Also, $|e^{\imath y}| = \sqrt{sin^2(y) + cos^2(y)} = 1$, where $y \in \mathbb{R}$, and for a random variable X with finite second moment,

$$\mathbb{E}(|X|) \leq \sqrt{\mathbb{E}(X^2)}. \tag{1.166}$$

Finally, note that

$$\begin{aligned}\mathbb{E}|c_n(t) - \mathbb{E}(c_n(t))|^2 &= \mathbb{E}[re(c_n(t)) - \mathbb{E}\{re(c_n(t))\}]^2 \\ &\quad + \mathbb{E}[im(c_n(t)) - \mathbb{E}\{im(c_n(t))\}]^2 \\ &= \mathbb{V}ar\,[re(c_n(t))] + \mathbb{V}ar\,[im(c_n(t))] \\ &= O\left(\frac{1}{n}\right)\end{aligned} \tag{1.167}$$

Then as $n \to \infty$,

$$\begin{aligned}&\mathbb{E}\left\{\sup_{x\in\mathbb{R}} |\widehat{f}_{PR}(x) - \mathbb{E}(\widehat{f}_{PR}(x))|\right\} \\ &= \mathbb{E}\left\{\sup_{x\in\mathbb{R}} \left|\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\imath tx}\psi_K(bt)\left\{c_n(t) - \mathbb{E}(c_n(t))\right\}dt\right|\right\} \\ &\leq \mathbb{E}\left\{\sup_{x\in\mathbb{R}} \frac{1}{2\pi}\int_{-\infty}^{\infty} |e^{-\imath tx}\psi_K(bt)[c_n(t) - \mathbb{E}(c_n(t))]|dt\right\} \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\psi_K(bt)| \cdot \mathbb{E}|c_n(t) - \mathbb{E}(c_n(t))|dt \\ &\leq \frac{1}{2\pi}\int_{-\infty}^{\infty} |\psi_K(bt)| \cdot \sqrt{\mathbb{E}|c_n(t) - \mathbb{E}(c_n(t))|^2}dt\end{aligned}$$

$$= O\left(\frac{1}{\sqrt{n}}\right)\int_{-\infty}^{\infty}|\psi_K(bt)|dt$$
$$= O\left(\frac{1}{b\sqrt{n}}\right). \tag{1.168}$$

In particular, the above quantity converges to zero if $nb^2 \to \infty$ as $n \to \infty$.

1.3.3 Bandwidth selection

Given a kernel K, the asymptotic property of $\widehat{f}_{PR}(x)$ depends on the smoothness of the pdf f near x and on the bandwidth b. This is evident from the asymptotic expressions (leading terms) for the mean squared error (AMSE) or the mean integrated squared error (AMISE) defined earlier. In particular, minimization of these quantities leads to algorithms for optimal bandwidth selection. For example, a *local optimum bandwidth* $b_{opt}^{(local)}(x)$ can be obtained by minimizing $AMSE(\widehat{f}_{PR}(x))$ with respect to b at every fixed x. Similarly, a *global optimum bandwidth* $b_{opt}^{(global)}$ can be obtained by minimizing $AMISE(\widehat{f}_{PR})$. One simply takes the derivative of the AMISE or the AMSE with respect to b and equates the resulting expression to zero, solving for b_{opt}. Fluctuations in f can be seen to affect b_{opt}. For instance, large values of $R(f^{(2)})$ and $f^{(2)}(x)$ lead to high AMISE and AMSE respectively, and consequently low $b_{opt}^{(global)}$ and $b_{opt}^{(global)}$. We have

$$b_{opt}^{(global)} = \underset{b}{\text{argmin}}\ AMISE(\widehat{f}_{PR}) \tag{1.169}$$

whereas

$$b_{opt}^{(local)}(x) = \underset{b}{\text{argmin}}\ AMSE(\widehat{f}_{PR}(x)). \tag{1.170}$$

Specifically,

$$b_{opt}^{(global)} = \left(\frac{R(K)}{R(f^{(2)})\mu_2^2(K)}\right)^{1/5} n^{-1/5} \text{ and} \tag{1.171}$$

$$b_{opt}^{(local)}(x) = \left(\frac{f(x)R(K)}{(f^{(2)}(x))^2\mu_2^2(K)}\right)^{1/5} n^{-1/5}. \tag{1.172}$$

When these formulas for the local optimum bandwidth and the global optimum bandwidth are substituted in the corresponding expressions for AMSE and AMISE respectively, we obtain the rate $n^{-4/5}$ for the mean squared error and also for the integrated mean squared error. For the local optimal choice of the bandwidth, we have

$$AMSE(\widehat{f}_{PR}(x))|_{b=b_{opt}^{(local)}} = n^{-4/5} \cdot \frac{5}{4}[\{f(x)\}^{4/5}\{f^{(2)}(x)\}^{2/5}\{\mu_2(K)\}^{2/5}\{R(K)\}^{4/5}] \tag{1.173}$$

and for the global optimal choice, the leading term in the asymptotic integrated mean squared error is

$$AMISE\,(\widehat{f}_{PR})|_{b=b_{opt}^{(global)}} = n^{-4/5} \cdot \frac{5}{4}[\{R(f^{(2)})\}^{1/5}\{\mu_2(K)\}^{2/5}\{R(K)\}^{4/5}]. \tag{1.174}$$

Of course, this $n^{-4/5}$ rate, though an improvement over the histogram, is still slower than the parametric n^{-1} rate (see, however, Hall and Marron 1987). The above formulas for the optimum bandwidth contain unknown functions such as the unknown pdf and its second derivative. As a result, these formulas cannot be used directly for real data and data-driven solutions are called for.

Numerous bandwidth selection procedures have been proposed in the literature. In what follows, we discuss a selection of four methods that are often used by practitioners. This selection is by no means complete. The reader is strongly recommended to read through Gasser et al. (1991), Sheather and Jones (1991), Jones et al. (1996), Loader (1999), Sheather (1992, 2004), and references therein for other bandwidth selection procedures and additional information on their asymptotic convergence.

In general, it may be said that there is no "best" method of bandwidth selection, and to understand the underlying structure of the unknown pdf, a good idea is to consider a sequence of bandwidths and compare results. This idea was used in another context by Silverman (1981), namely for testing multimodality of a density function where one exploits the idea that if a pdf is multimodal, then a large bandwidth is needed to arrive at a unimodal estimate.

Let m be the number of modes of the unknown pdf f. To test

$$H_o : m \leq k \text{ versus } H_1 : m > k, \tag{1.175}$$

where often $k = 1$, one defines a critical bandwidth b_c that is the minimum bandwidth required to obtain $\widehat{f}_{PR}$ with maximum k modes so that the null hypothesis is rejected if b_c is large. See Silverman (1981) for additional information.

1.3.3.1 Likelihood cross-validation

The idea here is to treat the bandwidth b as a parameter and choose b such that its likelihood expressed in terms of the kernel density estimator is maximized; see Stone (1974), Geisser (1975), Habbema et al. (1974), and Duin (1976); also see, for example, Hall (1982, 1987) and Titterington (1980). In order to have replicates of the density estimate, however, one observation is left out at each of the n steps.

Using bandwdth b, the *leave-one-out* density estimator is defined as

$$\widehat{f}_{PR,-i}(x) = \frac{1}{(n-1)b} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x - X_j}{b}\right) \tag{1.176}$$

which uses the observations $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ while leaving out X_i. Substituting $x = X_i$ one gets the *leave-one-out* density estimator of $f(X_i)$ as

$$\widehat{f}_{PR,-i}(X_i) = \frac{1}{(n-1)b} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{X_i - X_j}{b}\right). \tag{1.177}$$

At the second step, the *log-likelihood* is averaged over each choice of the omitted X_i, i.e.,

$$LCV(b) = \frac{1}{n} \sum_{i=1}^{n} log\{\widehat{f}_{PR,-i}(X_i)\}. \tag{1.178}$$

Finally, the optimum bandwidth is obtained as

$$\widehat{b}_{LCV} = \underset{b}{\text{argmax}}\ LCV(b). \tag{1.179}$$

Likelihood cross-validation is equivalent to minimization of the Kulback–Leibler loss function $I(f, \widehat{f}_{PR})$ in the sense that $LCV(b)$ is an asymptotically unbiased estimator of a constant minus the expected value of the Kulback–Leibler loss function. This can be seen roughly by noting that

$$\begin{aligned} I(f,\widehat{f}_{PR}) &= \int f(x) log\left(\frac{f(x)}{\widehat{f}_{PR}(x)}\right) dx \\ &= \int f(x) log(f(x))\, dx - \int f(x) log(\widehat{f}_{PR}(x))\, dx \end{aligned} \quad (1.180)$$

and the expected value of $LCV(b)$ is, as $n \to \infty$,

$$\begin{aligned} \mathbb{E}\{LCV(b)\} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\{log\widehat{f}_{PR,-i}(X_i)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}[\mathbb{E}_{X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}\{log\widehat{f}_{PR,-i}(X_i)\}] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n} \int_{-\infty}^{\infty} f(x) log\widehat{f}_{PR,-i}(x)\, dx \\ &\approx \mathbb{E}\left\{\int_{-\infty}^{\infty} f(x) log\widehat{f}_{PR}(x)\, dx\right\} \\ &= \int f(x) log(f(x))\, dx - \mathbb{E}\{I(f,\widehat{f}_{PR})\}. \end{aligned} \quad (1.181)$$

Implementation of the likelihood cross-validation approach to data is very simple and the method is intuitively appealing. Like the maximum likelihood estimation method, this procedure also is linked to the Kulback–Leibler information loss where two proability density functions are compared. However, its direct use without further considerations may create difficulties when the pdf f has infinite support whereas the kernel K does not. In this case, $I(f, \widehat{f}_{PR})$ may be infinite. The essential reason for this is due to the fact that $log(\widehat{f}_{PR}(x))$ will tend to $-\infty$ whenever $\widehat{f}_{PR}(x)$ approaches zero, which is likely to happen if, unlike f, K is restricted to have a bounded support. For similar reasons, the method may be sensitive to outliers in the data. For further discussions see Chow et al. (1983), Hall (1987), and Silverman (1986).

1.3.3.2 Least squares cross-validation

In this approach (Rudemo 1982 and Bowman 1984; also see Bowman et al. 1984, Hall 1983, and Stone 1984), one minimizes an unbiased estimator of the (shifted) integrated mean squared error of the kernel estimator of the density f. (This method is also known as the *unbiased cross-validation*). Thus consider the integrated mean squared error,

$$\begin{aligned} MISE(\widehat{f}_{PR}(x)) &= \int_{-\infty}^{\infty} E\{\widehat{f}_{PR}(x) - f(x)\}^2 dx \\ &= \int_{-\infty}^{\infty} E\{\widehat{f}_{PR}(x)\}^2 dx + \int_{-\infty}^{\infty} f^2(x)\, dx \\ &\quad - 2\int_{-\infty}^{\infty} E\{\widehat{f}_{PR}(x) f(x)\} dx. \end{aligned} \tag{1.182}$$

The aim is then to find the bandwidth b that minimizes $MISE(\widehat{f}_{PR}(x))$ or equivalently

$$\begin{aligned} MISE(\widehat{f}_{PR}(x)) - \int_{-\infty}^{\infty} f^2(x)\, dx &= \int_{-\infty}^{\infty} E\{\widehat{f}_{PR}(x)\}^2 dx \\ &\quad - 2\int_{-\infty}^{\infty} E\{\widehat{f}_{PR}(x) f(x)\} dx. \end{aligned} \tag{1.183}$$

It turns out that an unbiased estimator of the above quantity is

$$LSCV(b) = \int_{-\infty}^{\infty} \{\widehat{f}_{PR}(x)\}^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{PR,-i}(X_i) \tag{1.184}$$

where, as we have seen earlier, $\widehat{f}_{PR,-i}(X_i)$ is the *leave-one-out* density estimator evaluated at $x = X_i$ where

$$\widehat{f}_{PR,-i}(x) = \frac{1}{(n-1)b} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{x - X_j}{b}\right) \tag{1.185}$$

and $\widehat{f}_{PR}(x)$ is the kernel density estimator using the full set of observations.

To see why this is the case, it is obvious that $\int_{-\infty}^{\infty} \{\widehat{f}(x)\}^2 dx$ is an unbiased estimator of its expectation. As for the second term in

$LSCV(b)$, since $X_1, X_2, \ldots, X_n$ are iid,

$$\begin{aligned}
&\mathbb{E}\{\widehat{f}_{PR,-i}(X_i)\} \\
&\quad = \mathbb{E}_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}[\mathbb{E}_{X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}\{\widehat{f}_{PR,-i}(X_i)\}] \\
&\quad = \mathbb{E}_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}\left[\int_{-\infty}^{\infty}\widehat{f}_{PR,-i}(x)f(x)dx\right] \\
&\quad = \mathbb{E}\left[\int_{-\infty}^{\infty}\widehat{f}_{PR}(x)f(x)dx\right], \qquad (1.186)
\end{aligned}$$

so that

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{PR,-i}(X_i)\right\} = \int_{-\infty}^{\infty}\mathbb{E}\{\widehat{f}_{PR}(x)f(x)\}dx. \qquad (1.187)$$

Remarks on computation of LSCV(b)

Let $K_{(2)}(u)$ denote the convolution of $K(u)$ with itself, i.e.,

$$(K \otimes K)(v) = K_{(2)}(v) = \int_{-\infty}^{\infty}K(u)K(v-u)du. \qquad (1.188)$$

The above can be used to further simplify the expression for $LSCV(b)$ so that, for instance, integration of $\widehat{f}_{PR}^2$ can be avoided and an exact expression can be given as follows:

$$\begin{aligned}
\int_{-\infty}^{\infty}\{\widehat{f}_{PR}(x)\}^2dx &= \int_{-\infty}^{\infty}\left[\frac{1}{nb}\sum_{i=1}^{n}K\left(\frac{X_i-x}{b}\right)\right]^2dx \\
&= \frac{1}{n^2b^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\int_{-\infty}^{\infty}K\left(\frac{X_i-x}{b}\right)K\left(\frac{X_j-x}{b}\right)dx\right\} \\
&= \frac{1}{n^2b}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}K(u)K\left(\frac{X_i-X_j}{b}-u\right)du \\
&= \frac{1}{n^2b}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{(2)}\left(\frac{X_i-X_j}{b}\right) \qquad (1.189)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{PR,-i}(X_i) &= \frac{2}{n}\sum_{i=1}^{n}\frac{1}{(n-1)b}\sum_{\substack{j=1\\j\neq i}}^{n}K\left(\frac{X_i-X_j}{b}\right) \\
&= \frac{2}{n}\sum_{i=1}^{n}\frac{1}{(n-1)b}\left\{\sum_{j=1}^{n}K\left(\frac{X_i-X_j}{b}\right)-K(0)\right\}. \qquad (1.190)
\end{aligned}$$

In other words, the $LSCV(b)$ can be rewritten as

$$LSCV(b) = \frac{1}{n^2 b}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{(2)}\left(\frac{X_i - X_j}{b}\right) + \frac{2}{(n-1)b}K(0) - \frac{2}{n(n-1)b}\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{b}\right), \quad (1.191)$$

where $K(0)$ is the value of the kernel $K(u)$ evaluated at $u = 0$. When $n \to \infty$,

$$LSCV(b) \approx \frac{1}{n^2 b}\mathbf{1}'(\mathbf{A_2} - 2\mathbf{A_1} + 2\mathbf{A_0})\mathbf{1} \quad (1.192)$$

where $\mathbf{a}'$ denotes the transpose of a vector $\mathbf{a}$, $\mathbf{A_0}$, $\mathbf{A_1}$ and $\mathbf{A_2}$ are $n \times n$ matrices with the (i, j)th elements being $K(0)$, $K((X_i - X_j)/b)$ and $K_{(2)}((X_i - X_j)/b)$ respectively and $\mathbf{1}$ is a vector of 1's of length n. Since K is typcally a known pdf, its convolutions are also known. For example, consider a density that is closed under convolution, such as the Gaussian family. Then if K is the pdf of a standard normal distribution

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2} \quad (1.193)$$

then its convolution with itself $K_{(2)}$ is the pdf of a normal distribution with zero mean and variance 2, i.e.,

$$K_{(2)}(u) = \frac{1}{2\sqrt{\pi}}e^{-u^2/4}. \quad (1.194)$$

When K is symmetric, $\mathbf{A_1}$ and $\mathbf{A_2}$ will also be symmetric, resulting in further computational advantages.

1.3.3.3 Biased cross-validation

This approach is suggested by Scott and Terrell (1987). In this method, the focus is on AMISE, where $R(f^{(2)})$ is replaced by

$$R(\tilde{f}^{(2)}) = R(\hat{f}^{(2)}_{PR}) - R(K^{(2)})/(nb^5), \quad (1.195)$$

where

$$\hat{f}^{(2)}_{PR}(x) = \frac{d^2}{dx^2}\hat{f}_{PR}(x) = \frac{1}{nb^3}\sum_{i=1}^{n} K^{(2)}\left(\frac{X_i - x}{b}\right) \quad (1.196)$$

This has the advantage that it leads to an expected value equal to the AMISE plus an error of the order $O(1/n)$; see Theorem 3.2 in Scott and Terrell (1987). Here, $R(g) = \int g^2(u)du$ for an appropriately defined function g. Thus, instead of minimizing AMISE, one minimizes

$$BCV(b) = \frac{R(K)}{nb} + \frac{b^4}{4}\mu_2(K)R(\tilde{f}^{(2)}) \tag{1.197}$$

The ratio of the optimal bandwidth b_{BCV} minimizing the BCV criterion and b_{AMISE} that minimizes the AMISE is asymptotically normal, so

$$n^{-1/10}(b_{BCV}/b_{AMISE} - 1) \ \rightarrow \ N(0, \sigma^2_{BCV}) \tag{1.198}$$

as $n \rightarrow \infty$, with $\sigma^2_{BCV} > 0$. See Scott and Terrell (1987) for details. A similar asymptotic rule also exists for the LSCV criterion (see Hall and Marron 1987a and Scott and Terrell 1987):

$$n^{-1/10}(b_{LSCV}/b_{AMISE} - 1) \ \rightarrow \ N(0, \sigma^2_{LSCV}) \tag{1.199}$$

where for the Gaussian kernel, the ratio $\sigma^2_{LSCV}/\sigma^2_{BCV}$ is approximately 15.7; see Wand and Jones (1995). Also see Sheather (2004), among others, for an overview.

1.3.3.4 Reference to a known density

As in the case of a histogram and also for kernel density estimation, the integral of the squared derivative of f appearing in the formula for the global optimal bandwidth is replaced by an estimate (Silverman 1986, Scott 1979, 1992, and Deheuvels 1977) under the assumption that f is close to a known density. In other words, this is a plug-in approach, where the integral of the squared density derivative is estimated using a parametric approach. For example, if f is close to a normal density function (Silverman 1986) with zero mean and variance σ^2, then by differentiation,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \tag{1.200}$$

$$\frac{d}{dx}f(x) = -\frac{x}{\sigma^2}f(x) \tag{1.201}$$

$$\frac{d^2}{dx^2}f(x) = \frac{f(x)}{\sigma^2}\left(\frac{x^2}{\sigma^2} - 1\right) \tag{1.202}$$

so that

$$R(f^{(2)}) = \int_{-\infty}^{\infty} \left(\frac{d^2}{dx^2} f(x) \right)^2 dx$$
$$= \frac{1}{2\sqrt{\pi}\sigma^5} \int_{-\infty}^{\infty} \left(\frac{u^4}{4} - u^2 + 1 \right) \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right) dx$$
$$= \frac{3}{8\sqrt{\pi}\sigma^5} \tag{1.203}$$

which can be used in the formula for the $b_{opt}^{(global)}$ bandwidth given earlier, after substituting an estimate of σ. Since σ is the dispersion parameter, various options for estimating this quantity are available. As in the case of a histogram estimate, one option is to estimate σ as

$$\hat{\sigma} = min(s, IQR_{sample}/1.349),$$

where s is the sample standard deviation, whereas IQR_{sample} is the sample interquartile range and $IQR_{\phi} = 1.349$ is the interquartile range of the standard normal distribution; see Silverman (1986, p. 47).

Similar ideas can be used also when higher order derivatives of the density f are being estimated. In each case, one derives the formula for the optimum bandwidth, which involves an integral of the square of a higher order derivative of f. Then, as for estimating f, here also, using a reference density, a pilot bandwidth can be found, which may perhaps be used as a starting value for subsequent iterations in a bandwidth selection algorithm.

1.3.3.5 Kernel based plug-in methods

The plug-in method consists of substituting an estimate of the integral of the squared derivative of f in the formula for the global optimal bandwidth; see Woodroofe (1970). As for kernel based approaches to estimate the integral of the squared rth derivative of f, appearing in $AMISE\widehat{f}_{PR}(x)$, one idea is to consider an estimator of $f^{(r)}(x)$ that is based on the kernel $K^{(r)}$, such as the rth derivative of K in $\widehat{f}_{PR}(x)$ (Parzen 1962); also see Härdle et al. (1990), Gasser et al. (1985), Efromovich and Low (1996), and Efromovich and Samarov (2000).

An obvious first estimator of $R(f^{(r)})$ is $R(\widehat{f^{(r)}})$; i.e.,

$$R(\widehat{f^{(r)}}) = \int_{-\infty}^{\infty} (\widehat{f^{(r)}}(x))^2 dx \tag{1.204}$$

$$= \frac{1}{n^2 b^{2(r+1)}} \int_{-\infty}^{\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(r)}\left(\frac{X_i - x}{b}\right) \left(\frac{X_j - x}{b}\right) dx$$

$$= \frac{1}{n^2 b^{2r+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{-\infty}^{\infty} K^{(r)}(u) K^{(r)}\left(\frac{X_i - X_j}{b} - u\right) du$$

$$= \frac{1}{n^2 b^{2r+1}} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{(2)}^{(r)}\left(\frac{X_i - X_j}{b}\right) \tag{1.205}$$

where $K_{(2)}^{(r)}$ is the convolution of $K^{(r)}$ with itself.

Alternatively, under suitable conditions on f, integration by parts can be used to show that

$$R(f^{(r)}) = \int_{-\infty}^{\infty} (f^{(r)}(x))^2 dx$$

$$= (-1)^r \int_{-\infty}^{\infty} f^{(2r)}(x) f(x)\, dx$$

$$= (-1)^r \mathbb{E}(f^{(2r)}(X)). \tag{1.206}$$

This suggests the estimator (Hall and Marron 1987)

$$(-1)^r \frac{1}{n} \sum_{i=1}^{n} \widehat{f^{(2r)}}(X_i) \tag{1.207}$$

where $\widehat{f^{(2r)}}$ is an estimator of $f^{(2r)}$.

Note that irrespective of which method is used, while estimating a derivative of f, the bandwidth selection issue (for that function) comes up again. There can be several approaches to handle this problem. Scott et al. (1977) consider an iterative scheme, where in the ith iteration, $b^{(i)}$ is estimated from the formula for b_{opt} (local or global) using a pilot bandwidth $b^{(i-1)}$ and the method iterates until convergence. For a discussion see Silverman (1986, p. 60–61). One possible argument against the use of this method is the fact that the same bandwidth is used to estimate f as well as its derivatives. Various other authors have also suggested other plug-in approaches. Among them, the Sheather

and Jones (1991) method consists of writing b_2 as a function of b, where b_2 is the bandwidth used to estimate $f^{(2)}$, b being the bandwidth used to estimate f itself.

1.4 Multivariate density estimation

The multivariate generalization of the Parzen–Rosenblatt density estimator was due to Cacoullos (1966).

Consider iid multivariate observations $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})' \in \mathbb{R}^d, d \geq 1$, where $i = 1, 2, \ldots, n$, with a common d-dimensional pdf

$$f_d : \mathbb{R}^d \to \mathbb{R}. \tag{1.208}$$

Let f_d be three times partially differentiable, its third-order partial derivatives being bounded.

Then the formula for the Parzen–Rosenblatt density estimator can be extended to higher dimensions and the arguments for consistency follow in a similar manner.

Therefore consider the vector

$$\mathbf{u} = (u_1, u_2, \ldots, u_d)' \in \mathbb{R}^d, d \geq 1, \tag{1.209}$$

where $d \geq 1$ denotes the dimension of the vector and let

$$K_d : \mathbb{R}^d \to \mathbb{R} \tag{1.210}$$

be a d-dimensional kernel such that

$$\begin{aligned}
&K_d(\mathbf{u}) \geq 0,\\
&\int_{\mathbb{R}^d} K_d(\mathbf{u})\, d\mathbf{u} = 1,\\
&\int_{\mathbb{R}^d} u_i K_d(\mathbf{u})\, d\mathbf{u} = 0,\\
&\int_{\mathbb{R}^d} u_i^2 K_d(\mathbf{u})\, d\mathbf{u} = \mu_2(K_d) > 0,\\
&\int_{\mathbb{R}^d} |u_i|^3 K_d(\mathbf{u})\, d\mathbf{u} < \infty.
\end{aligned} \tag{1.211}$$

Let $\mathbf{B}$ be a $d \times d$ non-singular matrix and let its inverse and determinant be $\mathbf{B}^{-1}$ and $|\mathbf{B}|$ respectively. Given observations

$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, the Parzen–Rosenblatt density estimator of f_d is given by

$$\widehat{f}_d(\mathbf{x}) = \frac{1}{n|\mathbf{B}|} \sum_{i=1}^{n} K_d(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x})). \tag{1.212}$$

To simplify matters, one often uses a product kernel, where the d-dimensional kernel is simply a product of d univariate kernels. If one takes these univariate kernels to be the same as K, then

$$K_d(\mathbf{u}) = K(u_1)K(u_2)\cdots K(u_d) \tag{1.213}$$

Moreover, consider bandwidths $b_i > 0, i = 1, 2, \ldots, d$ such that $b_i \to 0$ and $nb_i \to \infty$ as $n \to \infty$. Then the density estimator using these bandwidths and the product kernel becomes

$$\widehat{f}_d(\mathbf{x}) = \frac{1}{nb_1b_2\cdots b_d} \sum_{i=1}^{n} \left[K\left(\frac{X_{i1} - x_1}{b_1}\right) K\left(\frac{X_{i2} - x_2}{b_2}\right) \cdots \right.$$
$$\left. K\left(\frac{X_{id} - x_d}{b_d}\right) \right] \tag{1.214}$$

where now the bandwidth matrix $\mathbf{B}$ is simply a diagonal matrix with positive diagonal elements

$$\mathbf{B} = diag(b_1, b_2, \ldots, b_d) \tag{1.215}$$

so that its determinant is $|\mathbf{B}| = b_1b_2\cdots b_d$. For instance, using the Gaussian kernel, we have

$$\widehat{f}_d(\mathbf{x}) = \frac{1}{n(\sqrt{2\pi})^d \prod_{j=1}^{d} b_j} \sum_{i=1}^{n} exp\left[-\frac{1}{2} \sum_{j=1}^{d} \left(\frac{X_{ij} - x_j}{b_j} \right)^2 \right] \tag{1.216}$$

and if we let $b_j = b$ for all j, then

$$\widehat{f}_d(\mathbf{x}) = \frac{1}{n(\sqrt{2\pi}b)^d} \sum_{i=1}^{n} exp\left[-\frac{1}{2b^2} \sum_{j=1}^{d} (X_{ij} - x_j)^2 \right]. \tag{1.217}$$

To derive the expressions for the bias and the variance of the multivariate density estimator f_d, we define the partial derivatives:

$$f_d^{(1)}(\mathbf{x}) = \left(\frac{\partial f_d(\mathbf{x})}{\partial x_1}, \frac{\partial f_d(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f_d(\mathbf{x})}{\partial x_d} \right)' \tag{1.218}$$

and

$$f_d^{(2)}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f_d(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f_d(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_2^2} & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_2 \partial x_d} \\ \cdots & \cdots & \cdots & \cdots & \\ \frac{\partial^2 f_d(\mathbf{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_d \partial x_2} & \frac{\partial^2 f_d(\mathbf{x})}{\partial x_d \partial x_3} & \cdots & \frac{\partial^2 f_d(\mathbf{x})}{\partial^2 x_d} \end{pmatrix}. \tag{1.219}$$

Then the expected value is

$$\mathbb{E}[\widehat{f}_d(\mathbf{x})] = \frac{1}{|\mathbf{B}|} \mathbb{E}[K_d(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x}))] \tag{1.220}$$

$$= \frac{1}{|\mathbf{B}|} \int_{\mathbb{R}^d} f_d(\mathbf{X}_i) K_d(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x}))\, d\mathbf{X}_i \tag{1.221}$$

$$= \int_{\mathbb{R}^d} f_d(\mathbf{x} + B\mathbf{u}) K_d(\mathbf{u})\, d\mathbf{u} \tag{1.222}$$

Now one may use the multidimensional Taylor series expansion,

$$f_d(\mathbf{x} + \mathbf{Bu}) = f_d(\mathbf{x}) + (\mathbf{Bu})' f^{(1)}(\mathbf{x}) + \frac{1}{2}(\mathbf{Bu})' f^{(2)}(\mathbf{x})(\mathbf{Bu}) + \cdots \tag{1.223}$$

so that, noting that the trace of a scalar matrix is a scalar itself, and for two matrices $\mathbf{C}$ and $\mathbf{D}$, $trace\,(\mathbf{CD}) = trace\,(\mathbf{DC})$ where the matrix products are permitted,

$$\mathbb{E}[\widehat{f}_d(\mathbf{x})] = f_d(\mathbf{x}) + \frac{1}{2} trace[\mathbf{B}' f^{(2)}(\mathbf{x})(\mathbf{B}) \int_{\mathbb{R}^d} \mathbf{uu}' K_d(\mathbf{u})\, d\mathbf{u}] + \cdots \tag{1.224}$$

Let $\mathbf{I}_d$ be the $d \times d$ identity matrix and $\mu_2(K_d)$ is a positive scalar defined earlier such that

$$\int_{\mathbb{R}^d} \mathbf{u}\mathbf{u}' K_d(\mathbf{u})\, d(\mathbf{u}) = \mu_2(K_d)\mathbf{I}_d. \tag{1.225}$$

Then for fixed $\mathbf{x}$, bias is

$$Bias\,[\widehat{f}_d(\mathbf{x})] = \frac{1}{2} trace\,[\mathbf{B}' f^{(2)}(\mathbf{x})\mathbf{B}]\mu_2(K_d) + o(trace\,[\mathbf{B}' f^{(2)}(\mathbf{x})\mathbf{B}]). \tag{1.226}$$

Special cases may be used to illustrate the above formula. For instance, if the bandwidth matrix $\mathbf{B}$ is a diagonal matrix, its diagonal elements being $b_1, b_2, \ldots, b_d$, we get, as $n \to \infty$, if $b_{max} \to 0$, where $b_{max} = max\{b_1, b_2, \ldots, b_d\}$,

$$Bias\,[\widehat{f}_d(\mathbf{x})] = \frac{1}{2}\mu_2(K_d) \sum_{i=1}^{d} b_i^2 \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} + o\left(b_{max}^2\right)$$

When all bandwidths are equal, i.e., $b_i = b$, then if $b \to 0$ as $n \to \infty$, we have

$$Bias\,[\widehat{f}_d(\mathbf{x})] = \frac{b^2}{2}\mu_2(K_d) \sum_{i=1}^{d} \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} + o(b^2).$$

As for the variance,

$$\begin{aligned}
\mathbb{V}ar\,[\widehat{f}_d(\mathbf{x})] &= \frac{1}{n|\mathbf{B}|^2}\mathbb{V}ar\,[K_d(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x}))] \\
&= \int_{\mathbb{R}^d} f_d(\mathbf{X}_i) K_d^2(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x}))\, d\mathbf{X}_i \\
&\quad - \left(\int_{\mathbb{R}^d} f_d(\mathbf{X}_i) K_d(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x}))\, d\mathbf{X}_i\right)^2 \\
&= A_1 + A_2.
\end{aligned} \tag{1.227}$$

As to the first term,

$$\begin{aligned}
A_1 &= \frac{1}{n|\mathbf{B}|^2} \int_{\mathbb{R}^d} f_d(\mathbf{X}_i) K_d^2(\mathbf{B}^{-1}(\mathbf{X}_i - \mathbf{x}))\, d\mathbf{X}_i \\
&= \frac{1}{n|\mathbf{B}|} \int_{\mathbb{R}^d} f_d(\mathbf{x} + \mathbf{B}\mathbf{u}) K_d^2(\mathbf{u})\, d\mathbf{u} \\
&= \frac{1}{n|\mathbf{B}|} f_d(\mathbf{x}) \int_{\mathbb{R}^d} K_d^2(\mathbf{u})\, d\mathbf{u} + o\left(\frac{1}{n|\mathbf{B}|}\right).
\end{aligned} \tag{1.228}$$

On the other hand, from previous calculations, it is easy to see that the second term is

$$A_2 = o\left(\frac{1}{n|\mathbf{B}|}\right). \tag{1.229}$$

Collecting terms, it follows that

$$\mathbb{V}ar\,[\widehat{f}_d(\mathbf{x})] = \frac{1}{n|\mathbf{B}|} f_d(\mathbf{x}) \int_{\mathbb{R}^d} K_d^2(\mathbf{u})\, d\mathbf{u} + o\left(\frac{1}{n|\mathbf{B}|}\right). \tag{1.230}$$

Thus for instance, if $\mathbf{B}$ is a diagonal matrix, with positive diagonal elements $b_1, b_2, \ldots, _d$, then

$$\mathbb{V}ar\,[\widehat{f}_d(\mathbf{x})] = \frac{1}{n\prod_{i=1}^d b_i} f_d(\mathbf{x}) \int_{\mathbb{R}^d} K_d^2(\mathbf{u})\, d\mathbf{u} + o\left(\frac{1}{n\prod_{i=1}^d b_i}\right). \tag{1.231}$$

In the special case, if $b_i = b$ for all i, then

$$\mathbb{V}ar\,[\widehat{f}_d(\mathbf{x})] = \frac{1}{nb^d} f_d(\mathbf{x}) \int_{\mathbb{R}^d} K_d^2(\mathbf{u})\, d\mathbf{u} + o\left(\frac{1}{nb^d}\right). \tag{1.232}$$

An optimal bandwidth selection algorithm can now be developed, as in the univariate case, by minimizing the leading terms in mse or MISE. As for consistency, note, for instance, when $\mathbf{B}$ is a diagonal matrix, if all else remains fixed,

$$b_{max} \to 0 \text{ and } n|\mathbf{B}| = nb_1 b_2 \cdots b_d \to \infty \tag{1.233}$$

as $n \to \infty$, then, for every fixed $\mathbf{x} \in \mathbb{R}^d$, $\widehat{f}_d(\mathbf{x})$ is weakly consistent. For further details see Scott (1992) and Wand and Jones (1995).

2

Nonparametric Regression

2.1 Introduction

Consider the bivariate random variable $(X, Y) \in \mathbb{R}^2$ where X denotes an explanatory variable and Y denotes the response or the dependent variable. Consider the observations

$$(x_i, y_i),\ i = 1, 2, \ldots, n \tag{2.1}$$

on the pair (X, Y). Although in this discussion, we let X be a scalar, the ideas presented here can easily be generalized to the multidimensional case, i.e., when $X \in \mathbb{R}^k$. Similarly, the response variable Y is in $\mathbb{R}$, though multivariate regression is also possible to consider. Moreover, let Y be continuous and in the sequel we will impose some further moment conditions. Nonparametric regression is concerned with the situation when the regression function, i.e., the conditional expected value of Y given X has an arbitrary shape, apart from satisfying some smoothness conditions.

Specifically, our interest lies in estimating the function m, which is the nonparametric regression function

$$m(x) = \mathbb{E}(Y|X = x) \tag{2.2}$$

where $\mathbb{E}$ denotes the conditional expectation of Y given $X = x$. For simplicity of notation, we write $X = x$ even when X may be a continuous random variable. In nonparametric regression, the aim is to estimate the function m, which, apart from some regularity conditions, is left unspecified.

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

We will consider the nonparametric regression model

$$y_i = m(x_i) + u_i,\ i = 1, 2, \ldots, n \tag{2.3}$$

where the regression errors u_i are independent of the design variables x_i and satisfy

$$\mathbb{E}(u_i) = 0 \tag{2.4}$$

$$\mathbb{V}ar(u_i) = \sigma^2,\ 0 < \sigma < \infty \tag{2.5}$$

$$\mathbb{C}ov(u_i, u_j) = 0,\ i \neq j. \tag{2.6}$$

We let m be three times continuously differentiable, its third derivative being bounded, i.e.

$$\sup|d^3 m(x)/dx^3| < \infty. \tag{2.7}$$

2.1.1 Method of least squares

The kernel estimators of $m(x)$ are linear estimators in the sense that they are weight averages of the observations on the response variable. This is also the case for linear models, which are parametric models with the regression function specified as being a linear function of regression coefficients β_0 and β_1:

$$m(x) = \beta_0 + \beta_1 x. \tag{2.8}$$

We take a brief look at this parametric model. An important special case is when X and Y are jointly normally distributed. Let their joint pdf be given by

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right\} \tag{2.9}$$

where $\mu_x, \mu_y \in \mathbb{R}$, $\sigma_x, \sigma_y \in \mathbb{R}_+$ and $\rho \in (-1, 1)$. Then the conditional pdf of Y given X is given by the normal density function

$$h(y|x) = \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} \times exp\left\{-\frac{1}{2\sigma_y^2(1-\rho^2)}\left(y - \mu_y - \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2\right\}, \tag{2.10}$$

the conditional mean of Y given $X = x$ being

$$m(x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x). \tag{2.11}$$

This is the familiar simple linear model, with intercept

$$\beta_0 = \mu_y - \frac{\rho\sigma_y}{\sigma_x}\mu_x \tag{2.12}$$

and slope

$$\beta_1 = \frac{\rho\sigma_y}{\sigma_x}. \tag{2.13}$$

In general, a simple linear model without the distributional assumption is simply

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{2.14}$$

where, given $x_1, \ldots, x_n$ assumed to be not all equal, the errors u_i follow the Gauss–Markov conditions, i.e., they have zero mean, constant variance, and are pairwise uncorrelated.

Of course, if, in addition, the errors are also normally distributed, then the maximum likelihood estimate of $m(x) = \beta_0 + \beta_1 x$ is the same as what one would get by substituting the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. These are obtained from

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\,\{Q(\beta)\} \tag{2.15}$$

where the quadratic form $Q(\beta)$ is simply the error sum of squares

$$Q(\beta) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.16}$$

and

$$\beta = (\beta_0, \beta_1)'. \tag{2.17}$$

The formulas for the least squares estimates of the regression coefficients are obtained by solving the normal equations:

$$\frac{\partial}{\partial\beta_0}Q(\beta)|_{\beta=\widehat{\beta}} = -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \tag{2.18}$$

and

$$\frac{\partial}{\partial \beta_1} Q(\beta)|_{\beta=\widehat{\beta}} = -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0, \tag{2.19}$$

leading to the least squares estimates:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \tag{2.20}$$

and

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = S_{xy}/S_{xx} \tag{2.21}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.22}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \text{ and } S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{2.23}$$

For $i = 1, 2, \ldots, n$, define

$$b_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.24}$$

and

$$a_i = \frac{1}{n} - \bar{x} \cdot b_i = \frac{1}{n} - \frac{\bar{x} \cdot (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2.25}$$

Given the values of the explanatory variable X, namely $x_1, \ldots, x_n$ and the sample size n, the weights a_i and b_i are computable quantities. In particular, a_i and b_i do not depend on the values of the response variable Y. We have

$$\sum_{i=1}^{n} a_i = 1, \ \sum_{i=1}^{n} a_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \ \sum_{i=1}^{n} a_i x_i = 0 \tag{2.26}$$

$$\sum_{i=1}^{n} b_i = 0, \ \sum_{i=1}^{n} b_i^2 = \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \ \sum_{i=1}^{n} b_i x_i = 1 \tag{2.27}$$

$$\sum_{i=1}^{n} a_i b_i = \frac{-\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.28}$$

Consequently,

$$\widehat{\beta}_0 = \sum_{i=1}^{n} a_i y_i = \beta_0 + \sum_{i=1}^{n} a_i u_i \tag{2.29}$$

and

$$\widehat{\beta}_1 = \sum_{i=1}^{n} b_i y_i = \beta_1 + \sum_{i=1}^{n} b_i u_i. \tag{2.30}$$

Finite sample properties of the least squares estimators follow easily. Given $x_1, \ldots, x_n$, they are unbiased since

$$\mathbb{E}\left(\sum_{i=1}^{n} a_i u_i\right) = \mathbb{E}\left(\sum_{i=1}^{n} b_i u_i\right) = 0 \tag{2.31}$$

Also, their variances are

$$\mathbb{V}ar(\widehat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \left\{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right\} \tag{2.32}$$

$$\mathbb{V}ar(\widehat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} b_i^2 = \frac{\sigma^2}{S_{xx}} \tag{2.33}$$

On the other hand, when $n \to \infty$, the least squares estimators become weakly consistent if

$$\sum_{i=1}^{n} a_i^2 \to 0 \text{ and } \sum_{i=1}^{n} b_i^2 \to 0 \text{ as } n \to \infty. \tag{2.34}$$

Moreover, the covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is given by

$$\mathbb{C}ov(\widehat{\beta}_0, \widehat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} a_i b_i = -\sigma^2 \frac{\bar{x}}{S_{xx}} \tag{2.35}$$

so that these estimators are asymptotically uncorrelated if

$$\sum_{i=1}^{n} a_i b_i \to 0 \text{ as } n \to \infty. \tag{2.36}$$

One can also express the above formulas using vector notation. Using standard terminology, the design matrix is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} \tag{2.37}$$

which has full rank since not all $x_1, \ldots, x_n$ are equal, so that in particular $\mathbf{X'X}$ is positive definite, i.e., has non-zero (positive) eigenvalues and the solution to (2.15) is unique. Moreover, the total variance of the estimators is

$$\sigma^2 \left(\sum_{i=1}^{n} a_i^2 + \sum_{i1}^{n} b_i^2 \right) = \sigma^2 tr(\mathbf{X'X})^{-1} = O(1/\lambda_{min}(\mathbf{X'X})). \tag{2.38}$$

Here $tr(\mathbf{A})$ denotes the trace of a matrix $\mathbf{A}$ and $\lambda_{min}(\mathbf{X'X})$ is the smallest eigenvalue of $\mathbf{X'X}$. In other words, due to Chebyshev's lemma, if

$$\lambda_{min}(\mathbf{X'X}) \to \infty \tag{2.39}$$

with $n \to \infty$, the least squares estimators are weakly consistent. For additional information see, among others, Drygas (1976), Eicker (1963), Lai et al. (1979), Rao (1973), and Sen and Srivastava (1990). For results under infinite variance, see, for example, Cline (1989).

If the errors in (2.14) are normally distributed, i.e., $u_i \sim iid\, N(0, \sigma^2)$, then being linear combinations of independent normal variables (see (2.29) and (2.30)), the estimated regression coefficients also have normal distributions. Specifically, given $x_1, \ldots, x_n$,

$$\widehat{\beta}_i \sim N(E\{\widehat{\beta}_i\}, \mathbb{V}ar(\widehat{\beta}_i)),\ i = 1, 2. \tag{2.40}$$

Once standardized, we have

$$(\widehat{\beta}_0 - \mathbb{E}(\widehat{\beta}_0))/\sqrt{\mathbb{V}ar(\widehat{\beta}_0)} = (\widehat{\beta}_0 - \beta_0)/\sqrt{\sigma^2(1/n + \bar{x}^2/S_{xx})} \sim N(0, 1) \tag{2.41}$$

and

$$(\widehat{\beta}_1 - \mathbb{E}(\widehat{\beta}_1))/\sqrt{\mathbb{V}ar(\widehat{\beta}_1)} = (\widehat{\beta}_1 - \beta_1)/\sqrt{\sigma^2/S_{xx}} \sim N(0, 1). \tag{2.42}$$

In addition, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ also have a joint bivariate normal distribution, the mean vector and the covariance matrix of this distribution being specified by the moments of the regression estimators given above. Similarly, the fitted curve

$$\widehat{m}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x \tag{2.43}$$

will also have a normal distribution. This fact is used to derive confidence intervals for the fitted curve. If the regression errors do not have a normal distribution, but $n \to \infty$, then the marginal distributions, the joint distribution, and all linear combinations of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ will be asymptotically normal.

As for an estimate of the error variance, it turns out that

$$\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} \{y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\}^2 \tag{2.44}$$

is an unbiased estimator of σ^2. This is easy to see since

$$\widehat{\sigma^2} = \frac{1}{n-2} \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right].$$

However,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \beta_1^2 S_{xx} + \sum_{i=1}^{n} (u_i - \bar{u})^2 + 2\beta_1 \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u})$$

so that

$$E\left[\sum_{i=1}^{n} (y_i - \bar{y})^2 \right] = \beta_1^2 S_{xx} + (n-1)\sigma^2.$$

On the other hand,

$$E\left[\widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] = \left\{ \mathbb{V}ar(\widehat{\beta}_1) + \beta_1^2 \right\} S_{xx} = \left\{ \sigma^2 / S_{xx} + \beta_1^2 \right\} S_{xx}$$
$$= \sigma^2 + \beta_1^2 S_{xx}.$$

Now, collecting terms, the result follows.

Adding distributional assumptions, when the error terms in the linear regression model are iid $N(0, \sigma^2)$), then

$$(\widehat{\beta}_0 - \beta_0) / \sqrt{\widehat{\sigma^2}(1/n + \bar{x}^2 / S_{xx})} \sim t_{n-2} \tag{2.45}$$

and

$$(\widehat{\beta}_1 - \beta_1)/\sqrt{\widehat{\sigma^2}/S_{xx}} \sim t_{n-2}. \tag{2.46}$$

Alternatively, in large samples, i.e., as $n \to \infty$, $\widehat{\sigma^2}$ is also consistent and this leads to the asymptotic normality of the least squares estimators:

$$(\widehat{\beta}_0 - \beta_0)/\sqrt{\widehat{\sigma^2}(1/n + \bar{x}^2/S_{xx})} \sim N(0,1) \tag{2.47}$$

and

$$(\widehat{\beta}_1 - \beta_1)/\sqrt{\widehat{\sigma^2}/S_{xx}} \sim N(0,1). \tag{2.48}$$

These facts can be used for testing and confidence intervals for the regression coefficients and, more generally, of the linear regression function. For instance, in large samples, an approximate $100(1-\alpha)\%$ confidence interval for β_1 can be given by

$$\widehat{\beta}_1 \pm z_{\alpha/2}\sqrt{\widehat{\sigma^2}/S_{xx}}, \tag{2.49}$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$-point of the standard normal distribution, i.e., $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the $N(0,1)$ distribution, and $0 < \alpha < 1$.

In a multiple regression model, however, the explanatory variable X is a vector in R^k, $k \geq 1$. Consider $k = p-1$ $(p > 1)$ explanatory variables $X^{(1)}, X^{(2)}, \ldots, X^{(p-1)}$ and the problem is to estimate the regression function m, where

$$\begin{aligned} m(x^{(1)}, x^{(2)}, \ldots, x^{(p-1)}) &= E(Y|X^{(1)} = x^{(1)}, X^{(2)} \\ &= x^{(2)}, \ldots, X^{(p-1)} = x^{(p-1)}) \\ &= \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} \ldots \beta_{p-1} x^{(p-1)}. \end{aligned}$$

The multiple regression model includes many examples. One specific case is polynomial regression, where considering X to be an explanatory variable of interest and defining $X^{(j)} = X^j$, the regression of Y on $X = x$ is

$$m(x) = E(Y|X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 \ldots \beta_{p-1} x^{p-1}, \tag{2.50}$$

which is a polynomial of degree $p-1$ in x. We come back to this special case again, in the context of *local polynomials*.

Thus let $\mathbf{y}$ = the vector of observations on the response variable Y, $\mathbf{X}$ = the design matrix containing values of the $p-1$ explanatory variables, β = the vector of p regression coefficients that are to be estimated, and $\mathbf{u}$ = the vector of errors. Thus,

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)',$$
$$\mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(p-1)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(p-1)} \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(p-1)} \end{pmatrix},$$
$$\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})' \text{ and}$$
$$\mathbf{u} = (u_1, u_2, \ldots, u_n)'.$$

The multiple regression model is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{2.51}$$

where the errors follow the Gauss–Markov conditions, i.e.,

$$E(\mathbf{u}) = \mathbf{0},\ \mathbb{C}ov(\mathbf{u}) = \sigma^2 \mathbf{I}_{n\times n},$$

where $\mathbf{0} = (0, 0, \ldots, 0)'$ is a vector of length n and $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix. In addition, we assume that the design matrix $\mathbf{X}$ (with n rows and p columns) has full rank, so that

$$rank(\mathbf{X}) = p,\ p < n.$$

As in the case of the simple linear model, the least squares estimator $\widehat{\beta}$ is defined as

$$\widehat{\beta} = \underset{\beta}{\text{argmin}}\ Q(\beta), \tag{2.52}$$

where the error sum of squares

$$Q(\beta) = \sum_{i=1}^{n} u_i^2 = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{2.53}$$

is convex in β, because the second derivative of $Q(\beta)$

$$\frac{\partial^2}{\partial\beta\partial\beta} Q(\beta) = 2[\mathbf{X}'\mathbf{X}] \tag{2.54}$$

is positive definite, so that the solution to (2.52) is unique and can be obtained from solving the normal equations

$$\frac{\partial}{\partial \beta} Q(\beta)|_{\beta=\widehat{\beta}} = 0. \tag{2.55}$$

In particular,

$$\widehat{\beta} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{y} = \beta + [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{u}, \tag{2.56}$$

so that $\widehat{\beta}$ is unbiased and

$$\mathbb{C}ov(\widehat{\beta}) = \sigma^2[\mathbf{X}'\mathbf{X}]^{-1}. \tag{2.57}$$

As for large sample properties, as in the case of simple linear regression,

$$\sum_{i=0}^{p-1} \mathbb{V}ar(\widehat{\beta}_i) = \sigma^2 tr([\mathbf{X}'\mathbf{X}]^{-1}) \tag{2.58}$$

so that $\widehat{\beta}_i,\ i = 0, 1, \ldots, p-1$, is weakly consistent if $\lambda_{min}(\mathbf{X}'\mathbf{X}) \to \infty$ as $n \to \infty$, where $\lambda_{min}(\mathbf{X}'\mathbf{X})$ is the minimum eigenvalue of $\mathbf{X}'\mathbf{X}$.

Substituting, we have the fitted hyperplane as

$$\widehat{E(\mathbf{y})} = \mathbf{X}\widehat{\beta} = \mathbf{H}\mathbf{y} \tag{2.59}$$

where $\mathbf{H}$ is the so-called Hat-matrix (an $n \times n$ matrix), namely

$$\mathbf{H} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'. \tag{2.60}$$

It is easy to establish that the Hat-matrix is (i) symmetric, i.e., $\mathbf{H}' = \mathbf{H}$, and (ii) idempotent, i.e. $\mathbf{H}^2 = \mathbf{H}\mathbf{H} = \mathbf{H}$. Also,

$$\begin{aligned} tr(\mathbf{H}) &= tr(\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}') = tr([\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{X}) \\ &= tr(\mathbf{I}_{p\times p}) = p. \end{aligned} \tag{2.61}$$

These facts have nice consequences, and lead to an unbiased estimator for the error variance. Define the residuals

$$\widehat{\mathbf{u}} = \mathbf{y} - E(\mathbf{y}) = \mathbf{y} - \mathbf{X}\widehat{\beta} = \mathbf{M}\mathbf{y} \tag{2.62}$$

where

$$\mathbf{M} = \mathbf{I}_{n\times n} - \mathbf{H}. \tag{2.63}$$

In particular, then, $\mathbf{M}$ is symmetric, idempotent, and $tr(\mathbf{M}) = n - p$, so that

$$\widehat{\sigma}^2 = \frac{1}{n-p}\widehat{\mathbf{u}}'\widehat{\mathbf{u}} \tag{2.64}$$

is an unbiased estimator of σ^2. This is easy to observe since $\mathbf{MX} = \mathbf{0}_{n\times n}$, so that

$$\widehat{\mathbf{u}} = \mathbf{My} = \mathbf{Mu}. \tag{2.65}$$

Therefore,

$$(n-p)E(\widehat{\sigma}^2) = E[(\mathbf{Mu})'\mathbf{Mu}] = E(\mathbf{u}'\mathbf{M}'\mathbf{Mu}) = E(\mathbf{u}'\mathbf{Mu}). \tag{2.66}$$

However, being a scalar, $E(\mathbf{u}'\mathbf{Mu}) = E(tr(\mathbf{u}'\mathbf{Mu})$, whereas $tr(\mathbf{u}'\mathbf{Mu}) = tr(\mathbf{Muu}')$ so that $E[tr(\mathbf{Muu}')] = tr[\mathbf{M}E(\mathbf{uu}')] = tr[\mathbf{M}\sigma^2 I_{n\times n}] = tr[\mathbf{M}\sigma^2] = (n-p)\sigma^2$.

When the regression errors are iid normal, i.e., if $u_i \sim iidN(0, \sigma^2)$, then consistency of $\widehat{\sigma}^2$ is easy to prove. This follows by noting that, since $\mathbf{M}$ has trace $n - p$ and M is idempotent, its rank is equal to $n - p$ as well, in which case (Rao 1973, p. 186),

$$\frac{1}{\sigma^2}\mathbf{u}'\mathbf{Mu} \sim \chi^2_{n-p} \tag{2.67}$$

implying

$$\mathbb{E}\left(\frac{\mathbf{u}'\mathbf{Mu}}{n-p}\right) = \frac{\sigma^2(n-p)}{n-p} = \sigma^2 \tag{2.68}$$

$$\mathbb{V}ar\left(\frac{\mathbf{u}'\mathbf{Mu}}{n-p}\right) = \frac{\sigma^4(n-p)}{(n-p)^2} \to 0, \text{ as } n \to \infty \tag{2.69}$$

so that weak-consistency follows from Chebyshev's inequality. When the errors are not normally distributed, let

$$\frac{\sum_{i=1}^n u_i^2}{n} \to \sigma^2 \text{ in probability, as } n \to \infty \tag{2.70}$$

hold. Then weak-consistency of $\widehat{\sigma}^2$ can be established by noting that (see Sen and Srivastava 1990, p. 47)

$$\widehat{\sigma}^2 = \frac{\mathbf{u}'\mathbf{Mu}}{n-p} = \frac{\mathbf{u}'(\mathbf{I}-\mathbf{H})\mathbf{u}}{n-p} \tag{2.71}$$

where by Markov's inequality, for $\delta > 0$,

$$P\left(\frac{\mathbf{u}'\mathbf{H}\mathbf{u}}{n-p} > \delta\right) \leq \frac{\mathbb{E}(\mathbf{u}'\mathbf{H}\mathbf{u})}{(n-p)\delta} = \frac{p\sigma^2}{(n-p)\delta} \to 0, \text{ as } n \to \infty. \tag{2.72}$$

For an extensive coverage of the theory of least squares estimation, see in particular Rao (1973). What we have considered above is the theory of the OLS (*ordinary least squares*) estimators, $\widehat{\beta}_{ols}$. In this case, the Gauss–Markov theorem ensures that $\widehat{\beta}_{ols}$ is BLUE (best linear unbiased estimator). When the errors are not uncorrelated, but have a covariance matrix, say Σ, in finite samples, the best linear unbiased estimator is not the OLS estimator but one that is obtained by pre-multiplying both sides of Equation (2.51) by the inverse of the square root of Σ. In particular, this leads to the WLS (*weighted least squares*) estimator, $\widehat{\beta}_{wls}$. Due to the Gauss–Markov theorem, in finite samples, unless Σ is a diagonal matrix, for $\mathbf{a} \in \mathbb{R}^p$, $\mathbf{a}'\widehat{\beta}_{wls}$ has a smaller variance than $\mathbf{a}'\widehat{\beta}_{ols}$, both being unbiased estimators of $\mathbf{a}'\beta$, so that the WLS is BLUE.

Some authors have studied efficiency of least squares estimators with time series data where the errors are no longer uncorrelated. Under stationarity and further suitable conditions, the OLS estimator turns out to be asymptotically efficient. This means, at least in these situations, as far as asymptotic efficiency is concerned, knowledge of the error covariances is not necessary; see Grenander (1954) and Yajima (1991) for further information. For related results under various types of correlations, in particular long-memory, see Beran et al. (2013).

2.1.2 Influential observations

Let h_{ii} be the ith diagonal element of $\mathbf{H}$. Then the following facts may be noted.

First of all, $h_{ii} \geq 0$. This is so because being the ith diagonal element of $\mathbf{H}$,

$$h_{ii} = \mathbf{x}_\mathbf{i}'[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{x}_\mathbf{i} \geq 0 \tag{2.73}$$

where $\mathbf{x}_\mathbf{i}'$ is the ith row of the design matrix $\mathbf{X}$.

In addition, $h_{ii} \leq 1$. This follows by noting that

$$\mathbb{V}ar(\hat{\mathbf{u}}) = \mathbf{M}\mathbb{C}ov(\mathbf{y})\,\mathbf{M}' = \sigma^2\mathbf{M}^2 = \sigma^2\mathbf{M}. \tag{2.74}$$

This means $\mathbb{V}ar(\widehat{u}_i)$ is the ith diagonal element of $\sigma^2\mathbf{M} = \sigma^2(\mathbf{I}_{n\times n} - \mathbf{H})$, so that $\mathbb{V}ar(\widehat{u}_i) = \sigma^2(1 - h_{ii})$. Since $\mathbb{V}ar(\widehat{u}_i)$ must be non-negative, we have $h_{ii} \leq 1$.

A high h_{ii} value, i.e., $h_{ii} \approx 1$, implies $\mathbb{V}ar(\widehat{u}_i) \approx 0$. On the other hand, $\mathbb{E}(\widehat{u}_i) = 0$. High h_{ii} would thus typically indicate a small $\widehat{u}_i$, i.e., a very good fit, and h_{ii} is termed the leverage. In some cases, however, an observation that is far from the majority in the design space, may result in a high leverage as well. In other words, high leverage need not necessarily imply a cause for concern; however, they are to be examined prior to further analysis.

The residuals $(\widehat{u}_i)$ as well as the leverages (h_{ii}) are therefore examined as part of routine regression diagnostics. High values of either of these quantities are worth investigating for judging the overall quality of the fit. An idea for a combined test is in Cook's distance (see Cook 1977, 1979). This quantity is computed for each observation number i as follows:

$$D_i = \frac{h_{ii}}{1 - h_{ii}} \frac{r_i^2}{p} \tag{2.75}$$

where r_i is the standardized residual

$$r_i = \frac{\widehat{u}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}} \tag{2.76}$$

and $\widehat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n} \widehat{u}_i^2$.

This quantity can be shown to be related to a test for comparing two estimates of the β vector, where in one case all observations are used and in the other case observation number i is deleted from the estimation procedure. In particular, a high value of D_i may indicate either an outlier or an observation with a high leverage and is worth a careful consideration.

2.1.3 Nonparametric regression estimators

In what follows, we relax the parametric assumptions on the regression function m and address a kernel based approach for

estimation of this function. Our interest is estimation of $m(x)$ at an arbitrary point x in the nonparametric regression model (2.3).

As in the case of kernel density estimation, here also the curve estimate is a convolution of a smooth function called the kernel K and a non-smooth stochastic component, namely the random observations.

Choice of the kernel affects smoothness of the resulting estimate because the smoothness properties of the kernel are transferred into the convolution. A second parameter that affects the quality of the estimator is the bandwidth. In particular there is a trade-off, i.e., a very large or a very small bandwidth results in sub-optimal estimators.

Optimality of the estimator can be defined in various different ways and one option is to consider the mean squared error or the quadratic loss function, which is related to convergence in probability of the estimate $\hat{m}(x)$ to the true but unknown $m(x)$ via Chebyshev's inequality. For L_1 norm based results, see Devroye (1987), though in the context of density estimation.

While addressing nonparametric regression, two cases are of typical interest:

(a) The fixed-design case, where the pairs (x_i, y_i) are observed at fixed values of $x_1, \ldots, x_n$ on a compact interval. These values of the explanatory variable are then treated as being non-random. In particular, the x_i values may be equidistant.
(b) The random-design case, where the iid pairs (x_i, y_i) are observed; i.e., $x_1, \ldots, x_n$ are random.

We focus on a selection of nonparametric regression estimators, namely, the Priestley–Chao regression estimator, the Nadaraya–Watson regression estimator, and the local polynomials regression estimator, and provide a brief description of the method of smoothing splines; another related approach is in (2.188), which we will also discuss briefly here. There is an extensive literature on nonparametric regression. In addition to the references appearing elsewhere in this book, also see Cheng and Lin (1981a, 1981b), Cheng et al. (1997), Clark (1977), Cleveland and Devlin (1988), Collomb (1981, 1985a, 1985b), Greblicki and Krzyzak (1980), Härdle and Marron (1985), Hastie and Loader (1993), and others. In particular, the textbooks by Wand and

Jones (1995), Bowman and Azzalini (1997), etc., contain additional information.

Generally speaking, the Priestley–Chao regression estimator is asymptotically unbiased for the fixed design case. However, in the case of a random design, the Priestley–Chao regression estimator converges in probability to the product $m(x)f(x)$, where $f(x)$ is the design density, i.e. it is the probability density function of the explanatory variable X and x is a fixed real number.

An obvious correction for the random design case is dividing the Priestley–Chao regression estimator by a consistent estimator of the design density function $f(x)$, for $f(x) > 0$. The resulting estimator is the Nadaraya–Watson regression estimator. As it turns out, however, the Nadaraya–Watson estimator is a special case of the class of local-polynomial regression estimators of degree $p = 0, 1, 2, \ldots$, the Nadaraya–Watson regression estimator being the so-called local-constant estimator ($p = 0$).

In the method based on local polynomials, the choice of p plays an important role, in particular in connection with boundary bias, and one may achieve improved (asymptotic) properties of the regression estimator near the boundaries by selecting higher order (local) polynomials.

We mention some nonparametric regression curve estimators and refer the reader to the cited references for further results. Additional results are presented in Chapters 3 to 5 where we deal with correlated observations.

2.2 Priestley–Chao regression estimator

Consider the fixed design case with observations (x_i, y_i), $i = 1, 2, \ldots, n$, on the explanatory variable X and the response variable Y, the x_i values being fixed and evenly spaced and in particular, $x_i = i/n$. Let the nonparametric regression model (2.3) hold.

Suppose that after plotting the data in a scatterplot, a nonlinear association emerges. This conditional mean relation between $X = x_i$ and Y is the expected value $m(x_i)$, and its estimation at some arbitrary point $x \in (0, 1)$ is of interest.

A kernel regression estimator is a weighted (local) average of the values of the response variable. The weights are chosen based

on a kernel and a bandwidth, both of which play important roles in the estimation process.

Typically, kernel estimates are not unbiased. Specific conditions are thus called for to ensure consistency. Conditions needed to achieve pointwise weak consistency are stated below. Additional conditions may be specified as needed, e.g., to achieve weak uniform consistency or other types of asymptotic properties.

The Priestley–Chao kernel regression estimator (Priestley and Chao 1972) of $m(x)$, $0 < x < 1$, is given by

$$\widehat{m}_{PC}(x) = \frac{1}{nb} \sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{b}\right) \tag{2.77}$$

where the bandwidth $b > 0$ and the kernel K satisfiy the following conditions:

(a) as $n \to \infty$, $b \to 0$ and $nb \to \infty$ and
(b) the kernel K is a continuous function such that $K(u) \geq 0$ for all $u \in \mathbb{R}$ and $K(u) = 0$ for all u such that $|u| > 1$; $\int_{-1}^{1} K(u)du = 1$ and $K(u) = K(-u)$ for all $u \in \mathbb{R}$.

A kernel such as the ones mentioned above falls in the category of kernels of *order 2.* This terminology is used in particular in the context of the so-called *higher order kernels* (see Gasser and Müller 1984), which are useful for estimating derivatives of the regression function. Derivative estimation is addressed in Chapter 3 on Trend Estimation using time series data and in the context of local polynomials later in this chapter.

Thus $\widehat{m}_{PC}(x)$ is a weighted average

$$\widehat{m}_{PC}(x) = \frac{1}{n} \sum_{i=1}^{n} y_i w_i(x) = \frac{1}{n} \sum_{i=1}^{n} m(x_i) w_i(x) + \frac{1}{n} \sum_{i=1}^{n} u_i w_i(x) \tag{2.78}$$

where the weights

$$w_i(x) = \frac{1}{b} K\left(\frac{x_i - x}{b}\right) \tag{2.79}$$

are chosen to satisfy some conditions as indicated via the assumptions on the bandwidth b and the kernel function K.

Note that while $K(u)$ has its support on $(-1, 1)$, $w_i(x)$ has its support on $(x_i - b, x_i + b)$. Alternatively, the y_i for which the corresponding x_i does not fall in the interval $(x - b, x + b)$ gets zero weight. In other words, the weighted average (2.77) has a local characteristic and, namely, $m(x)$ is estimated by taking the average of y_i values that have x_i values within b distance from x.

The choice of this bandwidth b thus becomes relevant. We need this bandwidth to be small so that the local properties of the mean function can be retained. On the other hand, when b is too small, the observation window does not contain many data points, so that the variance of the estimator gets inflated and the curve estimate becomes less smooth. In an extreme case, when b is near zero, the estimated curve will "follow the data", driven by randomness. As a result, we learn very little from our estimation procedure, since the statistical summary becomes inadequate.

The role of the bandwidth b can be assessed in a concrete manner, by analyzing the asymptotic properties of the estimator. For instance, it is easy to see that the expected value of the estimator is the same weighted average of the values $m(x_i)$ and thus need not equal $m(x)$ in finite samples. However, asymptotic unbiasedness can be proved so that $\sum_{i=1}^{n} m(x_i)w_i(x)/n$ approximately equals $m(x)$ with increasing sample size. Similarly, the variance of the estimator can be expressed as a function of b. This leads to ideas for optimal bandwidth selection and various data-driven algorithms.

First of all, note that as $n \to \infty$, for an integer $q = 0, 1, 2, \ldots$,

$$\frac{1}{nb} \sum_{i=1}^{n} \left(\frac{x_i - x}{b} \right)^q K\left(\frac{x_i - x}{b} \right) = \int_{-1}^{1} u^q K(u) du + O\left(\frac{1}{nb} \right). \tag{2.80}$$

Due to the differentiability condition on m, by Taylor series expansion,

$$m(x_i) = m(x) + (x_i - x)m^{(1)}(x) + \frac{(x_i - x)^2}{2!} m^{(2)}(x) + O(|x_i - x|^3). \tag{2.81}$$

Taking expectation and due to the assumptions on the kernel,

$$\begin{aligned}\mathbb{E}(\widehat{m}_{PC}(x)) &= \frac{1}{nb}\sum_{i=1}^{n} m(x_i)K\left(\frac{x_i - x}{b}\right) \\ &= \frac{1}{nb}\sum_{i=1}^{n}\left[m(x) + (x_i - x)m^{(1)}(x) + \frac{(x_i - x)^2}{2!}m^{(2)}(x)\right. \\ &\quad \left. + O(|x_i - x|^3)\right]K\left(\frac{x_i - x}{b}\right) \\ &= m(x) + \frac{b^2}{2!}m^{(2)}(x)\int_{-1}^{1} u^2 K(u)du + O\left(\frac{1}{nb}\right) + o(b^2)\end{aligned} \tag{2.82}$$

which leads to the asymptotic expression for the bias:

As $n \to \infty$, and under the conditions on b and K specified above,

$$\begin{aligned}Bias(\widehat{m}_{PC}(x)) &= \frac{b^2}{2}m^{(2)}(x)\int_{-1}^{1} u^2 K(u)du + o(b^2) \\ &\quad + O\left(\frac{1}{nb}\right).\end{aligned} \tag{2.83}$$

In addition to the conditions on the bandwidth b mentioned above, if $nb^3 \to \infty$ as $n \to \infty$, then

$$o(b^2) = O\left(\frac{1}{nb}\right), \tag{2.84}$$

so that the bias term reduces to

$$\frac{m^{(2)}(x)}{2}b^2\int_{-1}^{1} u^2 K(u)du + o(b^2). \tag{2.85}$$

As it turns out, the optimum bandwidth that minimizes the leading term in the asymptotic expression of the mean squared error satisfies this condition.

The asymptotic expression for the variance follows exactly along the same lines, by noting that

$$\mathbb{V}ar(\widehat{m}_{PC}(x)) = \frac{1}{n^2 b^2}\sum_{i=1}^{n}\left[K\left(\frac{x_i - x}{b}\right)\right]^2 \cdot \mathbb{V}ar(y_i) \tag{2.86}$$

which simplifies to

$$\mathbb{V}ar(\widehat{m}_{PC}(x)) = \frac{\sigma^2}{nb}\int_{-1}^{1} K^2(u)du + o\left(\frac{1}{nb}\right) \tag{2.87}$$

where we use the fact that as $n \to \infty$,

$$\frac{1}{nb}\sum_{i=1}^{n} K^2\left(\frac{x_i - x}{b}\right) = \int_{-1}^{1} K^2(u)du + o\left(\frac{1}{nb}\right). \tag{2.88}$$

2.2.1 Weak consistency

The above discussion shows that for every fixed x, both bias and variance of the Priestley–Chao regression estimator converge to zero, implying convergence of the mean squared error to zero.

Due to Chebyshev's inequality, this in turn implies pointwise weak consistency, i.e., for fixed x, for every $\epsilon > 0$, as $n \to \infty$,

$$P(|\widehat{m}_{PC}(x) - m(x)| > \epsilon) = 0. \tag{2.89}$$

In some applications, we may require uniform consistency, as, for instance, when estimating functionals of the regression function is of interest. This problem has been addressed among others by Nadaraya (1964), Devroye (1978), Schuster and Yakowitz (1979), Mack and Silverman (1982), and Bierens (1983, 1987); also see Ghosh (2014). Some of these authors follow up on the idea of a characteristic function based approach due to Parzen (1962), which we describe here.

Thus we are concerned with the property

$$\lim_{n\to\infty} P(\sup_x |\widehat{m}_{PC}(x) - m(x)| > \epsilon) = 0 \tag{2.90}$$

for every $\epsilon > 0$, and we would like to investigate at which rate the bandwidth b needs to converge to zero with increasing sample size, so that the above holds.

As in Chapter 1, consider a kernel K, which, in addition to the previously mentioned conditions, also satisfies the following:

Let K have a characteristic function ψ_K, i.e.,

$$\psi_K(t) = \int_{-\infty}^{\infty} e^{\imath tu} K(u)du, \ \sqrt{\imath} = -1, \ t \in \mathbb{R}. \tag{2.91}$$

Now suppose that ψ_K is absolutely integrable, i.e.,

$$\int_{-\infty}^{\infty} |\psi_K(t)|dt < \infty. \tag{2.92}$$

It is interesting to note that the uniform distribution does not have an absolutely integrable characteristic function. Density functions with absolutely integrable characteristic functions include the normal as well as Cauchy.

Due to the inversion theorem for characteristic functions, we can write

$$K(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\imath tu} \psi_K(t)dt. \tag{2.93}$$

Substitution yields

$$\widehat{m}_{PC}(x) = \frac{1}{2\pi} \frac{1}{nb} \sum_{j=1}^{n} \left\{ \int_{-\infty}^{\infty} e^{-\imath t(x_j - x)/b} \psi_K(t)dt \right\} y_j. \tag{2.94}$$

However, recalling the nonparametric regression model $y_j = m(x_j) + u_j$, where the u_j are zero mean and constant variance (σ^2) errors, and also due to the bias

$$\begin{aligned} \widehat{m}_{PC}(x) = m(x) + \frac{b^2}{2} m^{(2)}(x)\mu_2(K)o(b^2) + O\left(\frac{1}{nb}\right) \\ + \frac{1}{2\pi} \frac{1}{nb} \sum_{j=1}^{n} \left\{ \int_{-\infty}^{\infty} e^{-\imath t(x_j - x)/b} \psi_K(t)dt \right\} u_j. \end{aligned} \tag{2.95}$$

Recalling the assumption that m is three times continuously differentiable with finite derivatives, it is enough to show that, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(\sup_x |S_n(x)| > \epsilon) = 0 \tag{2.96}$$

where

$$\begin{aligned} S_n(x) &= \frac{1}{2\pi} \frac{1}{nb} \sum_{j=1}^{n} u_j \left\{ \int_{-\infty}^{\infty} e^{-\imath t(x_j - x)/b} \psi_K(t)dt \right\} \\ &= \frac{1}{2\pi} \frac{1}{b} \int_{-\infty}^{\infty} \psi_K(t) e^{\imath xt/b} \sum_{j=1}^{n} u_j e^{-\imath tx_j/b} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_K(bw) e^{\imath xw} \sum_{j=1}^{n} u_j e^{-\imath x_j w} dw, \end{aligned} \tag{2.97}$$

so that

$$\mathbb{E}(\sup_x |S_n(x)|) \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\psi_K(bw)| \cdot \mathbb{E}\left(\left| \sum_{j=1}^{n} u_j e^{-\iota x_j w} \right| \right) dw. \tag{2.98}$$

However,

$$\begin{aligned}
\mathbb{E}\left| \sum_{j=1}^{n} u_j e^{-\iota x_j w} \right| &= \mathbb{E}\left| \sum_{j=1}^{n} u_j \cos(x_j w) - \iota \sum_{j=1}^{n} u_j \sin(x_j w) \right| \\
&= \left\{ \mathbb{E}\left[\sum_{j=1}^{n} u_j \cos(x_j w) \right]^2 + \mathbb{E}\left[\sum_{j=1}^{n} u_j \sin(x_j w) \right]^2 \right\}^{1/2} \\
&= \left\{ \mathbb{V}ar\left[\sum_{j=1}^{n} u_j \cos(x_j w) \right] + \mathbb{V}ar\left[\sum_{j=1}^{n} u_j \sin(x_j w) \right] \right\}^{1/2} \\
&= \left\{ \frac{\sigma^2}{n^2} \sum_{j=1}^{n} (\cos^2(x_j w) + \sin^2(x_j w)) \right\}^{1/2} \\
&= \frac{\sigma}{\sqrt{n}}.
\end{aligned} \tag{2.99}$$

Moreover,

$$\begin{aligned}
\int_{-\infty}^{\infty} |\psi_K(bw)| dw &= \frac{1}{b} \int_{-\infty}^{\infty} |\psi_K(u)| du \\
&= O\left(\frac{1}{b} \right)
\end{aligned} \tag{2.100}$$

so that

$$\mathbb{E}(\sup_x |S_n(x)|) = O\left(\frac{1}{\sqrt{n}b} \right) \tag{2.101}$$

In other words, if

$$nb^2 \to \infty, \text{ as } n \to \infty \tag{2.102}$$

then

$$\mathbb{E}(\sup_x |S_n(x)|) \to 0 \tag{2.103}$$

and due to Markov's inequality, this implies that (2.96) holds for every $\epsilon > 0$. In other words, due to (2.95), $\widehat{m}_{PC}$ is uniformly consistent in probability as $n \to \infty$.

2.3 Local polynomials

Given iid pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$ on an explanatory variable X, and a response variable Y, we consider the nonparametric regression model

$$y_i = m(x_i) + u_i \tag{2.104}$$

with $\mathbb{E}(y_i|x_i) = m(x_i)$ and m is a smooth real-valued function. Also, $u_1, u_2, \ldots, u_n$ are independently distributed random variables with $\mathbb{E}(u_i|x_i) = 0$, $\mathbb{V}ar(u_i|x_i) = \sigma^2(x_i)$.

Often the explanatory variable will be in $\mathbb{R}^k$, but here we describe only the $k = 1$ case. For further generalization to $k > 1$ and other information, see in particular Fan and Gijbels (1996) and Fan et al. (1997) and references therein; also see, for example, Bickel and Li (2007), Breidt and Opsomer (2000), Opsomer and Ruppert (1997), Hastie and Loader (1993), Hastie and Tibshirani (1990), and others.

The main idea behind local polynomial smoothing is nonparametric estimation of m at $X = x$ using a (local) polynomial approximation of the function $m(x)$ in a small neighborhood of x. This leads to a local least squares solution with various advantages.

We assume that the regression function m is continuous. It is continuously differentiable $p + 1$ times ($p = 0, 1, 2, \ldots$), with finite derivatives. Using the Taylor series expansion in a small neighborhood around x, a pth-degree polynomial approximation of $m(x_i)$ is

$$m(x_i) = \sum_{j=0}^{p} (x_i - x)^j \beta_j(x) + O(|x_i - x|^{p+1}), \tag{2.105}$$

where

$$\beta_j(x) = \frac{m^{(j)}(x)}{j!},\ j = 0, 1, 2, \ldots, p, \tag{2.106}$$

where $\beta_0(x) = m(x)$. The problem of local polynomial estimation then reduces to estimation of the "local regression coefficients" $\beta_j(x)$, which are now functions, by minimizing the weighted error sum of squares

$$Q(x) = \sum_{i=1}^{n} \left\{ y_i - \sum_{j=0}^{p} (x_i - x)^j \beta_j(x) \right\}^2 w_i(x) \tag{2.107}$$

with respect to the vector $\beta \in \mathbb{R}^{p+1}$ where

$$w_i(x) = \frac{1}{b} K\left(\frac{x_i - x}{b}\right) \tag{2.108}$$

where K is a kernel and b is a bandwidth such that in particular, for $u \in \mathbb{R}$,

$$\begin{aligned} &K(u) \geq 0,\ K(u) = K(-u) \\ &\int_{-\infty}^{\infty} K(u)du = 1 \\ &\int_{-\infty}^{\infty} |u|^{p+1} K(u)du < \infty. \end{aligned} \tag{2.109}$$

As for the bandwidth b, as $n \to \infty$, $b \to 0$ and $nb \to \infty$, and more conditions may be added as needed.

Since, $\beta_0(x) = m(x)$ and $j!\beta_j(x) = m^{(j)}(x)$, estimation of the regression coefficients (functions) $\beta_j(x)$ automatically leads to estimation of the regression function $m(x)$ and its derivatives. We introduce new notation:

$$\begin{aligned} \mathbf{y} &= (y_1, y_2, \ldots, y_n)', \\ \mathbf{X}(x) &= \begin{pmatrix} 1 & x_1 - x & (x_1 - x)^2 & \ldots & (x_1 - x)^p \\ 1 & x_2 - x & (x_2 - x)^2 & \ldots & (x_2 - x)^p \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_n - x & (x_n - x)^2 & \ldots & (x_n - x)^p \end{pmatrix}, \\ \beta(x) &= (\beta_0(x), \beta_1(x), \beta_2(x), \ldots, \beta_p(x))', \end{aligned}$$

and

$$\begin{aligned} \mathbf{W}(x) &= \begin{pmatrix} w_1(x) & 0 & 0 & \ldots & 0 \\ 0 & w_2(x) & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \\ 0 & 0 & 0 & \ldots & w_n(x) \end{pmatrix} \\ &= \operatorname{diag}(w_1(x), w_2(x), \ldots, w_n(x)) \\ &= \operatorname{diag}\left(\frac{1}{b} K\left(\frac{x_1 - x}{b}\right), \frac{1}{b} K\left(\frac{x_2 - x}{b}\right), \ldots, \frac{1}{b} K\left(\frac{x_n - x}{b}\right)\right) \end{aligned}$$

so that

$$Q(x) = (\mathbf{y} - \mathbf{X}(x)\beta(x))' \, \mathbf{W}(x) \, (\mathbf{y} - \mathbf{X}(x)\beta(x)) \, . \tag{2.110}$$

For fixed x, differentiating $Q(x)$ with respect to $\beta(x)$ and equating to zero, one obtains the local polynomial estimator

$$\widehat{\beta}_{LP}(x) = \mathbf{A}(x)\mathbf{y}, \tag{2.111}$$

where

$$\widehat{\beta}_{LP}(x) = (\widehat{\beta}_0(x), \widehat{\beta}_1(x)\widehat{\beta}_2(x), \ldots, \widehat{\beta}_p(x))' \tag{2.112}$$

and

$$\mathbf{A}(x) = (\mathbf{X}'(x)\mathbf{W}(x)\mathbf{X}(x))^{-1}\mathbf{X}'(x)\mathbf{W}(x). \tag{2.113}$$

Here we have assumed that the matrix $\mathbf{X}'(x)\mathbf{W}(x)\mathbf{X}(x)$ is invertible. We have

$$\mathbf{X}'(x)\mathbf{W}(x) = \begin{pmatrix} w_1(x) & w_2(x) & w_3(x) & \ldots & w_n(x) \\ w_1(x)(x_1 - x) & w_2(x)(x_2 - x) & w_3(x)(x_3 - x) & \ldots & w_n(x)(x_n - x) \\ w_1(x)(x_1 - x)^2 & w_2(x)(x_2 - x)^2 & w_3(x)(x_3 - x)^2 & \ldots & w_n(x)(x_n - x)^2 \\ \ldots & \ldots & \ldots & \ldots & \\ w_1(x)(x_1 - x)^p & w_2(x)(x_2 - x)^p & w_3(x)(x_3 - x)^p & \ldots & w_n(x)(x_n - x)^p \end{pmatrix}, \tag{2.114}$$

so that

$$\mathbf{X}'(x)\mathbf{W}(x)\mathbf{X}(x) = \begin{pmatrix} \sum_{i=1}^n w_i(x) & \sum_{i=1}^n w_i(x)(x_i - x) & \ldots & \sum_{i=1}^n w_i(x)(x_i - x)^p \\ \sum_{i=1}^n w_i(x)(x_i - x) & \sum_{i=1}^n w_i(x)(x_i - x)^2 & \ldots & \sum_{i=1}^n w_i(x)(x_i - x)^{p+1} \\ \ldots & \ldots & \ldots & \ldots \\ \sum_{i=1}^n w_i(x)(x_i - x)^p & \sum_{i=1}^n w_i(x)(x_i - x)^{p+1} & \ldots & \sum_{i=1}^n w_i(x)(x_i - x)^{2p} \end{pmatrix}, \tag{2.115}$$

whereas

$$\mathbf{X}'(x)\mathbf{W}(x)\mathbf{Y} = \left(\sum_{i=1}^n w_i(x)y_i, \quad \sum_{i=1}^n w_i(x)(x_i - x)y_i, \quad \ldots, \quad \sum_{i=1}^n w_i(x)(x_i - x)^p y_i \right). \tag{2.116}$$

The estimate of the scalar-valued regression function $m(x)$ is given by $\widehat{\beta}_0(x)$, which can be conveniently written as

$$\widehat{\beta}_0(x) = \mathbf{s}'(x) \cdot \mathbf{y}, \tag{2.117}$$

where

$$\mathbf{s}'(x) = \mathbf{e}_0' \mathbf{A}(x) \tag{2.118}$$

and

$$\mathbf{e}_0' = (1, 0, \ldots, 0) \tag{2.119}$$

is a row vector with $(p+1)$ elements having 1 in its first position and zero elsewhere. More generally, defining the notation

$$\mathbf{e}_\nu' = (0, 0, \ldots 1, \ldots, 0) \tag{2.120}$$

to denote a row vector with $(p+1)$ elements having 1 in position $(\nu+1)$ and zero elsewhere, we can define the estimator

$$\widehat{\beta}_\nu(x) = \mathbf{e}_\nu' \mathbf{A}(x)\mathbf{y} \tag{2.121}$$

so that the νth derivative of the regression function $m(x)$ can be estimated from

$$\widehat{m}_{LP}^{(\nu)}(x) = \frac{\widehat{\beta}_\nu(x)}{\nu!} = \frac{\mathbf{e}_\nu' \mathbf{A}(x)\mathbf{y}}{\nu!}. \tag{2.122}$$

In particular, with $\nu = 0$, one obtains the intercept, the estimated regression function

$$\widehat{m}_{LP}(x) = \widehat{\beta}_0(x) = \mathbf{e}_0' \mathbf{A}(x)\mathbf{y}. \tag{2.123}$$

This estimator of $m(x)$ is the local polynomial estimator of degree p. Note that, if in the Taylor series expansion of $m(x_i)$ around $m(x)$, we let $p = 0$, then we get the local-constant estimator. It is in fact the Nadaraya–Watson estimator. Taking $p = 1$, one has the local-linear estimator, with $p = 2$, the local-quadratic, with $p = 3$, the local-cubic, and so on. The natural question then arises about the choice of the degree of the polynomial that is being used for the estimation. An insight into this problem is obtained from the asymptotic expression for the bias of the estimator, where a fundamental difference between the two cases, when $p - \nu$ is odd versus when $p - \nu$ is even, emerges. Here, ν is the order of the derivative of the regression function, estimation of which is of interest. To derive these properties, it

may be convenient to write the local-polynomial estimator as a kernel estimate. Of special interest are the so-called *equivalent kernels*.

2.3.1 Equivalent kernels

As we have seen above, the local-polynomial estimates of the regression function and its derivatives are obtained by considering a local multiple (polynomial) regression model and by using weighted least squares. In other words, the properties of the estimator can be derived by taking advantage of the theory of least squares. It is also possible to express the local polynomial estimators as kernel estimators. Using new notations (see, for example, Fan et al. 1997), let

$$\begin{aligned} S_n &= \mathbf{X}'(x)\mathbf{W}(x)\mathbf{X}(x) \\ &= \begin{pmatrix} S_0(x) & S_1(x) & \dots & S_p(x) \\ S_1(x) & S_2(x) & \dots & S_{p+1}(x) \\ \dots & \dots & \dots & \dots \\ S_p(x) & S_{p+1}(x) & \dots & S_{2p}(x) \end{pmatrix} \end{aligned} \tag{2.124}$$

and

$$\begin{aligned} \mathbf{T}_n &= \mathbf{X}'(x)\mathbf{W}(x)\mathbf{Y} \\ &= (T_0(x), T_1(x), \dots, T_p(x))' \end{aligned} \tag{2.125}$$

where

$$bS_j(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{b}\right)(x_i - x)^j,\ j = 0, 1, \dots, 2p, \tag{2.126}$$

$$bT_j(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{b}\right)(x_i - x)^j y_i,\ j = 0, 1, \dots, p. \tag{2.127}$$

We have

$$\begin{aligned} \widehat{\beta}_\nu(x) &= \mathbf{e}'_\nu \cdot \widehat{\beta}(x),\ \nu = 0, 1, \dots, p \\ &= \sum_{i=1}^{n} w_{\nu,n}\left(\frac{x_i - x}{b}\right) y_i \end{aligned} \tag{2.128}$$

where

$$w_{\nu,n}(t) = \mathbf{e}_\nu' \cdot S_n^{-1} \cdot (1, tb, (tb)^2, \dots, (tb)^p)' \frac{K(t)}{b}. \tag{2.129}$$

It is clear from (2.128) that $\widehat{\beta}_\nu(x)$ has the form of a kernel estimator where, however, the kernel $w_{\nu,n}(t)$ depends on the n observations $x_1, \dots, x_n$ on the explanatory variable X. In particular, this kernel satisfies the (finite sample) moment condition

$$\sum_{i=1}^{n} (x_i - x)^q w_{\nu,n}\left(\frac{x_i - x}{b}\right) = \delta_{\nu,q}, \ \nu, q = 0, 1, 2, \dots, p, \tag{2.130}$$

where $\delta_{\nu,q} = 1$ if $\nu = q$ and zero otherwise. This kernel representation can then be further exploited for derivation of various properties of the estimator.

As regards bias, note that if we take the expectation of $\widehat{\beta}_\nu(x)$ in (2.128) and expand $\mathbb{E}(y_i) = m(x_i)$ using a polynomial of degree up to p around $m(x)$, except for the contribution from the remainder of this Taylor series expansion, the rest of the terms lead to zero bias, even when n is finite. Of course, in case $m(x_i)$ is in fact a polynomial of degree p, as in the first $p + 1$ terms in (2.105), $\widehat{\beta}_\nu(x)$ is an unbiased estimator of $\beta_\nu(x)$. This is a direct consequence of the fact that $\widehat{\beta}_\nu(x)$ is a weighted least squares estimator. See Ruppert and Wand (1994) for further remarks.

As for derivative estimation, the choice of the degree of the polynomial is an important issue. Estimating the regression function using a polynomial of degree zero leads to the Nadaraya–Watson estimator. In other words, this is the *local constant* estimator. Based on asymptotic considerations, one can argue that the Nadaraya–Watson estimator will have some disadvantages compared to another estimator that uses a polynomial of another suitably chosen degree, e.g., the *local-linear* estimator with $p = 1$:

$p - \nu$ is odd:

$$\text{Bias}(\widehat{m^{(\nu)}}(x)) = a_1 \times \frac{m^{(p+1)}(x)}{(p+1)!} \nu! b^{p+1-\nu} + r_1, \tag{2.131}$$

$$r_1 = o(b^{p+1-\nu}); \tag{2.132}$$

$p - \nu$ is even:

$$\begin{aligned}&\text{Bias}(\widehat{m^{(\nu)}}(x))\\ &= a_2 \times \left\{ \frac{m^{(p+2)}(x)}{(p+2)!} + \frac{m^{(p+1)}(x)}{(p+1)!} \frac{f^{(1)}(x)}{f(x)} \right\} \nu! b^{p+2-\nu} + r_2,\end{aligned} \tag{2.133}$$

$$r_2 = o(b^{p+2-\nu}), \tag{2.134}$$

and a_1 and a_2 do not depend on n or b.

However, in both cases,

$p - \nu$ is odd or even:

$$\mathbb{V}ar(\widehat{m^{(\nu)}}(x)) = a_3 \times \frac{(\nu!)^2 \sigma^2}{n b^{1+2\nu} f(x)} + r_3, \tag{2.135}$$

$$r_3 = o\left(\frac{1}{n b^{1+2\nu}} \right) \tag{2.136}$$

and a_3 does not depend on n or b. Here $f(x)$ is the design density and $m^{(p+2)}(x)$ and $f^{(1)}$ are continuous in a neighborhood of x.

Thus, there is a theoretical difference between the two cases, i.e., when $p - \nu$ is odd and when it is even. In particular, when $p - \nu$ is even, $f^{(1)}(x)/f(x)$ appears in the asymptotic expression for the bias, through the additional term $\{m^{(p+1)}(x)/(p+1)!\} \times \{f^{(1)}(x)/f(x)\}$. In particular, this choice of the degree p (so that $p - \nu$ is even) allows the bias of the estimator to be affected by the design density (i.e., distribution of the points in the x-axis). This can be problematic especially near the boundary of the x-space. Thus, for estimating the νth derivative of the regression function $m(x)$ a remedy for the (boundary) bias problem is to select the degree p in such a way that the difference $p - \nu$ becomes odd, while at the same time keeping p low so that estimation of not too many terms is involved. For instance, all else remaining fixed, to attain an asymptotic bias rate of b^2 for estimating the first derivative of $m(x)$, the bias rule suggests taking $p = 2$. For details, see Fan and Gijbels (1996) and references therein.

In computations, one may give an approximate confidence interval for the regression function $m(x)$ as follows.

First of all, estimating $m(x)$ at the observed values of x (using the above method), namely at $x = x_1, x = x_2, \ldots, x = x_n$, and collecting the estimates in a column vector $\widehat{\mathbf{m}}$ we can write

$$\widehat{\mathbf{m}} = \begin{pmatrix} \widehat{m}(x_1) \\ \widehat{m}(x_2) \\ \cdots \\ \widehat{m}(x_n) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1' \mathbf{A}(x_1)\mathbf{y} \\ \mathbf{e}_1' \mathbf{A}(x_2)\mathbf{y} \\ \cdots \\ \mathbf{e}_1' \mathbf{A}(x_n)\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{s}'(x_1) \\ \mathbf{s}'(x_2) \\ \cdots \\ \mathbf{s}'(x_n) \end{pmatrix} \mathbf{y} = \mathbf{S}\mathbf{y}$$

where the matrix $\mathbf{S}$ defined as

$$\mathbf{S} = \begin{pmatrix} \mathbf{s}'(x_1) \\ \mathbf{s}'(x_2) \\ \cdots \\ \mathbf{s}'(x_n) \end{pmatrix}$$

is termed the *smoother matrix.* When $x_1, x_2, \ldots, x_n$ are fixed, we have the covariance matrix of the vector $\widehat{\mathbf{m}}$ as

$$\mathbb{C}ov(\widehat{\mathbf{m}}) = \mathbf{S}\mathbb{C}ov(\mathbf{y})\,\mathbf{S}' = \mathbf{S}\mathbf{S}'\sigma^2. \tag{2.137}$$

Finally, for $n \to \infty$, using the estimator for σ^2 as

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \widehat{m}(x_i)\}^2, \tag{2.138}$$

an asymptotic $100(1-\alpha)\%$ confidence interval (*ignoring bias*) for the regression function at $x = x_i$ may be given as

$$\widehat{m}(x_i) \pm z_{\alpha/2}\widehat{\sigma}\sqrt{(\mathbf{S}\mathbf{S}')_{ii}} \tag{2.139}$$

where $(\mathbf{S}\mathbf{S}')_{ii}$ is the ith diagonal element of $\mathbf{S}\mathbf{S}'$ and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution, i.e., $1 - \Phi(z_{\alpha/2}) = \alpha/2$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

2.4 Nadaraya–Watson regression estimator

Consider the nonparametric regression model with independent observations $(x_i, y_i,)$, $i = 1, 2, \ldots, n$, on the pair of random variables (X, Y). Let the marginal probability density function of X be $f(x)$ and the conditional mean of Y given $X = x$ be $m(x)$. In

particular, consider the nonparametric regression model

$$y_i = m(x_i) + u_i,\ i = 1, 2, \ldots, n \tag{2.140}$$

with iid errors $u_i,\ i = 1, 2, \ldots, n$ having mean

$$\mathbb{E}(u_i) = 0 \tag{2.141}$$

and variance

$$\mathbb{V}ar(u_i) = \sigma^2, \tag{2.142}$$

where $0 < \sigma < \infty$. The regression function is

$$\mathbb{E}(y_i|x_i) = m(x_i). \tag{2.143}$$

Suppose that m is smooth and the problem is a nonparametric estimation of this function.

The Nadaraya–Watson estimator (Nadaraya 1964 and Watson 1964) is particularly designed for the situation when the values of the design variable X, namely $x_1, x_2, \ldots, x_n$, are random. It is given by

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{b}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{b}\right)} \tag{2.144}$$

where the bandwidth b and the kernel K satisfy the following conditions:

- As $n \to \infty$, $b \to 0$ and $nb \to \infty$.
- The kernel K is bounded and continuous, three times continuously differentiable, with bounded derivatives. Moreover,
  - $K(u) \geq 0$ for all $u \in R$
  - $\int K(u)du = 1$, $K(u) = K(-u)$ for all $u \in R$
  - $\int |u|^3 K(u)du < \infty$

Note that the Nadaraya–Watson estimator can be derived as a local-constant estimator, i.e., as a local-polynomial estimator based on a polynomial of degree $p = 0$. This is easy to see since, for fixed x, let $\theta = m(x)$. To estimate θ, using a local polynomial approach, we minimize the quadratic form $Q(\theta)$ and obtain

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \{Q(\theta)\} \tag{2.145}$$

where the weighted error sum of squares to be minimized is

$$Q(\theta) = \sum_{i=1}^{n} (y_i - \theta)^2 w_i(x). \tag{2.146}$$

Differentiation of $Q(\theta)$ with respect to θ yields the estimator

$$\widehat{\theta} = \widehat{m}_{NW}(x) \tag{2.147}$$

as defined in (2.144).

Now recall the formula for the Priestley–Chao regression estimator of the regression function $m(x)$ and also the formula for the Parzen–Rosenblatt nonparametric density estimator of the design density $f(x)$. We recollect these formulas below:

Priestley–Chao regression estimator:

$$\widehat{m}_{PC}(x) = \frac{1}{nb} \sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{b}\right) \tag{2.148}$$

Parzen–Rosenblatt density estimator:

$$\widehat{f}_{PR}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x_i - x}{b}\right) \tag{2.149}$$

Then the Nadaraya–Watson estimator can be written as

$$\widehat{m}_{NW}(x) = \widehat{m}_{PC}(x)/\widehat{f}_{PR}(x) \tag{2.150}$$

As we have seen earlier in this chapter, when the $x_1, \ldots, x_n$ are evenly spaced and fixed with $x_i = i/n,\ i = 1, 2, \ldots, n$, the bias of the Priestley–Chao estimator of $m(x)$ is of the order $O(b^2)$ and its variance is of the order $O(1/(nb))$. Thus, with $x_i = i/n$, $i = 1, 2, \ldots, n$,

$$\mathbb{E}(\widehat{m}_{PC}(x)) = m(x) + O(b^2), \tag{2.151}$$

$$\mathbb{V}ar(\widehat{m}_{PC}(x)) = O(1/(nb)), \tag{2.152}$$

so that in this case $\widehat{m}_{PC}(x)$ is a consistent estimator of $m(x)$. However, if $x_1, \ldots, x_n$ are iid random variables, with a common pdf f, then $\widehat{m}_{PC}(x)$ is a consistent estimator of $m(x)f(x)$. In other words, a multiplicative factor $f(x)$ appears. As a result, division of $\widehat{m}_{PC}(x)$ by a consistent estimator of $f(x)$ becomes

necessary. As we have seen in Chapter 1 on Density Estimation, for every fixed x, the bias and the variance of the Parzen–Rosenblatt density estimator converge to zero, at the rates $O(b^2)$ and $O(1/(n))$ respectively. In other words, pointwise weak consistency of $\widehat{f}_{PR}(x)$ follows due to Chebyshev's inequality. Finally, weak consistency of each of the estimators $\widehat{m}_{PC}(x)$ and $\widehat{f}_{PR}(x)$, combined with Slutsky's lemma when $f(x) > 0$, ensures pointwise weak consistency of $\widehat{m}_{NW}(x)$.

To see these explicitly, we start with $\widehat{m}_{PC}(x)$ in the random design case. Denote $\mu_2(K) = \int u^2 K(u)du$ and $R(K) = \int K^2(u)du$. First of all,

$$\begin{aligned}
&\mathbb{E}(\widehat{m}_{PC}(x)) = \frac{1}{nb}\mathbb{E}\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{b}\right) \\
&= \frac{1}{b}\int_{-\infty}^{\infty} m(z)K\left(\frac{z-x}{b}\right)f(z)dz \\
&= \int_{-\infty}^{\infty}\left[m(x) + bu\, m^{(1)}(x) + \frac{b^2u^2}{2}m^{(2)}(x) + \cdots\right] \\
&\quad \times\left[f(x) + bu\, f^{(1)}(x) + \frac{b^2u^2}{2}f^{(2)}(x) + \ldots\right]K(u)du \\
&= m(x)f(x) + \mu_2(K)\frac{b^2}{2}[m(x)f^{(2)}(x) + f(x)m^{(2)}(x) \\
&\quad + 2m^{(1)}(x)f^{(1)}(x)] + o(b^2) \\
&= m(x)f(x) + \mu_2(K)\frac{b^2}{2}\frac{\partial^2}{\partial x^2}\{m(x)f(x)\} + o(b^2). \qquad (2.153)
\end{aligned}$$

As for the variance,

$$\begin{aligned}
&\mathbb{V}ar(\widehat{m}_{PC}(x)) = \frac{1}{nb^2}\mathbb{V}ar\left(y_i K\left(\frac{x_i - x}{b}\right)\right) \\
&= \frac{1}{nb^2}\left[\mathbb{E}\left(y_i^2 K^2\left(\frac{x_i - x}{b}\right)\right) - \mathbb{E}^2\left(y_i K\left(\frac{x_i - x}{b}\right)\right)\right] \\
&= \frac{1}{nb^2}\left[\int_{-\infty}^{\infty}(\sigma^2 + m^2(x_i))K^2\left(\frac{x_i - x}{b}\right)f(x_i)dx_i\right. \\
&\quad \left. - b^2\{m(x)f(x) + O(b^2)\}^2\right] \\
&= \frac{1}{nb}f(x)R(k)(\sigma^2 + m^2(x)) + o\left(\frac{1}{nb}\right). \qquad (2.154)
\end{aligned}$$

This then leads to the pointwise weak consistency of the Priestly–Chao estimator, so that for every fixed x, as $n \to \infty$, $\widehat{m}_{PC}(x)$ converges in probability to $m(x)f(x)$. On the other hand, consistency of the Priestly-Rosenblatt estimator was established in Chapter 1 on Density Estimation and we may recall that for every fixed x, as $n \to \infty$, $\widehat{f}_{PR}(x)$ converges in probability to $f(x)$. Thus, for $f(x) > 0$, by Slutsky's lemma it follows that, as $n \to \infty$, $\widehat{m}_{NW}(x)$ converges in probability to $m(x)f(x)/f(x) = m(x)$, so that the Nadaraya–Watson estimator gives rise to a consistent estimate of the regression function $m(x)$.

The covariance between $\widehat{m}_{PC}(x)$ and $\widehat{f}_{NW}(x)$ is also of interest:

$$
\begin{aligned}
&\mathbb{C}ov(\widehat{m}_{PC}(x), \widehat{f}_{NW}(x)) \\
&= \mathbb{E}(\widehat{m}_{PC}(x)\widehat{f}_{NW}(x)) - \mathbb{E}(\widehat{m}_{PC}(x))\mathbb{E}(\widehat{f}_{NW}(x)) \\
&= \mathbb{E}(\widehat{m}_{PC}(x)\widehat{f}_{NW}(x)) \\
&\quad - \left[m(x)f(x) + \frac{b^2}{2}\mu_2(K)(m^{(2)}(x)f(x) + f^{(2)}(x)m(x)\right. \\
&\quad \left. + 2m^{(1)}(x)f^{(1)}(x)) + o(b^2)\right] \\
&\quad \times \left[f(x) + \frac{b^2}{2}\mu_2(K)f^{(2)}(x) + o(b^2)\right] \\
&= \mathbb{E}(\widehat{m}_{PC}(x)\widehat{f}_{NW}(x)) - m(x)f^2(x) \\
&\quad - \frac{b^2}{2}\mu_2(K)[m^{(2)}(x)f^2(x) + 2m(x)f(x)f^{(2)}(x) \\
&\quad + 2f(x)f^{(1)}(x)m^{(1)}(x)] + o(b^2) \qquad (2.155)
\end{aligned}
$$

whereas

$$
\begin{aligned}
&\mathbb{E}(\widehat{m}_{PC}(x)\widehat{f}_{NW}(x)) \\
&= \mathbb{E}\left[\frac{1}{n^2b^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i K\left(\frac{x_i - x}{b}\right)K\left(\frac{x_j - x}{b}\right)\right] \\
&= \frac{1}{n^2b^2}\mathbb{E}\left[\sum_{i=1}^{n} y_i K^2\left(\frac{x_i - x}{b}\right) + \sum_{i\neq j=1}^{n} y_i K\left(\frac{x_i - x}{b}\right)K\left(\frac{x_j - x}{b}\right)\right] \\
&= \frac{1}{n^2b^2}\left[n\int_{-\infty}^{\infty} m(z)K^2\left(\frac{z - x}{b}\right)f(z)dz\right]
\end{aligned}
$$

(continued to next page)

$$
\begin{aligned}
&+\frac{1}{n^2b^2}\left[n(n-1)\left(\int_{-\infty}^{\infty}m(z)K\left(\frac{z-x}{b}\right)f(z)dz\right)\right.\\
&\left.\times\left(\int_{-\infty}^{\infty}K\left(\frac{z-x}{b}\right)f(z)dz\right)\right]\\
&=\frac{1}{n^2b^2}[nb\{m(x)f(x)R(K)+O(b^2)\}]\\
&+\frac{1}{n^2b^2}\left[n(n-1)b^2\{m(x)f(x)+a_1(b)\}\{f(x)+a_2(b)\}\right]\\
&=m(x)f^2(x)+\frac{1}{nb}m(x)f(x)R(K)+a_3(b) \qquad (2.156)
\end{aligned}
$$

where

$$c(b)=\mu_2(K)\frac{b^2}{2} \qquad (2.157)$$

$$
\begin{aligned}
a_1(b)&=c(b)[f(x)m^{(2)}(x)+m(x)f^{(2)}(x)+2m^{(1)}(x)f^{(1)}(x)]\\
&\quad+o(b^2) \qquad (2.158)
\end{aligned}
$$

$$a_2(b)=c(b)f^{(2)}(x)+o(b^2) \qquad (2.159)$$

$$
\begin{aligned}
a_3(b)&=c(b)f(x)[2m(x)f^{(2)}(x)+2m^{(1)}(x)f^{(1)}(x)\\
&\quad+f(x)m^{(2)}(x)]+o(b^2). \qquad (2.160)
\end{aligned}
$$

Combining,

$$\mathbb{C}ov(\widehat{m}_{PC}(x),\widehat{f}_{NW}(x))=\frac{1}{nb}m(x)f(x)R(K)+o\left(\frac{1}{nb}\right). \qquad (2.161)$$

To derive the asymptotic expressions for the (unconditional) bias and the variance of $\widehat{m}_{NW}(x)$, define the function $\psi:\mathbb{R}\times\mathbb{R}_+\to\mathbb{R}$ such that

$$\widehat{m}_{NW}(x)=\psi(\widehat{m}_P(x),\widehat{f}_{PR}(x))=\widehat{m}_P(x)/\widehat{f}_{PR}(x) \qquad (2.162)$$

Noting that $\widehat{m}_{NW}(x)$ is a ratio of two means, for fixed x, expanding around $\psi(m(x)f(x),f(x))=m(x)$ using Taylor series expansion, we have

$$
\begin{aligned}
&\psi(\widehat{m}_P(x),\widehat{f}_{PR}(x))=\psi(m(x)f(x),f(x))\\
&+(\widehat{m}_{PC}(x)-m(x)f(x))-(\widehat{f}_{PR})x)-f(x))\frac{m(x)}{f(x)}+\cdots \qquad (2.163)
\end{aligned}
$$

Taking the expected value and combining terms, we have

$$\begin{aligned}&\mathbb{E}[\psi(\widehat{m}_P(x), \widehat{f}_{PR}(x))] = \mathbb{E}[\widehat{m}_{NW}(x)] \\ &= m(x) + \mu_2(K)\frac{b^2}{2}\left(m^{(2)}(x) + 2\frac{m^{(1)}(x)f^{(1)}(x)}{f(x)}\right) + o(b^2)\end{aligned} \tag{2.164}$$

whereas the variance becomes

$$\begin{aligned}&\mathbb{V}ar[\psi(\widehat{m}_P(x), \widehat{f}_{PR}(x))] = \mathbb{V}ar[\widehat{m}_{NW}(x)] \\ &= \frac{1}{f^2(x)}\mathbb{V}ar[\widehat{m}PC(x)] + \frac{m^2(x)}{f^2(x)}\mathbb{V}ar[\widehat{f}PR(x)] \\ &\quad - 2\frac{m(x)}{f^2(x)}\mathbb{C}ov[\widehat{m}PC(x), \widehat{f}_{PR}(x)] \\ &= \frac{1}{nb}\frac{\sigma^2 R(K)}{f(x)} + o\left(\frac{1}{nb}\right)\end{aligned} \tag{2.165}$$

In other words, for every fixed x, $\widehat{m}_{NW}(x)$ converges to $m(x)$ in probability with $n \to \infty$. As indicated in the discussion about the local-polynomial based estimation, being a local-constant estimator for the regression function, $\widehat{m}_{NW}(x)$ also suffers from the boundary problem in its bias. The presence of the design density can inflate the bias, in particular in the boundary region of the x-space.

2.5 Bandwidth selection

There are various reviews of bandwidth selection procedures. Some references are Benedetti (1977), Chiu (1989), Gijbels and Goderniaux (2004b), Herrmann (1997, 2000), Herrmann et al. (1992), Loader (1999), Schucany (2004), Jones et al. (1991, 1996), Köhler et al. (2014), Müller et al. (1987), Park and Marron (1990), Rice (1984), Sheather (1992), Sheather and Jones (1991), and others, as well as the list of references in Wand and Jones (1995). Here we only address the asymptotic properties of the regression estimator, followed by a brief look at the LSCV method.

There are two main streams of approaches for selecting an optimum bandwidth. In one approach, the idea is to minimize the residual sum of squares and, for this, using residuals from

cross-validation (CV), the generalized cross-validation (gcv), and the minimum unbiased risk estimation. The second approach involves substituting estimates of the unknown quantities in the expression for the asymptotic mean squared error. This is the so-called plug-in approach.

To address the main issue for a cross-validation based method, we see that the expressions for bias and the variance of the estimator of the regression function can be used to derive the formula for the optimal bandwidth b. This formula is obtained by minimizing the leading term of the asymptotic expression of the mean squared error (mse) of $\widehat{m}_{PC}(x)$. In fact, formulas for both "local optimal bandwidth" for each fixed x and the "global optimal bandwidth" can be derived. Specifically, collecting the leading terms in the asymptotic expression for the mean squared error, where $mse = Bias^2 + Variance$, define

$$AMSE(\widehat{m}_{PC}(x)) = \left(\frac{b^2}{2} m^{(2)}(x) \int_{-1}^{1} u^2 K(u) du\right)^2 + \frac{\sigma^2}{nb} \int_{-1}^{1} K^2(u) du \tag{2.166}$$

Note the opposing effect of the bandwidth b on bias and variance, the so-called bias–variance trade-off, where all else remains fixed, increasing b reduces variance but increases bias and vice versa. Considering, however, the mean squared error, one approach is to minimize $AMSE$ so that

$$\frac{d}{db} AMSE(\widehat{m}_{PC}(x)) = 0 \tag{2.167}$$

leads to the formula for the locally optimum bandwidth

$$b_{opt}^{(local)}(x) = \left\{ \frac{\sigma^2 R(K)}{(m^{(2)}(x))^2 \mu_2^2(K)} \right\}^{1/5} n^{-1/5} \tag{2.168}$$

where we have used the notations $R(K) = \int K^2(u) du$ and $\mu_2^2(K) = \int u^2 K(u) du$.

In contrast to having a variable bandwidth that depends on the location x, one may in some situations opt for the constant or the global bandwidth. One way to do this is to consider the mean integrated squared error (MISE) and, in particular, the leading term in its asymptotic expansion. This is then given by simply

integrating out x in the formula for $AMSE(x)$ above. We then define

$$AMISE(\widehat{m}_{PC}) = \int AMSE(\widehat{m}_{PC}(x))dx$$
$$= \left(\frac{b^2}{2}R(m^{(2)})\mu_2(K)\right)^2 + \frac{\sigma^2}{nb}R(K) \tag{2.169}$$

where, using previous notation, $R\left(m^{(2)}\right) = \int \left(m^{(2)}(x)\right)^2 dx$. Differentiating $AMISE$ with respect to b and equating to zero, the formula for the global optimum bandwidth is obtained. This formula is given by

$$b_{opt}^{(global)} = \left\{\frac{\sigma^2 R(K)}{R(m^{(2)}(x))\mu_2^2(K)}\right\}^{1/5} n^{-1/5}. \tag{2.170}$$

What is in particular worth noting is the role of the second derivative of m on the optimum bandwidth. In particular, large values of the second derivative, high local variations in $m(x)$, leads to smaller optimum bandwidths. Moreover, substituting $b_{opt}^{(global)}$ or $b_{opt}^{(local)}(x)$ into $AMISE$ or $AMSE(x)$ respectively yields

$$\text{Best asymptotic rate } = \alpha \times C(K) \times n^{-4/5} \tag{2.171}$$

where α does not depend on the kernel K and

$$C(K) = [R(K)]^{4/5}\left[\mu_2^2(K)\right]^{1/5}. \tag{2.172}$$

The rate of decay in (2.171) is proportional to $n^{-4/5}$. Thus, in contrast to typical parametric estimation procedures, where the rate of convergence is n^{-1}, the best rate of convergence to zero for these nonparametric curve estimates, under the conditions mentioned above, is slower.

To make the above formula usable, in practice one would have to substitute estimates of σ^2 and $m^{(2)}(x)$. For instance, an estimate of σ^2 may be based on the mean residual errors squared. To estimate the variance of the regression estimate, Gasser (1986) considers a finite-difference based approach. Another approach is to consider smoothing the squared residuals using an appropriate kernel. This idea is suggested, for instance, in Menéndez et al. (2013) for correlated data. Moreover, estimation of $m^{(2)}(x)$

will again involve optimality considerations. Substituting the estimated unknown quantities into the formula for the optimum bandwidth leads to the so-called *plug-in* approaches. One option is to twice differentiate the formula for $\widehat{m}_{PC}(x)$ with respect to x, which essentially leads to an estimation procedure based on the higher order kernels, such as

$$\widehat{m}^{(2)}_{PC}(x) = \frac{1}{nh^3} \sum_{i=1}^{n} K^{(2)} \left(\frac{x_i - x}{h} \right) y_i \tag{2.173}$$

where $K^{(r)}(u)$ is the rth derivative of $K(u)$ with respect to u and h is a pilot bandwidth. This approach would thus require using differentiable kernels K, and, for instance, the uniform kernel where $K(u) = 1/2$, if $|u| \leq 1$ and $K(u) = 0$, if $|u| \geq 1$, would not be useful. The problem with this approach is that a pilot bandwidth h is needed and hence the bandwidth selection problem is shifted to choosing an optimum h. One option is to take $h = b$, which, however, need not be the ideal choice. For independent errors, Gasser et al. (1991) consider

$$h = bn^{\alpha} \tag{2.174}$$

where n^{α} is an inflation factor, $\alpha > 0$, and argue that

$$h = bn^{1/10} \tag{2.175}$$

is the ideal choice, b being the bandwidth used to estimate the initial regression function m. This idea is further modified when the errors are correlated. See Herrmann et al. (1992) for further information on an algorithm.

The method of local polynomials provides another approach to derivative estimation. Also see Gasser and Müller (1984) who address derivative estimation using the general class of higher order kernels.

As mentioned earlier, formulas for both "local optimal bandwidth" for each fixed x and the "global optimal bandwidth" can be derived by minimizing the leading term in the asymptotic expression for the mean squared error of $\widehat{m}_{PC}(x)$. To make this method data-driven, one option would be to plug-in estimates of the unknown functions and parameters estimated from the given set of observations. Another approach is not to use the

asymptotic expression of the mse of $\widehat{m}_{PC}(x)$ but rather use cross-validation. This is described below.

Consider the following sums of squares corresponding to the nonparametric regression model above and the estimator $\widehat{m}_{PC}(\cdot)$

$$\text{Error sum of squares} = \text{ESS}(b) = \sum_{i=1}^{n}(y_i - m(x_i))^2 = \sum_{i=1}^{n} u_i^2 \tag{2.176}$$

$$\text{Average squared error} = \text{ASE}(b) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_{PC}(x_i) - m(x_i))^2 \tag{2.177}$$

$$\text{Average residual sum of squares} = \text{ARSS}(b)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}_{PC}(x_i))^2. \tag{2.178}$$

Then the expression for ARSS(b) can be written as

$$\text{ARSS}(b) = \frac{1}{n}\text{ESS}(b) + \text{ASE}(b) - \frac{2}{n}\sum_{i=1}^{n} u_i(\widehat{m}_{PC}(x_i) - m(x_i)). \tag{2.179}$$

Taking the expectation of both sides of the last equation,

$$\mathbb{E}(\text{ARSS}(b)) = \sigma^2 + \mathbb{E}(\text{ASE}(b)) - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}(u_i(\widehat{m}_{PC}(x_i) - m(x_i))). \tag{2.180}$$

However,

$$\mathbb{E}(u_i m(x_i)) = m(x_i)\mathbb{E}(u_i) = 0, \tag{2.181}$$

so that

$$\mathbb{E}(\text{ARSS}(b)) = \sigma^2 + \mathbb{E}(\text{ASE}(b)) - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}(u_i\widehat{m}_{PC}(x_i)) \tag{2.182}$$

Also,

$$\begin{aligned}\mathbb{E}(u_i\widehat{m}_{PC}(x_i)) &= \mathbb{E}\left(u_i\frac{1}{nb}\sum_{j=1}^{n}y_jK\left(\frac{x_j-x_i}{b}\right)\right)\\ &= \mathbb{E}\left(u_i\cdot\frac{1}{nb}\sum_{j=1}^{n}(m(x_j)+u_j)K\left(\frac{x_j-x_i}{b}\right)\right)\\ &= \mathbb{E}\left(u_i\cdot\frac{1}{nb}(m(x_i)+u_i)K\left(\frac{x_i-x_i}{b}\right)\right)\\ &= \mathbb{E}\left(\frac{1}{nb}{u_i}^2K(0)\right)=\frac{\sigma^2}{nb}K(0).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}(\text{ARSS}(b)) &= \sigma^2+\mathbb{E}(\text{ASE}(b))-\frac{2}{n}\sum_{i=1}^{n}\frac{\sigma^2}{nb}K(0)\\ &= \sigma^2+\mathbb{E}(\text{ASE}(b))-\frac{2\sigma^2}{nb}K(0)\end{aligned}$$

Thus ARSS(b) can be taken as an unbiased estimator of $\mathbb{E}(\text{ASE}(b))+\sigma^2-(2\sigma^2/nb)K(0)$. However, although we would like to find b that minimizes $\mathbb{E}(\text{ASE}(b))$ or, in a data-driven method, an unbiased estimate of $\mathbb{E}(\text{ASE}(b))$, plus perhaps a constant (such as σ^2) that does not depend on b, the presence of b in the denominator (see the right-hand sides of the above equations) makes this approach not meaningful. Direct minimization of ARSS(b) leads to selection of the smallest b on the chosen grid so that this bandwidth selection algorithm breaks down.

The remedy for this is to use a revised version of the ARSS(b) that eliminates the term $(2\sigma^2/nb)K(0)$ from its expectation; this leads to the least squares cross-validation approach. To start with, we define the cross-validation function (*this is simply a revised version of the ARSS(b) defined above*)

$$CV(h)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{m}_{PC,-i}(x_i))^2. \tag{2.183}$$

Here, the notation $\widehat{m}_{PC,-i}(x_i)$ is defined as

$$\widehat{m}_{PC,-i}(x_i)=\frac{1}{nb}\sum_{j=1,j\neq i}^{n}y_jK\left(\frac{x_j-x_i}{b}\right). \tag{2.184}$$

In other words, this is the *leave-one-out* Priestley–Chao kernel regression estimator. Using the same argument as before, one can now show that

$$\mathbb{E}(CV(b)) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\widehat{m}_{PC,-i}(x_i) - m(x_i))^2 \qquad (2.185)$$

In particular, note that this time the quantity $(2\sigma^2/nb)K(0)$ no longer occurs.

Since in large samples

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_{PC,-i}(x_i) - m(x_i))^2\right) &\approx \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{m}_{PC}(x_i) - m(x_i))^2\right) \\ &= \mathbb{E}(\mathrm{ASE}(b)), \qquad (2.186)\end{aligned}$$

the quantity $CV(b)$ can be used in a bandwidth selection algorithm.

Thus the LSCV criterion can now be given by

$$h_{opt}^{LSCV} = \underset{b}{\operatorname{argmin}}\ CV(b). \qquad (2.187)$$

2.6 Further remarks

2.6.1 Gasser–Müller estimator

There are also other kernel estimators, as for instance the estimators due to Theo Gasser and his colleagues; see, for example, Gasser and Müller (1984). Such an estimator is of the form

$$\begin{aligned}\widehat{m}_{GM}(x) &= \sum_{i=1}^{n} y_i \int_{s_{i-1}}^{s_i} \frac{1}{b} K\left(\frac{x-u}{b}\right) du,\ x \in [a_1 + b, a_2 - b] \\ &= \sum_{i=1}^{n} y_i \left\{ F_K\left(\frac{x - s_{i-1}}{b}\right) - F_K\left(\frac{x - s_i}{b}\right)\right\} \\ &= \sum_{i=1}^{n} y_i \tilde{w}_i(x, b) \qquad (2.188)\end{aligned}$$

where F_K is the cumulative distribution function corresponding to the kernel K,

$$\tilde{w}_i(x, b) = \left\{ F_K\left(\frac{x - s_{i-1}}{b}\right) - F_K\left(\frac{x - s_i}{b}\right) \right\}, \tag{2.189}$$

and $x_1 < \cdots < x_n \in [a_1, a_2]$, $-\infty < a_1 < a_2 < \infty$ satisfy

$$F(x_i) = (i - .5)/n, \; i = 1, 2, \ldots, n, \tag{2.190}$$

F being the cumulative distribution function corresponding to the design density f and $s_1, \ldots, s_{n-1}$ are the mid-points

$$s_i = 0.5(x_i + x_{i+1}), \tag{2.191}$$

$$s_0 = a_1, \tag{2.192}$$

$$s_n = a_2. \tag{2.193}$$

Moreover, the sum of the weights $\tilde{w}_i$ over $i = 1, 2, \ldots, n$ converges to unity as $n \to \infty$, where $b \to 0$ as $n \to \infty$ and x is fixed. For related bandwidth selection procedures, see, for instance, Herrmann (1997) and references therein.

2.6.2 Smoothing splines

2.6.2.1 The model

Given pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$, on the response variable Y and the explanatory variable X, we consider the regression model

$$y_i = m(x_i) + u_i$$

where the u_i are independently and identically distributed (iid) errors with zero mean and variance σ^2 and $m(x)$ is the regression function evaluated at x, i.e., it is the (conditional) mean of Y given $X = x$. Our problem is to estimate m. Now consider the naive least squares problem.

Find a function m that minimizes the residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n} \{y_i - m(x_i)\}^2. \tag{2.194}$$

The obvious solution to this problem is any function $\hat{m}$ such that matches the data, i.e.

$$\hat{m}(x_i) = y_i. \tag{2.195}$$

so that the residuals are exactly equal to zero, i.e. $\hat{u}_i = y_i - \hat{m}(x_i) = 0,\ i = 1, 2, \ldots, n$. From a statistical point of view, this solution is not satisfactory. The solution is heavily affected by the randomness and the distinction between the rough and the smooth parts is not adequate. This leads to an idea based on the use of a roughness penalty and hence the method of cubic splines. For background information, see Diggle (1990), Eubank (1988, 2000), Silverman (1984a, 1984b) and Wahba (1990).

A measure of roughness is the integrated squared second derivative

$$R(m^{(2)}) = \int_{-\infty}^{\infty} \{m^{(2)}(x)\}^2 dx, \tag{2.196}$$

where $m^{(2)}(x)$ is the second derivative of the regression function m. This leads to the following *penalized least squares* problem.

For $\lambda \geq 0$, find the function m that minimizes

$$\sum_{i=1}^{n} \{y_i - m(x_i)\}^2 + \lambda \int_{-\infty}^{\infty} \{m^{(2)}(x)\}^2 dx. \tag{2.197}$$

2.6.2.2 The parameter λ

λ plays the role of the *smoothing parameter* by balancing *Taylor series-of-fit* and *smoothness*. In particular $\lambda = 0$ corresponds to a perfect Taylor series-of-fit. This is the solution to the naive least squares problem. The case $\lambda = \infty$ corresponds to linear regression: $\hat{m}(x) = \alpha + \beta x$. In this case, the second derivative is zero. As λ ranges from 0 to ∞, the estimate or the fitted curve ranges from the most complex model (perfect Taylor series-of-fit) to the simplest model (linear model).

If the sample size n is at least 2, then it can be shown that there is a unique, computable minimizer for the above criterion, which we may denote by $\widehat{m}_{cspline}$, which is a *cubic spline* on the interval $[x_{(1)}, x_{(n)}]$, where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the n ordered statistics of the x observations $x_1, x_2, \ldots, x_n$. Being a cubic spline, the estimator $\widehat{m}_{cspline}(x)$ has the following properties: (i) it has a *continuous first derivative* everywhere, (ii) it is *linear* for $x < x_{(1)}$ and $x > x_{(n)}$, and (iii) it is a cubic function between each successive pair of the ordered values of x. To compute this cubic spline, note that $\widehat{m}_{cspline}(x)$ can be written as a weighted average of the

y-values, namely

$$\widehat{m}_{cspline}(x) = \sum_{i=1}^{n} w_i(x, \lambda) y_i, \tag{2.198}$$

where the weights w_i are appropriately defined.

2.6.2.3 Approximation and computation of the cubic spline

Silverman (1984) pointed out that the smoothing spline can be expressed as a kernel regression estimator with a variable bandwidth. In particular, when away from the boundary and with λ relatively small, the cubic spline estimator can be written as

$$\widehat{m}_{cspline}(x) = \sum_{i=1}^{n} w_i(x) y_i. \tag{2.199}$$

Here the weights $w_i(x) = w_i(x, \lambda; x_1, x_2, \ldots, x_n)$ can be approximated by

$$w_i(x) \approx \frac{1}{f(x_i)} \frac{1}{nb(x_i)} K\left(\frac{x_i - x}{b(x_i)}\right) \tag{2.200}$$

where b is a "local bandwidth" $b(x_i)$ at $x = x_i$ that depends on the design density $f(x_i)$ at $x = x_i$, the sample size n, and the smoothing parameter λ, whereas K is a symmetric kernel. Specifically,

$$b(x_i) = \left[\frac{\lambda}{nf(x_i)}\right]^{1/4} \tag{2.201}$$

and

$$K(u) = \frac{1}{2} exp\left(-\frac{|u|}{2}\right) sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right). \tag{2.202}$$

2.6.2.4 Selection of the optimum smoothing parameter λ

The next problem then is to find the optimal λ depending on the unknown regression function m as well as the error variance. One option is to minimize the squared error loss

$$L(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [m(x_i) - \hat{m}_\lambda(x_i)]^2. \tag{2.203}$$

In order to make a data-driven choice for λ, L must be approximated. This is possible in particular due to the generalized cross-validation (GCV) criterion due to Craven and Wahba (1979):

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\mathrm{RSS}(\hat{m}_\lambda) \Big/ \left(1 - \frac{1}{n}\sum_{i=1}^{n} w_i(x_i)\right)^2 \tag{2.204}$$

where RSS is the *residual sum of squares*, namely

$$\mathrm{RSS}(\hat{m}_\lambda) = \sum_{i=1}^{n} [\, y_i - \hat{m}_\lambda(x_i)]^2. \tag{2.205}$$

It turns out that $\mathrm{GCV}(\lambda)$ is an estimator for $L(\lambda) + \sigma^2$, so that minimizing $\mathrm{GCV}(\lambda)$ is equivalent to minimizing $L(\lambda)$.

Other options for finding the optimal λ includes the *cross-validation* criterion, which may turn out to be computationally intensive. There one minimizes

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \{y_i - \hat{m}_{\lambda,-i}(x_i)\}^2 \tag{2.206}$$

where $\hat{m}_{\lambda,-i}(x)$ is the leave-one-out cubic spline estimator of $m(x)$ excluding observation number i, i.e., it is computed excluding the pair (x_i, y_i). Finally, $\hat{m}_{\lambda,-i}(x_i)$ is the leave-one-out estimator $\hat{m}_{\lambda,-i}$ *evaluated at* x_i.

2.6.3 Kernel efficiency

As in the case of density estimation, minimizing the asymptotic expression for the mean squared error with respect to the kernel K leads to the Epanechnikov kernel as being the optimum choice (see Epanechnikov 1969 and Hodges and Lehmann 1956). This kernel is named after Epanechnikov due to its first use and derivation in the case of density estimation. Silverman (1986) tabulates efficiency of other kernels compared to the Epanechnikov kernel:

$$\begin{aligned} K_{epane}(u) &= (3/4)(1 - u^2/5)/\sqrt{5},\ |u| < \sqrt{5} \\ &= 0,\ \text{otherwise.} \end{aligned} \tag{2.207}$$

He defines *efficiency* to be the quantity (see Silverman 1986, p. 42)

$$eff(K) = (C(K_{epane})/C(K))^{5/4} \tag{2.208}$$

where C is defined in (2.172). To obtain the formula for the "best" kernel, an approach is to minimize $AMSE(x)$ or $AMISE$ over all kernels $K(\cdot)$ that satisfy the conditions:

(1) $\int K(u)du = 1$, (2)$K(u) \geq 0$, (3) $\int uK(u)du = 0$, and (4)$\int u^2 K(u)du = 1$.

The resulting kernel is given by formula (2.207); see Hodges and Lehmann (1956). Thus all other things remaining fixed, the kernel K_{epane} leads to the minimum asymptotic integrated mean squared error. In particular, the *efficiency* of any other kernel K can be defined in terms of the ratio

$$\text{eff}(K) = \left[\frac{C(K_{epane})}{C(K)}\right]^{5/4} = \frac{3}{5\sqrt{5}}\left[\int u^2 K(u)du\right]^{-1/2} \times \left[\int K^2(u)du\right]^{-1} \tag{2.209}$$

Some examples are given in Silverman (1986, p. 43), whose calculations using some popular kernels show that the above efficiency is more than 90%. These commonly used kernels are:

- Uniform:

$$K(u) = 1/2,\ |u| < 1,\ K(u) = 0,\ \text{otherwise.}$$

- Gaussian:

$$K(u) = (1/\sqrt{2\pi})exp(-u^2/2),\ u \in \mathbb{R}.$$

- Triangular:

$$K(u) = 1 - |u|,\ |u| < 1,\ K(u) = 0,\ \text{otherwise.}$$

- Biweight:

$$K(u) = (15/16)(1 - u^2)^2,\ |u| < 1,\ K = 0,\ \text{otherwise,}$$

etc. This means that the choice of the kernel is rather to be governed by other considerations such as differentiability and existence of moments.

3

Trend Estimation

3.1 Time series replicates

Consider the trend estimation problem when the observations are serially correlated, i.e., when one has time series data. Such data are very common in many areas of applications, and examples include climate, geophysics, ecology, engineering, medicine, economics, and so on. A time series is generated when a process is monitored over time and observations are recorded. The most important feature of such data is that the observations are serially correlated. In some cases, the observed series may be treated as a time series, as, for instance, in the case of observations from deep ice cores or from a horizontal track in an ecological study.

More generally, several series may be available, as, for instance, in Figure 3.2a–c, and the problem may be to obtain an estimate of the common trend. Thus suppose that k series are available, the length of each series being n. The problem is estimation of the common mean. When one has exactly one series, then $k = 1$ (see, for instance, Figure 3.1). To formulate the problem, one writes down a nonparametric regression model, where the regression function is the trend function of interest. Statistical properties of the error term are decisive of the properties of the estimated trend.

For simplicity of this discussion, we let the error processes for the ith individual series ($i = 1, \ldots, k$) be stationary. Needless to say, more complex cases may be envisaged. In addition, we let the k series be independent. Obviously, there may be situations where this is not the case, for instance if the series are from

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

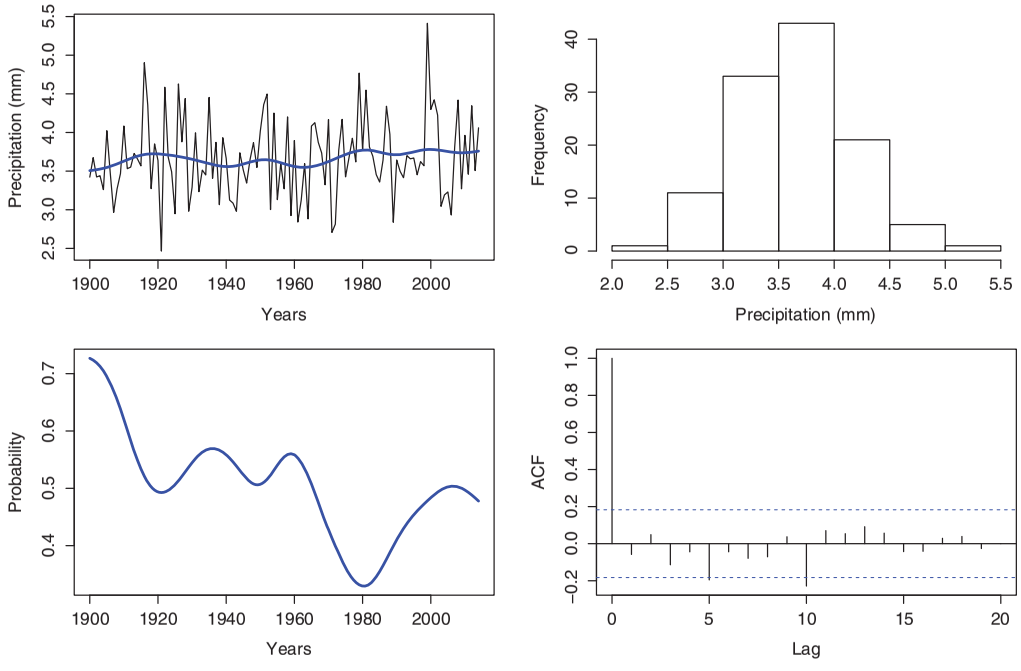

Figure 3.1 Yearly means of daily precipitation totals (05:40 am–05:40 am following day in mm) in Arosa (1840 m asl; 770730/183320, canton Graubünden), Switzerland: (top left) time series and estimated trend function, (top right) histogram of the raw precipitation values, (below left) probability of not exceeding the sample median (3.64 mm), and (below right) autocorrelation function of the residuals (detrended precipitation series). *Source:* Data from MeteoSchweiz, Switzerland.

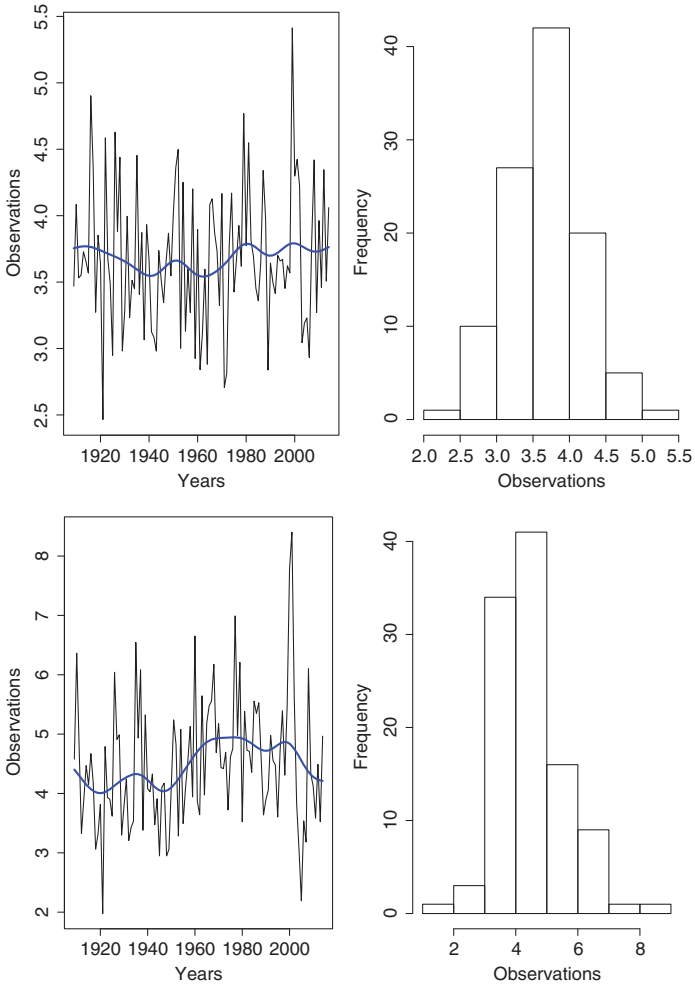

Figure 3.2(a) Yearly means of daily precipitation totals (05:40 am–05:40 am following day in mm) in climate stations in Switzerland. Time series and estimated trend function and histogram of the observations. Top: Arosa (1840 m asl; 770730/183320, canton Graubünden). Bottom: Bernina Pass (2307 m asl; 798660/143180, canton Graubünden). *Source:* Data from MeteoSchweiz, Switzerland.

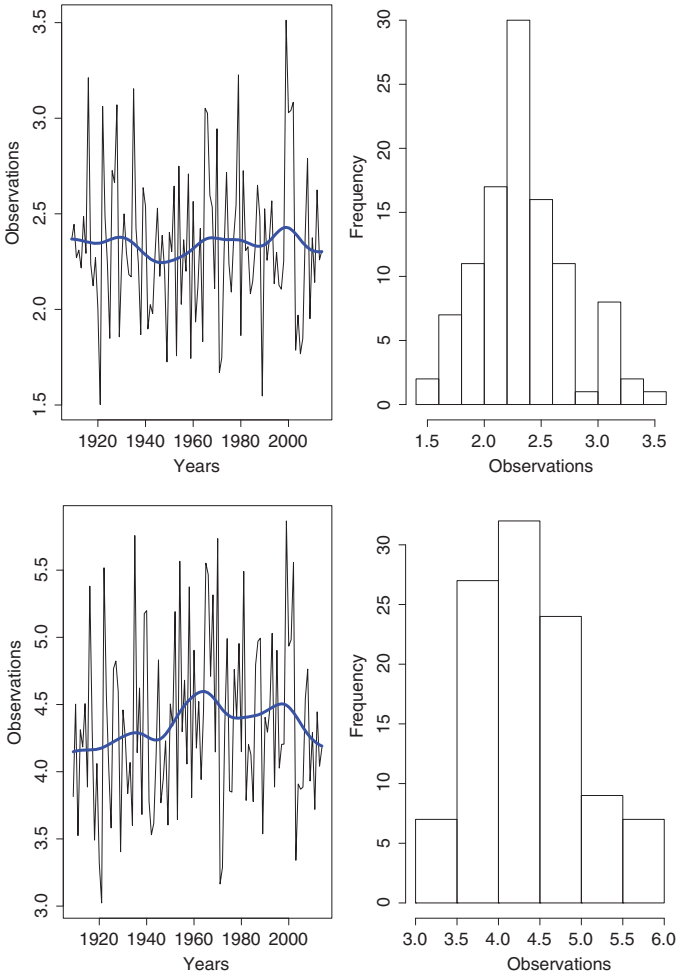

Figure 3.2(b) Continued from Figure 3.2(a): Top: Chur (556 m asl; 759471/193157, canton Graubünden). Bottom: Elm (958 m asl; 732265/198425, canton Glarus). *Source:* Data from MeteoSchweiz, Switzerland.

nearby locations. In such a case, the underlying model would have to accommodate (spatial) correlations among the k time series.

We look at an ANOVA type model (see Equation 3.2 later) for the replicated series and assume that the additive effects m_i sum up to zero. The common trend m is then estimated by

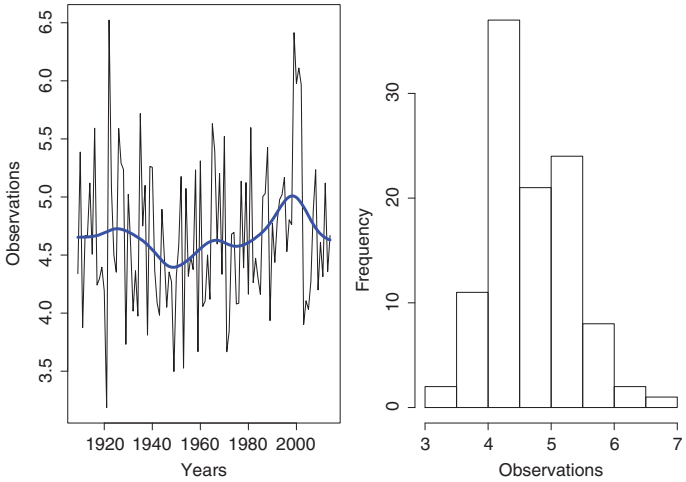

Figure 3.2(c) Continued from Figure 3.2(a) and (b): Averages of the yearly values from the four stations. Time series with estimated trend and histogram of the data. *Source:* Data from MeteoSchweiz, Switzerland.

smoothing the average of the k series. As in the previous chapter, we apply kernel smoothing to the averaged series. We then examine the properties of the trend estimator under varying assumptions about the correlations structure in the data. In particular, we consider the situation where the errors of each individual series is stationary and has short-memory, long-memory, or anti-persistence correlations. These three correlation types lead to varying rates of convergence of the variance of the sample mean of a stationary process u_i to zero. This is determined by the limiting behavior of $S_n = \sum_{k=-n}^{n} \mathrm{cov}(u_i, u_{i+k})$, as $n \to \infty$; (see Beran et al. 2013 for additional information; also see in particular, Beran and Ocker 1999). Asymptotically optimal bandwidth and mean integrated squared error are then derived where the length of each series n tends to infinity.

The trend estimator mentioned here is based on the averaged series, $\bar{y}_j, j = 1, 2, \ldots, n$. Alternatively, one can estimate the trend in each individual series and then take the average of the trend estimates from the various replicates. For further remarks see Ghosh (2001), who also presents some simulation studies.

As for estimation issues with replicated time series, Diggle and Wasel (1997) address spectral estimation. In contrast, Ghosh

(2001) considers the problem of nonparametric estimation of a common trend function. We discuss these results here, keeping in mind that the approach can be extended to more complex cases such as when the errors are non-stationary (e.g., in a time-dependent Gaussian subordinated model) or when the data are spatially correlated (e.g., relaxing the assumption of independence of the replicates).

In our $k \geq 1$ series, we let the within-series correlation structures be governed by the fractional differencing parameters $\delta_i \in (-1/2, 1/2)$, corresponding to anti-persistence ($\delta_i < 0$), short-range dependence ($\delta_i = 0$), or long-range dependence ($\delta_i > 0$). At the next step, a nonparametric estimate of the common trend is defined, followed by a discussion involving derivation of the formulas for the optimal bandwidth and optimal mean integrated squared error.

In order to proceed with estimation of the common trend function, we need to impose smoothness conditions. There are many examples where the trend is smooth. Such trends can be deterministic or stochastic. In this note, we address the case of deterministic trends. However, there may be trend-like patterns created by the random fluctuations of the time series observations around the deterministic trend function. This can happen, for instance, when there are slowly decaying long-term correlations in the data. When there is long-range dependence, trend estimation can be difficult because these slowly decaying correlations contribute to a slower rate of convergence of kernel estimates; see, for instance, Hall and Hart (1990) and Csörgő and Mielniczuk (1995).

Fractional autoregressive processes as in Equation (3.1) (see Granger and Joyeux 1981 and Hosking 1980) are examples of stationary models that exhibit spurious trend-like behavior in the data:

$$\phi_p(B)(1 - B)^{\delta} u_i = v_i. \tag{3.1}$$

Here u_i is a zero-mean stationary stochastic process, B denotes the backshift operator, so that $B^r u_i = u_{i-r}$, where r is an integer, ϕ_p is a polynomial of degree $p = 0, 1, 2, \ldots$, with its roots outside the unit circle, and $v_i, i = 1, 2, \ldots, n$ are iid $N(0, \sigma^2)$ random variables, and, finally, $-0.5 < \delta < 0.5$ is known as the fractional

differencing parameter and the trend function m is a smooth function ($m \in \mathbb{C}^2[0,1]$). Varying δ, different correlation types can be obtained. For instance, $p = 0, \delta = 0$ implies $u_i \sim$ iid $N(0, \sigma^2)$, $\delta > 0$ implies u_i has long-memory, and $\delta < 0$ implies u_i is anti-persistent. For detailed information on these correlation types see Beran et al. (2013). Such a process is often used as a model for the error in a nonparametric regression model with time series data. An example of a simulated fractional $ARIMA(0, \delta, 0)$ series is shown in Figure 3.3a,b, along with its estimated spectral density function, namely the periodogram. The time series is simulated using the function *fracdiff.sim* in the R-package *fracdiff*. In this example, $\delta = 0.3$ is chosen so that the series exhibits long-memory. The pole at zero can be seen in both plots of the periodogram.

3.1.1 Model

We consider the following nonparametric regression model for replicated time series data. Let $y_{i,j}, i = 1, 2, \dots, k, j = 1, 2, \dots, n$ denote evenly spaced time series observations from $k \geq 1$ replicates, j denoting time and $t_j = j/n$ denoting rescaled time:

$$y_{i,j} = m(t_j) + m_i(t_j) + u_{i,j}, t_j = j/n, m, m_i \in C^2[0,1]. \quad (3.2)$$

Moreover, let

$$\sum_{i=1}^{k} m_i(t) = 0, t \in [0,1], \quad (3.3)$$

i.e., the additive effects m_i that model the deviations of individual trends from the overall trend m in the different replicates add up to zero. Of course, this simplified assumption may be changed in a more complex model comprising interactions. The ith error process $u_{i,j}$ is assumed to be a stationary zero mean process with finite second moments. Let the k series be independent of each other. Also let $u_{i,j}, j = 1, 2, \dots, n$, have a spectral density

$$f_i(\lambda) \sim C_i |\lambda|^{-2\delta_i} \quad (3.4)$$

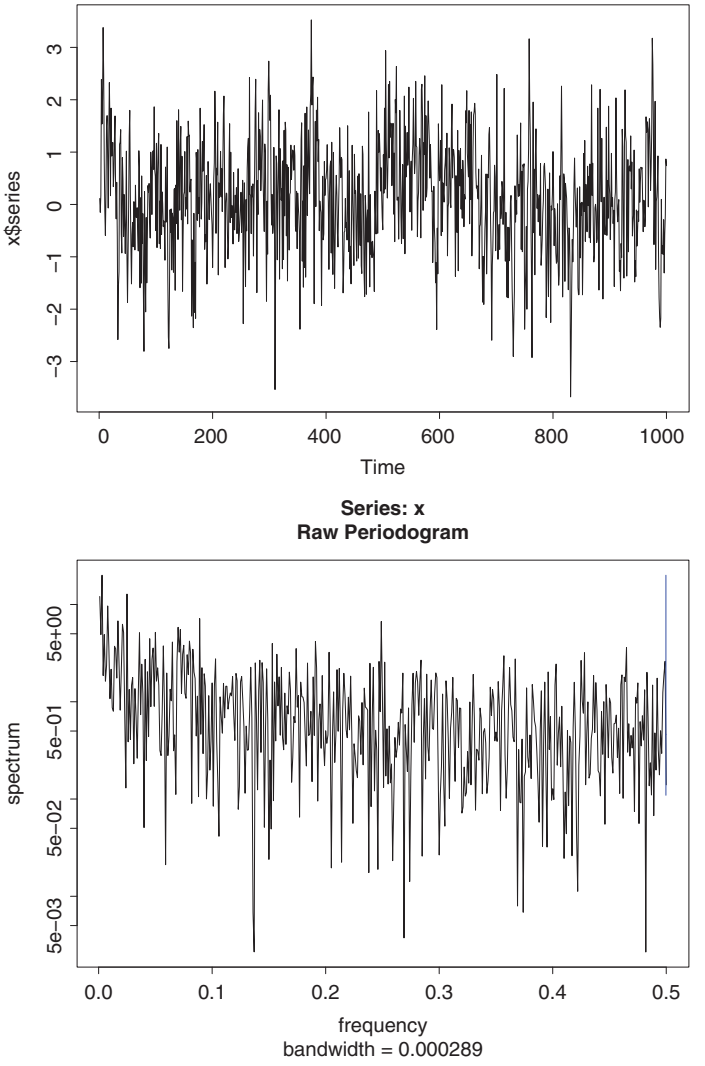

Figure 3.3(a) Simulated fractional *ARIMA*(0, 0.3, 0) series with $n = 1000$ observations. The pole at zero in the periodogram indicates long-memory. The R-package *fracdiff* is used for simulating the time series. Top: simulated series with $n = 1000$ observations; Bottom: periodogram.

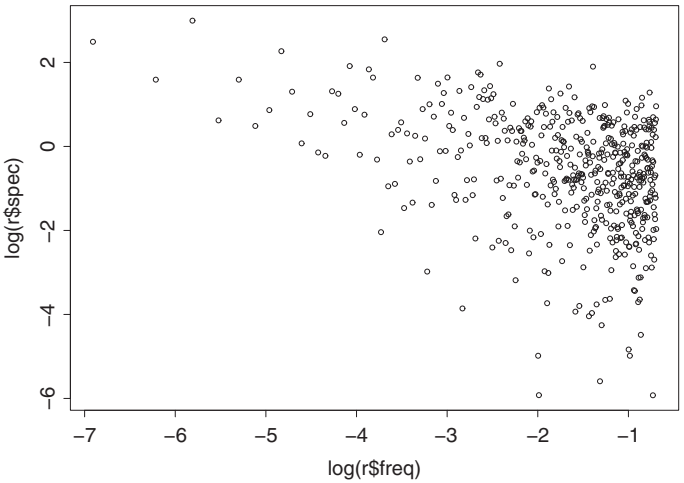

Figure 3.3(b) Continued from Figure 3.3(a). Periodogram of simulated FARIMA series, log(periodogram) plotted against log(frequency).

as $\lambda \to 0$. Here, $\delta_i \in (-1/2, 1/2)$ and C_i is positive. Then the covariances decay hyperbolically as

$$\gamma_i(h) = \mathbb{C}ov(u_{i,j}, u_{i,l}) \sim D_i |h|^{2\delta_i - 1} \tag{3.5}$$

as $h \to \infty$ for $\delta_i \neq 0$, where

$$D_i = \frac{sin(\pi\delta_i)\Gamma(1 - 2\delta_i)}{(1 + 2\delta_i)} C_i. \tag{3.6}$$

In (3.4) to (3.6), if $\delta_i > 0$, then the spectral density f_i has a pole at zero and the sum of the autocorrelations diverges, i.e., u_i has long-range dependence. On the other hand, if $\delta_i = 0$, then the spectral density at zero is finite and the sum of all autocorrelations is finite and non-zero. In this case, u_i has short-range dependence. Finally, if $\delta_i < 0$, then the spectral density at zero is zero and the sum of all autocorrelations is equal to zero; i.e., the negative and positive autocorrelations cancel each other. In this case, u_i is anti-persistent.

3.1.2 Estimation of common trend function

A kernel smoothed estimate of m may be given by

$$\widehat{m}(t) = \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{t_j - t}{b}\right) \bar{y}_j \tag{3.7}$$

where $t \in (0, 1)$ and $\bar{y}$ is the average of the time series observations from all replicates, i.e.,

$$\bar{y}_j = k^{-1} \sum_{i=1}^{k} y_{i,j} = m(t_j) + \bar{u}_j$$

and, similarly, $\bar{u}_j = k^{-1} \sum_{i=1}^{k} u_{i,j}$.

Usually, general kernels such as those mentioned in Chapter 2 may be considered. Typically the kernel K will be a symmetric pdf, satisfying some moment conditions and additional regularity conditions. For simplicity of our discussion here, we consider the rectangular kernel on the compact interval $(-1, 1)$. Thus

$$K(u) = \frac{1}{2} 1\{-1 \le u \le 1\}. \tag{3.8}$$

The bandwidth b, on the other hand, is such that $b \to 0$ and $nb \to \infty$ as $n \to \infty$.

3.1.3 Asymptotic properties

3.1.3.1 Conditional mean squared error: exact expression

To start with, consider the exact expressions for bias, variance, and the mean squared error for $\widehat{m}$. Let K be as in (3.8) and consider $\widehat{m}$ defined in (3.7). The expression for bias follows from definition so that

$$bias = B_{n,k}(t) = \mathbb{E}(\widehat{m}(t) - m(t)) = \frac{1}{2nb} \sum_{j=n(t-b)}^{n(t+b)} m(t_j) - m(t). \tag{3.9}$$

As for the variance, noting that

$$\sum_{i,j=1}^{n} a(i-j) = \sum_{h=-(n-1)}^{(n-1)} (n - |h|) a(h) \tag{3.10}$$

for a function $a(\cdot)$, and since

$$\mathbb{C}ov(\bar{u}_j, \bar{u}_l) = \frac{1}{k^2} \sum_{i=1}^{k} \gamma_i(j-l), \tag{3.11}$$

we have

$$\mathbb{V}ar(\widehat{m}(t)) = \frac{1}{(2nbk)^2} \sum_{i=1}^{k} \sum_{j,l=n(t-b)}^{n(t+b)} \gamma_i(j-l) \tag{3.12}$$

$$= \frac{1}{(2nbk)^2} \sum_{i=1}^{k} \sum_{j',l'=1}^{2nb+1} \gamma_i(j'-l'). \tag{3.13}$$

The last expression follows by substituting

$$j' = j - n(t-b) + 1 \text{ and } l' = l - n(t-b) + 1. \tag{3.14}$$

Then applying (3.10), the exact expressions for the variance is

$$\begin{aligned} variance \ &= V_{n,k,\delta_1,\ldots,\delta_k} = \frac{1}{k^2(2nb)^2} \sum_{i=1}^{k} \sum_{h=-2nb}^{2nb} v(i,h) \\ \text{where } v(i,h) &= (2nb + 1 - |h|)\gamma_i(h). \end{aligned} \tag{3.15}$$

The mean squared error at fixed t is then by definition

$$mse(t) = B^2_{n,k}(t) + V_{n,k,\delta_1,\ldots,\delta_k} \tag{3.16}$$

Finally, the mean integrated squared error (MISE) is calculated by integrating the $mse(t)$ over $(\beta, 1-\beta)$ where $0 < \beta < 1/2$:

$$MISE_{\{\delta_1,\ldots,\delta_k\}} = \int_{\beta}^{1-\beta} (B_{n,k}(t))^2 dt + (1-2\beta) V_{n,k,\delta_1,\ldots,\delta_k}. \tag{3.17}$$

For fixed n and k, given $\delta_1, \ldots, \delta_k$, the MISE is thus given by

$$\begin{aligned} MISE_{\{\delta_1,\ldots,\delta_k\}} = &\int_{\beta}^{1-\beta} \left(\frac{1}{2nb} \sum_{j=n(t-b)}^{n(t+b)} m(t_j) - m(t) \right)^2 dt \\ &+ (1-2\beta) \frac{1}{(2nbk)^2} \sum_{i=1}^{k} \sum_{h=-2nb}^{2nb} v(i,h) \end{aligned} \tag{3.18}$$

3.1.3.2 Spectral density and covariance of $\bar{u}$

As regards the ith stationary process $\{u_{i,j}, j = 1, 2, \ldots, n\}$, f_i denotes its spectral density with the fractional differencing parameter $\delta_i, i = 1, 2, \ldots, k$. The behavior of f_i at the origin is of the type

$$f_i((\lambda) \sim C_i|\lambda|^{-2\delta_i}, \lambda \to 0. \tag{3.19}$$

The autocovariances of $\bar{u}_j, j = 1, 2, \ldots, n$, are given by

$$\begin{aligned}
&\gamma(|j-l|) = \mathbb{C}ov(\bar{u}_j, \bar{u}_l) \\
&= \mathbb{C}ov\left(\frac{1}{k}\sum_{i=1}^{k} u_{i,j}, \frac{1}{k}\sum_{i=1}^{k} u_{i,l}\right) = \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{C}ov(u_{i,j}, u_{i,l}) \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\gamma_i(|j-l|).
\end{aligned} \tag{3.20}$$

The corresponding spectral density is

$$f_{\bar{u}}(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma(h)e^{ih\lambda} = \frac{1}{k^2}\sum_{i=1}^{k} f_i(\lambda). \tag{3.21}$$

3.1.3.3 Asymptotic rates for bias, variance, and optimal bandwidth

To derive asymptotic results, we make a note of the following:

- We have a finite number of replicates, i.e., k is fixed and finite.
- The fractional differencing parameters are non-random, i.e., $\delta_1, \delta_2, \ldots, \delta_k$ are fixed. When $\delta_i = 0$, γ_i satisfies

$$\sum_{h=-n}^{n}|h|\gamma_i(h) = O(n), n \to \infty. \tag{3.22}$$

For other combinations of fixed or random δ's versus a finite or infinite number of replicates, see Ghosh (2001).

Note that if δ is the largest fractional differencing parameter, i.e., $\delta = max\{\delta_1, \delta_2, \ldots, \delta_k\}$ then it is also the fractional differencing parameter for the sample mean process, i.e.,

$$f_{\bar{u}}(\lambda) \sim \frac{1}{k^2}C|\lambda|^{-2\delta}\ as\ |\lambda| \to 0, \tag{3.23}$$

where $C = \sum_{i:\delta_i=\delta} C_i$ and the C_i are defined in (3.4). Now define

$$\nu(\delta) = \frac{2^{2\delta}\Gamma(1-2\delta)sin(\pi\delta)}{\delta(2\delta+1)} \tag{3.24}$$

for $\delta \neq 0$, and

$$\nu(0) = \lim_{\delta\to 0} \nu(\delta) = \pi \tag{3.25}$$

(see Beran and Ocker 1999). In what follows, we discuss the role of the type of dependence in the convergence rate for the trend estimate. Under short memory (i.e. $\delta = 0$), the global optimal bandwidth b_{opt} is of the order $n^{-1/5}$ and the AMISE is of the order $n^{-4/5}$, which are the same rates for independent observations. In case of long-range dependence, on the other hand (i.e., when $\delta > 0$), $\widehat{m}$ converges to g at a slower rate. Finally, anti-persistence ($\delta < 0$) implies a faster rate of convergence. Similar results have been obtained for $k = 1$ elsewhere. See, for instance, Chiu (1989), Altman (1990), Herrmann, Gasser, and Kneip (1992), Hall and Hart (1990), Csörgő and Mielniczuk (1995), and Beran and Ocker (1999).

When k is fixed, Taylor series expansion along with the properties of the kernel and the bandwidth b can be used to show that as $n \to \infty$ the square of the bias converges to

$$\frac{b^4}{36}\int_{\beta}^{1-\beta} (g^{(2)}(t))^2 dt + o(b^4). \tag{3.26}$$

On the other hand, the integrated variance of the estimator (second term in the AMISE) equals

$$(1-2\beta)\frac{1}{k^2}\sum_{i=1}^{k}(a_{n,i} + b_{n,i} + c_{n,i}), \tag{3.27}$$

where

$$\begin{aligned} a_{n,i} &= \frac{1}{(2nb)^2}\sum_{u=-2nb}^{2nb} 2nb\gamma_i(u), \\ b_{n,i} &= \frac{1}{(2nb)^2}\sum_{u=-2nb}^{(2nb)} \gamma_i(u), \text{and} \\ c_{n,i} &= \frac{1}{(2nb)^2}\sum_{u=-2nb}^{2nb} |u|\gamma_i(u). \end{aligned} \tag{3.28}$$

We now examine the above terms under long-memory, short-memory, and anti-persistence.

Long-memory: $0 < \delta < 0.5$: In this case, since $2\delta_i - 1 > -1$ and $nb \to \infty$,

$$\lim_{nb\to\infty} \sum_{u=-2nb}^{2nb} \gamma_i(u) = \infty, \tag{3.29}$$

so that

$$\sum_{u=-2nb}^{2nb} \gamma_i(u) - \gamma(0) \sim 2D_i \sum_{u=1}^{2nb} |u|^{2\delta_i - 1} \sim (2nb)^{2\delta_i} 2D_i q(i)$$
$$\text{where, } q(i) = \int_o^1 x^{2\delta_i - 1} dx. \tag{3.30}$$

This implies

$$a_{n,i} = \frac{D_i}{\delta_i}(2nb)^{2\delta_i - 1} + o((nb)^{2\delta_i - 1}). \tag{3.31}$$

Clearly, $b_{n,i} = o(a_{n,i})$. Finally, since

$$\gamma_i(u) \sim D_i |u|^{2\delta_i - 1}, \text{ as } u \to \infty, \tag{3.32}$$

$\sum_{u=-2nb}^{2nb} |u|\gamma_i(u)$ diverges so that

$$c_{n,i} \sim \frac{1}{(2nb)^2} 2D_i \sum_{u=1}^{2nb} u^{2\delta_i} \tag{3.33}$$

which leads to

$$c_{n,i} = \frac{2D_i}{(2\delta_i + 1)}(2nb)^{2\delta_i - 1} + r_{i,n} \tag{3.34}$$

where $r_{i,n} = o((nb)^{2\delta_i - 1})$.

Antipersistence: $-0.5 < \delta < 0$: In this case,

$$\sum_{u=-\infty}^{\infty} \gamma_i(u) = 0 \tag{3.35}$$

so that

$$\sum_{u=-2nb}^{2nb} \gamma_i(u) = -2 \sum_{u=2nb+1}^{\infty} \gamma_i(u). \tag{3.36}$$

Also, $\sum_{u=-\infty}^{\infty} |u|\gamma_i(u)$ diverges. Using arguments as for the long-memory case, $a_{n,i}$, $b_{n,i}$ and $c_{n,i}$ have the same limits.

Short-memory: $\delta = 0$: In this case, applying the relation between γ and the spectral density f,

$$\sum_{u=-\infty}^{\infty} \gamma_i(u) = 2\pi f_i(0) \tag{3.37}$$

where $f_i(0) = C_i$. This means

$$a_{n,i} \sim \frac{2\pi C_i}{2nb}, b_{n,i} = o(a_{n,i}), \text{and } \; c_{n,i} = o(1/(nb)). \tag{3.38}$$

We may summarize the above findings. Let $n \to \infty$, so that $b \to 0$ and $nb \to \infty$. Let $\delta = max\{\delta_1, \delta_2, \ldots, \delta_k\}$ and $C_\delta = \sum_{i:\delta_i=\delta} C_i$, where C_i is defined in (3.4) and ν is defined in (3.24).

- The asymptotic expression for the MISE is given by

$$MISE = AMISE + r_n, \tag{3.39}$$

where the leading term is

$$\begin{aligned} AMISE &= \frac{b^4}{36} \int_\beta^{1-\beta} (g^{(2)}(t))^2 dt \\ &\quad + \frac{1}{k^2}(1-2\beta) \sum_{i=1}^{k} (nb)^{2\delta_i - 1} \nu(\delta_i) C_i \\ &= \frac{b^4}{36} \int_\beta^{1-\beta} (g^{(2)}(t))^2 dt \\ &\quad + \frac{1}{k^2}(1-2\beta)(nb)^{2\delta-1} \nu(\delta) C_\delta \end{aligned} \tag{3.40}$$

and the remainder is

$$\begin{aligned} r_n &= o(b^4) + \frac{1}{k^2} \sum_{i=1}^{k} o((nb)^{2\delta_i - 1}) \\ &= max(b^4, (nb)^{2\delta-1}) \end{aligned} \tag{3.41}$$

The global optimal bandwidth is then obtained by minimizing the AMISE with respect to b. We have

$$b_{opt} = \left[\frac{9(1-2\beta)(1-2\delta)v(\delta)C_\delta}{\int_\beta^{1-\beta}(g^{(2)}(t))^2 dt} \right]^{1/(5-2\delta)} n^{(2\delta-1)/(5-2\delta)} k^{-2/(5-2\delta)}. \tag{3.42}$$

Substituting $b = b_{opt}$, the optimal rate of *AMISE* can be obtained. This rate is of the order $O(n^{(8\delta-4)/(5-2\delta)} k^{-8/(5-2\delta)})$.

3.2 Irregularly spaced observations

When time series observations are irregularly spaced in time and trend estimation is of interest, one option is to consider a continuous index stochastic process as a model for the error in the nonparametric regression model.

In applications, such data may occur due to missing observations or due to the nature of the specific area of investigation, as, for instance, in the palaeo sciences such as, palaeo climate (e.g., stable oxygen isotopes or other gases in the ice cores of the polar regions; e.g. Johnsen et al. 1997 and Schwander et al. 2000) or palaeo ecology (e.g. Tinner and Lotter, 2001). In the palaeo sciences, fossil data are obtained from deep stratigraphic cores or ice cores, going as deep as several hundred to several thousand meters below ground. For the purpose of scientific interpretation, depth is transformed to the age of the sample (years before present) using various dating and calibration techniques. However, due to the nonlinearity of the age–depth relation, a reason for which being compaction of material over time, the age of the samples are unevenly distributed in the core, even though the points in the depth axis, where data summaries are recorded, are equidistant. Time series so obtained can thus be seen as being unevenly distributed in time.

In some situations, the time series may cover very long time spans. In the palaeo sciences, for instance, the entire span may cover several thousand years. Therefore, the assumption that the marginal distribution of the error remains unchanged over very long spans of time may be doubtful. A flexible way to incorporate changing marginal distributions is to consider a time-dependent

Gaussian subordination model. This leads to having a very rich class for the marginal distributions of the underlying stochastic process. Here we address trend estimation when the data are Gaussian subordinated and the stochastic process is a continuous indexed process.

For further information and references to irregularly spaced time series data analysis, see Parzen (1984), West (1994), Haslett et al. (2006), and references therein.

For properties of empirical processes arising from nonlinear functionals of stationary Gaussian processes see Csörgő and Mielniczuk (1996), Major (1981) and Taqqu (1975, 1979). For background information on probabilistic and statistical aspects of long-memory processes see Beran (1994), Doukhan et al. (2003), Embrechts and Maejima (2002), and Künsch (1986), among others.

For trend estimation of time series of long-memory processes see, for instance, Hall and Hart (1990), Csörgő and Mielniczuk (1995), Ray and Tsay (1997), and Beran and Feng (2002), as well as the treatment of this problem in Section 3.1.

Suppose that we have a data set where these conditions mentioned above apply, the trend function having a complex shape so that an adequate formulation of the trend in terms of a finite number of parameters is either not possible or difficult to guess. Smoothing of the observed time series may then be an option for estimating the underlying trend, and here we address this problem. See Menéndez et al. (2013) for additional information and data examples. Let us assume that

- The nonparametric regression errors are Gaussian subordinated (Taqqu 1975), being an unknown transformation G of a latent Gaussian process Z for every fixed time t.
- For every fixed t, the unknown transformation is monotonically increasing. The reason for considering monotonicity is simplicity. Even if the transformation is monotone, the resulting class of the marginal distributions is still very broad. Monotonicity of the transformation means the Hermite rank of the transformed process is 1. The Hermite rank of a process $G(Z)$ is l, if l is the smallest positive integer such that

$$\int_{-\infty}^{\infty} G(z)H_l(z)\exp(-z^2/2)dz \neq 0. \tag{3.43}$$

It is easy to see that if $G(\cdot, t)$ is monotone increasing for all t, then by partial integration (where the Hermite polynomial of degree l is H_l and for $l = 1, H_1(z) = z$)

$$\int_{-\infty}^{\infty} G(z) z \exp(-z^2/2) dz = \int_{-\infty}^{\infty} G^{(1)}(z) \exp(-z^2/2) dz > 0. \tag{3.44}$$

If the monotone function G is not differentiable,

$$\int_{-\infty}^{\infty} G^{(1)}(z) \exp(-z^2/2) dz = \int_{-\infty}^{\infty} \exp(-z^2/2) dG(z). \tag{3.45}$$

- The regression errors may have long-range dependence. The reason for specifically addressing this correlation type is the fact that many geophysical applications have examples of time series with long-range dependence. See Beran (1994) and Beran et al. (2013) for examples.

The Hermite rank of the error process plays an important role. In particular, it affects asymptotic formulas such as the mean square error of the curve estimate. These formulas are later used to obtain a data-driven bandwidth selection algorithm, requiring estimation of unknown parameters and functions appearing in the mean squared error. When the function G is monotone, not only the Hermite rank is known there are also other advantages. For instance, estimation of the first Hermite coefficient is facilitated.

The trend estimation problem addressed here generalizes to other related problems, such as nonparametric estimation of the distribution function. One defines a suitable indicator function, which is then smoothed to nonparametrically estimate the cdf. The same theory as developed in the main theorems of this section applies also here, because the Gaussian subordination assumption also extends to the indicator functions.

3.2.1 Model

As mentioned above, in the case of irregularly spaced time series data, we envisage a zero mean continuous index stochastic

process $u(T)$, $T \in \mathbb{R}_+$. Let the observations be available at time points $T_1, T_2, \ldots, T_n$, n denoting the sample size, and let

$$y_i = y(T_i) \tag{3.46}$$

denote an observation number $i \geq 1$. Our interest lies in estimation of a smooth trend function $m(t) = \mathbb{E}(Y(T))$ appearing in (3.48), where t denotes rescaled time. Similarly, let

$$u_i = u(T_i). \tag{3.47}$$

In the jargon of the palaeo example considered earlier, T_i may be called the age of the sample at depth i. We thus have the following nonparametric regression model:

$$y_i = m(t_i) + u_i\,, i = 1, \ldots, n, \tag{3.48}$$

where $t_i = T_i/T_n$ are the rescaled times and $m(t), t \in \mathbb{C}^2[0, 1]$ is a smooth function. The regression errors u_i are centered, i.e., they have zero mean and finite variance. Let us assume that the variance of u_i may change smoothly as a function of time. In other words, let

$$\mathbb{V}ar(u_i) = \sigma^2(t_i) > 0. \tag{3.49}$$

3.2.1.1 Gaussian subordination

As for distributional assumptions, consider a continuous index zero mean and unit variance latent stationary Gaussian process $Z(T)$, $T \in \mathbb{R}_+$ such that the following is satisfied:

$$u_i = G(Z_i, t_i), \tag{3.50}$$

where $Z_i = Z(T_i)$. As mentioned earlier, we focus particularly on the case when $u(T)$, or, more specifically, $Z(T)$ has long-range dependence. However, similar results can also be derived under the assumptions of short-range dependence or anti-persistence.

As in the previous section, the assumption of long-range dependence means that the autocovariances decay slowly, and in particular hyperbolically, as follows:

$$\mathbb{C}ov(Z(T), Z(T + S)) \sim C_Z|S|^{2H-2}, S \to \infty, \tag{3.51}$$

where $\frac{1}{2} < H < 1$ is the long-memory parameter or the Hurst coefficient.

In equation (3.50),

$$G : \mathbb{R} \times [0,1] \to \mathbb{R} \tag{3.52}$$

is square integrable with respect to the normal density function. Specifically,

$$\mathbb{E}(G(Z,t)^2) < \infty, \mathbb{E}(G(Z,t)) = 0, \quad (Z \sim \mathcal{N}(0,1)) \tag{3.53}$$

for all $t \in [0,1]]$. In particular, for every fixed t, we let $G(\cdot, t)$ be monotonically increasing and left-continuous.

Since G is square-integrable with respect to the standard normal density function, G can be expanded (in the mean squared sense) using Hermite polynomials. In particular, we have

$$u_i = \sum_{l=q}^{\infty} \frac{c_l(t_i)}{l!} H_l(Z_i). \tag{3.54}$$

In Equation (3.54), c_l are Hermite coefficients, which we assume to be in $\mathbb{C}^2[0,1]$:

$$c_l(t_j) = \mathbb{E}(u_j H_l(Z_j)), t_j \in [0,1], \tag{3.55}$$

$j = 1, 2, \ldots, n, l \geq 1$, H_l are Hermite polynomials, l being the degree of the polynomial, and $q \geq 1$ is the Hermite rank of G, which, for simplicity, we assume to be a constant.

The Hermite polynomials have zero means and, for $l = l'$,

$$\mathbb{C}ov(H_l(Z_i), H_l(Z_j)) = l!(\mathbb{C}ov(Z_i, Z_j))^l, \tag{3.56}$$

whereas they are orthogonal when $l \neq l'$, i.e.,

$$\mathbb{C}ov\left(H_l(Z_i), H_{l'}(Z_j)\right) = 0. \tag{3.57}$$

Since $\mathbb{V}ar(Z_i) = 1$,

$$\mathbb{V}ar(H_l(Z_i)) = l!. \tag{3.58}$$

Moreover, due to (3.54), the autocovariances of the error process can be written as

$$\mathbb{C}ov(u_i, u_j) = \sum_{l=q}^{\infty} \frac{c_l(t_i)c_l(t_j)}{l!} (\mathbb{C}ov(Z_i, Z_j))^l \tag{3.59}$$

and, in particular,

$$\mathbb{V}ar(u_i) = \mathbb{V}ar(G(Z_i, t_i)) = \sum_{l=q}^{\infty} \frac{c_l^2(t_i)}{l!}. \tag{3.60}$$

Rewriting (3.59),

$$\begin{aligned}\mathbb{C}ov(u_i, u_j) = &\frac{c_q(t_i)c_q(t_j)}{(q)!}(\mathbb{C}ov(Z_i, Z_j))^q \\ &+ \sum_{l=q+1}^{\infty} \frac{c_l(t_i)c_l(t_j)}{l!}(\mathbb{C}ov(Z_i, Z_j))^l.\end{aligned} \tag{3.61}$$

Now, if $|T_i - T_j| \to \infty$ but $t_i, t_j \to t$, then applying (3.51), the first term in (3.61) dominates, i.e.,

$$\mathbb{C}ov(u_i, u_j) \sim C_Z^q c_q(t)^2 |T_i - T_j|^{(2H-2)q}. \tag{3.62}$$

Therefore, from (3.61) it is clear that if the long-memory parameter in the latent Gaussian process Z_i is H, then the errors u_i will also have long-memory if and only if

$$-1 < q(2H - 2) \tag{3.63}$$

or, equivalently

$$H > 1 - \frac{1}{2q}. \tag{3.64}$$

Also, since $G(\cdot, t)$ is assumed to be monotone increasing for all t, due to (3.44),

$$q = 1. \tag{3.65}$$

This property can be exploited further to develop a bandwidth selection algorithm. We discuss this in the sequel.

3.2.2 Derivatives, distribution function, and quantiles

In certain situations, derivatives of order $\nu \geq 0$ of the trend function m are needed. An example is estimation of rapid change points. Rapid change points are points of time where the first derivative of the trend function exceeds a prespecified threshold, with further technical conditions on the other derivatives of m. Another application is estimation of mode, where mode

is defined as the point of time where the first derivative of the function m is equal to zero.

In what follows we consider estimation of the νth derivative of the trend m. To estimate the trend itself, set $\nu = 0$.

Since the time points are irregularly spaced, consider the Priestley–Chao kernel estimator defined by

$$\widehat{m}^{(\nu)}(t) = \frac{(-1)^{\nu}}{b^{\nu+1}} \sum_{i=1}^{n} (t_i - t_{i-1}) K^{(\nu)} \left(\frac{t_i - t}{b} \right) y_i, \tag{3.66}$$

where we set $t_0 = 0$.

A related problem is to estimate threshold exceedance probabilities, in particular as a function of time. This translates to the estimation of distribution functions and quantiles. See Ghosh et al. (1997) and Ghosh and Draghicescu (2002b) for further examples. In particular, Ghosh and Draghicescu (2002b) address the nonparametric prediction problem; also see Beran and Ocker (1999) for related information.

For a given cut-off value $-\infty < y < \infty$, define the indicator process

$$I_i(y) = \begin{cases} 1, & y_i \le y \\ 0, & otherwise \end{cases} \tag{3.67}$$

where

$$\mathbb{E}(I_i(y)) = P(y_i \le y) = F(t_i; y). \tag{3.68}$$

Here F is the (marginal) cumulative probability distribution function of y_i at $t_i = T_i / T_n$. We assume that $F(t; y) : [0, 1] \times \mathfrak{R} \to [0, 1]$ is a smooth function of t for every fixed y. Since y is Gaussian subordinated so is I_i via the function $\tilde{G}$ say. Having a bounded variance allows for a Hermite expansion

$$I_i(y) - F(t_i; y) = \tilde{G}(Z_i, t_i) = \sum_{l=\tilde{q}}^{\infty} \frac{\tilde{c}_l(t_i, y)}{l!} H_l(Z_i), \tag{3.69}$$

where $\tilde{q}$ is the Hermite rank of $\tilde{G}$ and $\tilde{c}_l(t_i, y)$ are Hermite coefficients. We assume that $\tilde{q}$ is a constant for all y and t_i, although generalizations are possible. For estimation, smoothing of the

indicator function $I_i(y)$ leads to a nonparametric estimate of the non-exceedance probability $F(t;y)$, which is assumed to be continuous in both t and in y. Therefore we define the estimator

$$\widehat{F}(t;y) = \frac{1}{b}\sum_{i=1}^{n}(t_i - t_{i-1})K\left(\frac{t_i - t}{b}\right)I_i(y). \tag{3.70}$$

Finally, $\widehat{F}(t;y)$ may be inverted to obtain a consistent nonparametric estimate of the time-varying quantile function at time $t \in (0,1)$:

$$\widehat{Q}(\alpha;t) = \inf\{y \in \mathbb{R} | \widehat{F}(t;y) \geq \alpha\}, \alpha \in (0,1). \tag{3.71}$$

The quantile functions have various obvious applications. The extreme quantiles are of particular interest in studies related to extrema of processes. The interquartile range (IQR) function $\widehat{Q}(0.75;t) - \widehat{Q}(0.25;t)$ is a popular choice for a summary statistic and the above method provides a way to estimate this quantity nonparametrically when the quantiles may be functions of time.

3.2.2.1 Technical conditions

The kernel K satisfies the following conditions (see Gasser and Müller 1984):

1. $K \in C^{\nu+1}[-1,1]$;
2. $K(x) \geq 0, K(x) = 0\ (|x| > 1), \int_{-1}^{1} K(x)dx = 1$;
3. $\forall x, y \in [-1,1], |K^{(\nu)}(x) - K^{(\nu)}(y)| \leq L_0|x-y|$ where $L_0 \in \mathbb{R}^+$ is a constant;
4. K is of order $(\nu, k), \nu \leq k-2$, where k is a positive integer, i.e.,

$$\int_{-1}^{1} K^{(\nu)}(x)x^j dx = \begin{cases} (-1)^\nu \nu!, & j = \nu \\ 0, & j = 0, \ldots, \nu-1, \nu+1, \ldots, k-1 \\ \theta, & j = k \end{cases} \tag{3.72}$$

where $\theta \neq 0$ is a constant.

5. Condition on $K^{(j)}$:

$$K^{(j)}(1) = K^{(j)}(-1) = 0, \forall\, j = 0, 1, \ldots, \nu - 1. \tag{3.73}$$

The last two conditions above imply the following (also see Lemma 1 in Gasser and Müller 1984):

$$\int_{-1}^{1} K(x)x^j dx = \begin{cases} 1, & j = 0 \\ 0, & j = 1, \ldots, k - \nu - 1 \\ (-1)^{\nu}\theta\frac{(k-\nu)!}{k!}, & j = k - \nu \end{cases}. \tag{3.74}$$

Below we mention some additional conditions as well as some of the technical conditions discussed earlier. The first condition in (A1) implies a type of local stationarity in the errors u_i that allows for slow and smooth changes in their marginal distribution function. (A2) implies long-memory in Z_i being inherited by the subordinated process $u_i = G(Z_i, t)$. (A5) ensures that all time points are distinct, as well as no extreme clustering of time points. The first condition in (A6) ensures asymptotic unbiasedness of $\widehat{m}^{(\nu)}(t)$, whereas the second and third conditions imply convergence of the variance of the curve estimate to zero. (A7) is used in the derivation of the asymptotic expression of the mean squared error. Condition (A2) is also required in (A8).

- (A1) The Hermite coefficients: $c_k(t) = E[G(Z, t)H_k(Z)]$ are continuously differentiable with respect to $t \in [0, 1]$.
- (A2) $1 - (2q)^{-1} < H < 1$.
- (A3) The trend function m: $m \in C^{\nu+1}[0, 1]$.
- (A4) The time points where data are recorded: $0 \le T_1 \le T_2 \le \cdots \le T_n$, $t_i = T_i/T_n \in [0, 1]$.
- (A5) Conditions on the spacings between two consecutive time points: $\alpha_n^{-1} \le t_j - t_{j-1} \le \beta_n^{-1}$ where $\alpha_n \ge \beta_n > 0$ and $\beta_n \to \infty$. The equidistant case is included here: simply set $\alpha_n = \beta_n = n$.
- (A6) Conditions on the bandwidth: as $n \to \infty$, $b \to 0$, $bT_n \to \infty$, and $b\beta_n \to \infty$.
- (A7) Further conditions on the bandwidth: $\lim_{n\to\infty} (b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2} = 0$.
- (A8) Additional conditions on the kernel: $K \in C^{\nu+1}[0, 1]$ with $0 < c_{\nu+1} = \sup_{u\in[0,1]} |K^{(\nu+1)}(u)| < \infty$.

3.2.3 Asymptotic properties

Under the assumptions mentioned above, the asymptotic expressions for bias, variance, and the mean squared error for the trend derivative estimate can be given as follows. First of all, let the sample size $n \to \infty$. For simplicity, we let the kernel K have compact support. However, this assumption can be relaxed and details can be worked out also in particular when the kernel is sufficiently regular and has non-compact support.

Let $t \in (0, 1)$. Define the quantities I and J as

$$I_q(t) = \frac{c_q^2(t)}{q!} C_Z^q \int_{-1}^{1} \int_{-1}^{1} K^{(\nu)}(u) K^{(\nu)}(v) |u - v|^{(2H-2)q} du\, dv \tag{3.75}$$

and

$$J_{\nu,k}(t) = \frac{g^{(k)}(t)}{k!} \int_{-1}^{1} K^{(\nu)}(u) u^{k-\nu} du. \tag{3.76}$$

Then the expression for the bias later in (3.96) of the trend estimator $\widehat{m}^{(\nu)}(t)$ can be derived as in Chapter 2 on Nonparametric Regression. In particular, one uses a Taylor series expansion of m and the properties of the kernel and the bandwidth.

To derive the variance of the estimator, define

$$V_{i,j} = Cov(y_i, y_j) = \sum_{l=q}^{n} \frac{c_l(t_i) c_l(t_j)}{l!} \gamma_Z^l(T_i - T_j). \tag{3.77}$$

However, $-1 < (2H - 2)q < 0$ and

$$\gamma_Z(T_i - T_j) \sim C_Z |T_i - T_j|^{2H-2}. \tag{3.78}$$

This means

$$Cov(y_i, y_j) \sim \frac{c_q^2(t)}{q!} \gamma_Z^q(T_i - T_j) \tag{3.79}$$

for $i, j \in U_b(t)$ with $U_b = k \in \mathbb{N} : |t - T_k / T_n| \le b$. We then have

$$\begin{aligned} & b^{2\nu} (T_n b)^{(2-2H)q} \mathbb{V}ar(\widehat{m}^{(\nu)}(t)) \\ &= b^{-2} (T_n b)^{(2-2H)q} \sum_{i,j=1}^{n} \left[(t_i - t_{i-1})(t_j - t_{j-1}) K^{(\nu)} \left(\frac{t - t_i}{b} \right) \right. \\ & \left. \times\ K^{(\nu)} \left(\frac{t - t_j}{b} \right) V_{i,j} \right] \end{aligned} \tag{3.80}$$

Now consider the double sum

$$S_n = b^{-2}(T_n b)^{(2-2H)q} \sum_{i \neq j} \left[(t_i - t_{i-1})(t_j - t_{j-1}) K^{(\nu)} \left(\frac{t_i - t}{b} \right) \right.$$
$$\left. \times K^{(\nu)} \left(\frac{t_j - t}{b} \right) |T_i - T_j|^{(2H-2)q} \right]. \tag{3.81}$$

Since $K(u) = 0$ for $|u| > 1$, we have

$$S_n = \sum_{i:|T_i - tT_n| \leq b} K^{(\nu)} \left(\frac{t_i - t}{b} \right) \frac{t_i - t_{i-1}}{b} [S_{i,1} + S_{i,2}] \tag{3.82}$$

where

$$S_{i,1} = \sum_{j \in A_i} K^{(\nu)} \left(\frac{t_j - t}{b} \right) \cdot \left(\frac{t_i - t_j}{b} \right)^{(2H-2)q} \frac{t_j - t_{j-1}}{b}, \tag{3.83}$$

$$S_{i,1} = \sum_{j \in B_i} K^{(\nu)} \left(\frac{t_j - t}{b} \right) \cdot \left(\frac{t_i - t_j}{b} \right)^{(2H-2)q} \frac{t_j - t_{j-1}}{b} \tag{3.84}$$

and

$$A_i = \{ j \in \mathbb{N} : 1 \leq j \leq i-1, |T_i - tT_n| \leq b \}, \tag{3.85}$$

$$B_i = \{ j \in \mathbb{N} : i+1 \leq j \leq n, |T_i - tT_n| \leq b \}. \tag{3.86}$$

Introduce the notation

$$h_n(x) = K^{(\nu)} \left(\frac{t_j - t}{b} \right) \cdot \left(\frac{t_i - t_j}{b} \right)^{(2H-2)q}. \tag{3.87}$$

Then we have

$$S_{i,1} = \int_{t_1/b}^{t_{i-1}/b} h_n(x)dx + \sum_{j \in A_i} h_n'(x_j) \left(\frac{t_j - t_{j-1}}{b} \right)^2 \tag{3.88}$$

$$= \int_{t_1/b}^{t_{i-1}/b} h_n(x)dx + r_{n,i,1} \tag{3.89}$$

and similarly for $S_{i,2}$. Here $t_{j-1}/b \le x_j \le t_j/b$ and

$$h_n'(x) = g_{n,1}(x) + g_{n,2}(x) \tag{3.90}$$

with

$$g_{n,1}(x) = K^{(\nu+1)}\left(\frac{t_j - t}{b}\right) \cdot \left(\frac{t_i - t_j}{b}\right)^{(2H-2)q}, \tag{3.91}$$

$$g_{n,2}(x) = K^{(\nu)}\left(\frac{t_j - t}{b}\right) \cdot \left(\frac{t_i - t_j}{b}\right)^{(2H-2)q-1} (2-2H)q. \tag{3.92}$$

By assumption we have $\alpha_n^{-1} \le \left|t_j - t_{j-1}\right| \le \beta_n^{-1}$, $-1 < (2H-2)q < 0$, and

$$0 \le \sup_{u\in[0,1]} |K^{(\nu+1)}(u)| = c_{\nu+1} < \infty. \tag{3.93}$$

In addition, $b\beta_n \to \infty$ implies $b\alpha_n \to \infty$. This means, denoting $j_1 = [tT_n - b\alpha_n]$ and $j_2 = [tT_n + b\alpha_n]$, an upper bound is

$$\begin{aligned}
&\left|\sum_{j\in A_i} g_{n,1}(x_j)\left(\frac{t_j - t_{j-1}}{b}\right)^2\right| \\
&\le c_{\nu+1} b^{-2}\beta_n^{-2} \sum_{j=j_1}^{j_2} \left(\frac{t_i - t_j}{b}\right)^{(2H-2)q} \\
&\le c_{\nu+1} b^{-2}\beta_n^{-2} \sum_{j=1}^{[2b\alpha_n]} \left(\frac{j}{b\alpha_n}\right)^{(2H-2)q} \\
&= c_{\nu+1} b^{-1}\alpha_n\beta_n^{-2} \sum_{j=1}^{[2b\alpha_n]} \left(\frac{j}{b\alpha_n}\right)^{(2H-2)q} \frac{1}{b\alpha_n} \\
&\le c_{\nu+1} b^{-1}\alpha_n\beta_n^{-2} \int_o^2 x^{(2H-2)q} dx.
\end{aligned} \tag{3.94}$$

This means that, when $(2H-2)q > -1$ and $\lim_{n\to\infty} b^{-1}\alpha_n\beta_n^{-2} = 0$, there is a uniform (in i) upper bound on the remainder term $r_{n,i,1}$. Note that $1 + (2-2H)q > 1$ and $b\alpha_n \to \infty$ so that $\lim_{n\to\infty} b\alpha_n(b\beta_n)^{-2} = 0$ follows from the assumption that

$\lim_{n\to\infty}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2} = 0$. Similarly, for the remainder term $r_{n,i,2}$ in $g_{n,2}$ we have

$$\begin{aligned}
&\left|\sum_{j\in A_i} g_{n,2}(x_j)\left(\frac{t_j - t_{j-1}}{b}\right)^2\right| \\
&\le c_{\nu+1}(b\beta_n)^{-2}\sum_{j=j_1}^{j_2}\left(\frac{t_i - t_j}{b}\right)^{(2H-2)q-1} \\
&\le c_{\nu+1}(b\beta_n)^{-2}\sum_{j=1}^{[2b\alpha_n]}\left(\frac{j}{b\alpha_n}\right)^{(2H-2)q-1} \\
&= c_{\nu+1}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2}\sum_{j=1}^{[2b\alpha_n]} j^{(2H-2)q-1} \\
&\le c_{\nu+1}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2}\sum_{j=1}^{\infty} j^{(2H-2)q-1} \qquad (3.95)
\end{aligned}$$

so that under the assumption that $H < 1$ and $\lim_{n\to\infty}(b\alpha_n)^{1+(2-2H)q}(b\beta_n)^{-2} = 0$ there is a uniform (in i) upper bound on the remainder term $r_{n,i,2}$. Analogous arguments apply to $S_{i,2}$.

This means that the S_n converges to the corresponding double integral and $c_q^2(t)/q!C_Z \cdot S_n$ converges to the asymptotic variance in (3.97) below.

In summary, the asymptotic expressions for bias and variance of the trend derivative estimator are given by:

Bias:

$$\mathbb{E}(\widehat{m}^{(\nu)}(t)) - m^{(\nu)}(t) = b^{k-\nu}J_{\nu,k(t)} + o(b^{k-\nu}), \qquad (3.96)$$

Variance:

$$\mathbb{V}ar(\widehat{m}^{(\nu)}(t)) = b^{-2\nu}(T_n b)^{(2H-2)q}I_q(t) + o(b^{-2\nu}(T_n b)^{(2H-2)q}). \qquad (3.97)$$

The leading term in the mean squared error is:

$$\text{AMSE} = \text{ Sum of the leading terms in Bias and Variance} \qquad (3.98)$$

so that

$$mse = AMSE + r_n, \tag{3.99}$$

where the remainder term is

$$r_n = b^{-2\nu}(T_n b)^{(2H-2)q} I_q(t) + o(\max(b^{2(k-\nu)}, b^{-2\nu}(T_n b)^{(2H-2)q})). \tag{3.100}$$

A formula for an asymptotically optimal bandwidth follows immediately by minimizing $AMSE$ with respect to b. This is given as:

Local optimal bandwidth:

$$b_{opt} = \left[\frac{2\nu + (2-2H)q}{2(k-\nu)} \frac{I_q}{J_{\nu,k}^2}\right]^{1/(2k+(2-2H)q)} T_n^{(2H-2)q/(2k+(2-2H)q)}. \tag{3.101}$$

A typical case is $\nu = 0$, $k = 2$, and $q = 1$, as for instance when trend estimation is of interest, with Gaussian errors and using a Gaussian kernel. In this case, the rate of decay of the optimum bandwidth to zero is $T_N^{(2H-2)/(6-2H)}$. When the observations are evenly spaced, $T_n = n$. If also $H = 1/2$, i.e., the data have short-memory, then the rate becomes $n^{-1/5}$, the same as for iid observations.

3.2.3.1 Central limit theorem

Menéndez et al. (2013) state the central limit theorem when t is fixed. When $t_1, \ldots, t_k$ are distinct points, see Menéndez et al. (2010) for the multivariate central limit theorem. These results concern the asymptotic distribution of the centered curve estimator. Since G is monotonically increasing, the Hermite rank of G is $q = 1$ so that the first Hermite coefficient is $c_1(t_i) \neq 0$. In this case, the asymptotic distribution of the centered curve estimator is normal. In addition, the estimates at the different but fixed values $t_1, \ldots, t_k$ are asymptotically independent. A similar limit theorem can be derived for $q \geq 2$, but with a non-normal limiting distribution corresponding to the marginal distribution of a Hermite process of order q.

Define the quantities $I_{\nu,q}^*(t)$ and $J_{\nu,k}^*$ as follows:

$$I_{\nu,q}^*(t) = C_{\nu,q} c_q^2(t) C_Z^q \tag{3.102}$$

where q is the Hermite rank of G, c_q is the qth Hermite coefficient, and C_Z appears in the error covariance $Cov(u_i, u_j)$ whereas $C_{\nu,q}$ is

$$C_{\nu,q} = \frac{1}{q!}\int_{-1}^{1}\int_{-1}^{1} K^{(\nu)}(u)K^{(\nu)}(v)|u-v|^{q(2H-2)}du\,dv. \quad (3.103)$$

J^* is defined in terms of J (see 3.76):

$$J^*_{\nu,k} = J_{\nu,k}\frac{k!}{(k-\nu)!}. \quad (3.104)$$

CLT fixed t:

Let the Hermite rank q be equal to 1. Then for every fixed $t \in (0,1)$, as $n \to \infty$,

$$\xi_n(t) = a_n^{(1)}(t)\left[\widehat{m}^{(\nu)}(t) - m^{(\nu)}(t) - b^{k-\nu}m^{(k)}(t)J^*_{\nu,k}\right] \quad (3.105)$$

converges to a standard normal variable where

$$a_n^{(q)}(t) = T_n^{1-H}b^{1-H+\nu}I^*_{\nu,q}(t)^{-\frac{1}{2}} \quad (3.106)$$

and $J^*_{\nu,k}$ and $I^*_{\nu,q}(t)$ are defined in (3.104) and (3.102) respectively.

To see why this is the case, note that we can write

$$\widehat{m}^{(\nu)}(t) = m^{(\nu)}(t) + r_n + \zeta_{n,q} \quad (3.107)$$

where

$$r_n = (b^{k-\nu}/(k-\nu)!)m^{(k)}(t)\int_{-1}^{1}K(u)u^{k-\nu}du + o(b^{k-\nu}) + O(1/(b^{1+\nu}\beta_n)) \quad (3.108)$$

and

$$\zeta_{n,q} = (1/b^{\nu+1})\sum_{i=1}^{n}(t_i - t_{i-1})K^{(\nu)}((t_i-t)/b) \times \sum_{l=q}^{\infty}(c_l(t)/l!)H_l(Z_i). \quad (3.109)$$

We also know that, as $n \to \infty$,

$$\mathbb{V}ar[\zeta_{n,q+1}] = O((T_n b)^{(2H-2)(q+1)}/b^{2\nu}) \quad (3.110)$$

and

$$\mathbb{V}ar[a_n^{(q)}\zeta_{n,q+1}] = O\big((T_n b)^{(2H-2)}\big) \to 0 \tag{3.111}$$

where $a_n^{(q)}$ is defined in (3.106). Also,

$$\mathbb{E}(\zeta_{n,q+1}) = 0. \tag{3.112}$$

Thus, by Chebychev's inequality, $a_n^{(q)}\zeta_{n,q+1}$ converges to zero in probability as $n \to \infty$. When $q = 1$, define

$$X_n = a_n^{(1)}[\zeta_{n,1} - \zeta_{n,2}]. \tag{3.113}$$

Then

$$X_n = \{a_n^{(1)}/b^{(\nu+1)}\} \sum_{i=1}^{n} (t_i - t_{i-1}) K^{(\nu)}((t_i - t)/b) c_1(t_i) Z_i. \tag{3.114}$$

Since $\mathbb{E}(X_n) = 0$ and $\mathbb{V}ar(X_n) \sim 1$, being a linear combination of jointly normal variables, X_n is asymptotically standard normal. The result follows by noting that

$$a_n^{(1)}\zeta_{n,1} = X_n + a_n^{(1)}\zeta_{n,2} \tag{3.115}$$

where the second term converges to zero in probability.

3.2.3.2 Covariance between trend derivatives

Let $\eta \geq \nu \geq 0$, $x = h - b \geq 0$, and $x/h = O(1)$, where h and b are bandwidths, satisfying similar conditions as b, and $h \geq b > 0$. Specifically, we let h be the bandwidth for estimating $\widehat{m}^{(\eta)}(t)$ and let b be the bandwidth for estimating $\widehat{m}^{(\nu)}(t)$. Then under as $n \to \infty$,

$$\begin{aligned}
\mathbb{C}ov(\widehat{m}^{(\nu)}(t), \widehat{m}^{(\eta)}(t)) &\sim g(\nu, \eta, b, h, t) \\
&= (-1)^{\nu+\eta} \frac{(T_n h)^{(2H-2)q}}{b^{\nu} h^{\eta}} \times L_{\nu,\eta,q}(t)
\end{aligned}$$

where $$L_{\nu,\eta,q}(t) = D_{\nu,\eta,q} c_q^2(t) C_Z^q$$

and $$D_{\nu,\eta,q} = \frac{1}{q!} \int_{-1}^{1} \int_{-1}^{1} \Big[K^{(\nu)}(u) K^{(\eta)}(v) \times |(u - v) - u\frac{x}{h}|^{(2H-2)q} \Big] du\, dv$$

To see this, note that we can write

$$T_i - T_j = T_n h[(1 - x/h)(t_i - t)/b - (t_j - t)/h] \tag{3.116}$$

where $x = h - b \geq 0$, x/h is asymptotically bounded, and b and h are bandwidths following conditions of the theorem. Then using the argument as before, as $n \to \infty$,

$$\begin{aligned} &\mathbb{C}ov(\widehat{m}^{(\nu)}(t), \widehat{m}^{(\eta)}(t)) \\ &\sim \left\{ \frac{(-1)^{\nu+\eta}}{b^{\nu+1}h^{\eta+1}} \right\} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(t_i - t_{i-1})(t_j - t_{j-1}) K_i^{(\nu)} K_j^{(\eta)} \frac{c_q^2(t)}{q!} \right. \\ &\left. \times C_Z^q |T_i - T_j|^{(2H-2)q} \right] \end{aligned} \tag{3.117}$$

where

$$K_i^{(\nu)} = K^{(\nu)} \left(\frac{t_i - t}{b} \right) \tag{3.118}$$

and

$$K_j^{(\eta)} = K^{(\eta)} \left(\frac{t_j - t}{h} \right). \tag{3.119}$$

Equation (3.116) now follows by approximating the above sum in (3.117) by a double-integral.

3.2.3.3 Confidence interval for the trend

If the Hermite rank q equals 1, a pointwise asymptotic $100(1 - \alpha)\%$ confidence band for the trend $m(t)$ ignoring bias is

$$\widehat{m}(t) \pm z_{\alpha/2} / \widehat{a}_n^{(1)}(t). \tag{3.120}$$

Here $\widehat{a}_n^{(1)}(t)$ is a consistent estimate of $a_n^{(1)}(t)$ as defined in (3.106), $z_{\alpha/2}$ being the upper $\alpha/2$-point of the $N(0, 1)$ distribution.

Substituting $\nu = 0$ and $q = 1$,

$$a_n^{(1)}(t) = T_n^{1-H} b^{1-H} I_{0,1}(t)^{-1/2} \tag{3.121}$$

where $I_{0,1}(t)$ is defined in (3.102) and this quantity needs to be estimated, involving estimation of H, $c_1(t)$ and C_Z.

When $q = 1$ and if $\widehat{m}$ is a consistent estimate of m, a bias corrected confidence interval for the trend function $m(t)$ may be given by

$$\left\{\widehat{m}(t) - \frac{b^2}{2}\widehat{m}^{(2)}(t)\mu_2\right\} \pm \frac{1}{d_n(t)} z_{\alpha/2}, \tag{3.122}$$

where $$\mu_2 = \int_{-1}^{1} u^2 K(u) du,$$

and $$d_n(t) = \left\{\widehat{g}(0,0,b,b,t) + \frac{b^4}{4}\mu_2^2 \widehat{g}(2,2,h,h,t) - b^2\mu_2 \widehat{g}(0,2,b,h,t)\right\}^{1/2}$$

and g is defined in (3.116). To obtain a confidence interval ignoring bias, the trend may be estimated using a sub-optimal bandwidth (e.g., Hall 1992, p. 207).

Note that when $\nu = \eta = 0$ and $q = 1$, $D_{0,0,1} = C_{0,1}$. This is later used for computing the confidence interval for the trend $m(t)$ and also for bandwidth selection.

3.2.4 Bandwidth selection

3.2.4.1 Global optimal bandwidth

Consider the leading term of the mean squared error $mse(\widehat{m}^{(\nu)}(t))$ given in (3.98). To obtain the formula for the global optimal bandwidth in (3.125), the mean integrated squared error is minimized. Specifically, the asymptotic mean integrated squared error (AMISE) is

$$\text{AMISE}(\widehat{m}^{(\nu)}) = \int_0^1 \text{AMSE}(\widehat{m}^{(\nu)}(t)) dt. \tag{3.123}$$

The global optimal bandwidth is then obtained as

$$b_{opt} = \underset{b}{argmin}\, \text{AMISE}(\widehat{m}^{(\nu)}), \tag{3.124}$$

the formula for which is derived by differentiating $\text{AMISE}(\widehat{m}^{(\nu)})$ and by equating the resulting expression to zero. In a similar manner, the expression for the local optimal bandwidth can be obtained as a function of time t, where one would minimize the leading term of $\text{AMSE}(\widehat{m}^{(\nu)}(t))$.

The formula for the global optimal bandwidth minimizing the mean integrated squared error is thus given by

$$b_{opt} = \left[\frac{R(c_q)}{R(m^{(k)})}\right]^{1/\delta} \left[\frac{2\nu - (2H-2)q}{2(k-\nu)} \frac{C_{\nu,q} C_Z^q}{J_{\nu,k}^2}\right]^{1/\delta} [T_n]^{(2H-2)q/\delta} \tag{3.125}$$

where $\delta = 2k - (2H-2)q$ and for a square integrable function f, $R(f) = \int_0^1 f^2(u)du$.

3.2.4.2 Data-driven approach

As mentioned earlier, we assume $G(\cdot, t)$ to be monotone increasing so that $q = 1$. One can then use an iterative plug-in approach (see, for example, Sheather and Jones 1991) for an optimal bandwidth estimation.

Step 1: Let b_{init} be the initial bandwidth. Compute $\widehat{m}(t)$ using $b = b_{init}$ followed by computing the residuals $\hat{u}_i = y_i - \widehat{m}(t_i)$.

Step 2: Compute estimates $\widehat{H}$, $\widehat{C}_Z$, and $\widehat{c}_1(t)$ from the residuals $\hat{u}_i = y_i - \widehat{m}(t_i)$.

Step 3: Compute an updated bandwidth $b_{updated}$ by plugging in the estimates from **Step 2** in (3.125).

Step 4: Repeat steps 1 to 3 until convergence.

3.2.4.3 Further on estimation

To estimate H and C_Z in step 2, one may use the periodogram of the scaled residuals $\widehat{u}_{scaled,i} = \widehat{u}_i / \widehat{\sigma}(t_i)$ where, for a given bandwidth b_s and a kernel K_s,

$$\widehat{\sigma^2}(t) = \frac{1}{b_s} \sum_{i=1}^{n} (t_i - t_{i-1}) K_s\left(\frac{t_i - t}{b_s}\right) \widehat{u}_i^2. \tag{3.126}$$

Additional details on computations can be found in Menéndez (2013). For instance, one may use a periodogram that is defined for irregularly spaced data (see Masry 1978) or the Lomb–Scargle periodogram (see Press et al. 1992, pp. 575–584), popular in bioinformatics and geology. For extensions of this method, see, for instance, Lévy-Leduc et al. (2008). Estimation of H and C_Z can then be done by, for instance, fitting a straight line

in log-log coordinates to periodogram ordinates close to zero frequency or by the approximate maximum likelihood method (Haslett and Raftery 1989). These are implemented in the *R*-package as *fracdiff*. For additional background information, see Robinson (1995) and Geweke and Porter-Hudak (1983).

3.2.4.4 Estimation of the first Hermite coefficient

In order to estimate $R(c_1) = \int c_1^2(t)dt$ in step 2, where c_1 is the first Hermite coefficient in G, one may consider exploiting the monotonicity property of G, leading to an estimate of Z.

This is done as follows. Being monotonically increasing,

$$Z_i = \Phi^{-1}(F(t_i, u_{scaled,i})). \tag{3.127}$$

Here F denotes the marginal distribution function of the standardized errors and

$$u_{scaled}(tT_n) = u(tT_n)/\sigma(t) \tag{3.128}$$

(at least if the latter is continuous and strictly increasing) at t. Substitution of $\widehat{F}$ leads to a nonparametric "estimate" or a "proxy" for the latent series Z_i. Finally, since the first Hermite coefficient $c_l(t_j)$ is the expected value of $u_j H_l(Z_j)$, $c_1(t)$ may be estimated by smoothing as follows:

$$\widehat{c}_1(t) = \frac{1}{b_c}\sum_{j=1}^{n}(t_j - t_{j-1})K_c\left(\frac{t_j - t}{b_c}\right)\widehat{u}_j\widehat{Z}_j, t \in (0,1), \tag{3.129}$$

where b_c denotes a bandwidth and K_c is a kernel.

On the other hand, since G is monotone, H can be estimated either from $\hat{u}_i$'s or from the estimated latent Gaussian series.

Note that by extending Dehling and Taqqu (1989) to include time, the consistency of $\widehat{F}(t, y)$ holds uniformly in t and y. Another approach would be to consider a kernel that has an absolutely integrable characteristic function. Since $\widehat{m}(t)$ is consistent,

$$\widehat{u}_j = u_j + u_{1n,j} \tag{3.130}$$

and

$$\widehat{u}_{scaled}(T_j) = u_{scaled}(T_j) + u_{2n,j} \tag{3.131}$$

where $u_{1n,j}$ and $u_{2n,j}$ converge to zero in probability as $n \to \infty$. Moreover, as mentioned above under suitable conditions, uniform convergence may be attained (see Hall and Hart 1990 (Theorem 2.1), Parzen 1962, Ghosh 2014, and others). Similarly, since

$$\widehat{F}(t, e + \nu_{1n}) = \frac{1}{b} \sum_{i=1}^{n} (t_i - t_{i-1}) K\left(\frac{t_i - t}{b}\right) I_i(e + \nu_{1n}),$$

and the Hermite polynomial expansion of the centered process I_i holds, due to the regularity conditions on the Hermite coefficients and on F,

$$\widehat{F}(t, e + \nu_{1n}) = F(t, e) + \nu_{2n} \tag{3.132}$$

for every $e \in \mathbb{R}$, so that

$$\widehat{Z}_j = Z_j + w_n \tag{3.133}$$

where ν_{1n}, ν_{2n}, and w_n converge to zero uniformly in probability as $n \to \infty$. Moreover, note that the leading part of the mean integrated squared error of the estimated trend function $MISE(\widehat{m})$ can be estimated consistently by plugging in the estimates of the unknown quantities. Finally, the continuous mapping theorem can be applied to arrive at the following assertion:

Suppose that $\widehat{H}$ and $\widehat{C}_Z$ are consistent estimates of H and C_Z respectively, the technical conditions mentioned earlier hold, and the Hermite coefficients $\tilde{c}_l(t, y)$ as well as $F(t, y)$ are continuously differentiable functions of t and y. Moreover, let $\partial/\partial y F(t, y) = f(t, y)$ and $\partial^2/\partial t^2 F(t, y)$ exist.

Then as $n \to \infty$, $\widehat{c}_1(t)$ is a weakly consistent estimator of $c_1(t)$ and the ratio of the estimated b_{opt} and the true b_{opt} converges to one in probability.

The bandwidth selection method discussed here is based on the assumption that G is monotone so that the Hermite rank q is equal to one. Also, the estimated $Z(T)$ cannot be used for estimating the Hermite rank q or for carrying out a Taylor series-of-fit test for unit Hermite rank. However, the Hermite coefficients can be estimated consistently.

For discrete time processes with stationary marginal distributions, Ray and Tsay (1997) and Beran and Feng (2002) propose

bandwidth selection methods; also see references therein. When the regression errors are iid, or when standardized have stationary marginal distribution, see Wand and Jones (1995) for bandwidth selection procedures. Some of these ideas were also presented earlier in Chapter 2 on Nonparametric Regression. The formula for the optimal bandwidth for the iid case can be seen as a special case (substitute $H = 1/2$, $T_n = n$, and $\nu = 0$ to address trend estimation using a kernel of order 2, i.e., $k = 2$) so that the rate of the optimal bandwidth becomes $n^{-1/5}$. In the case of long-memory ($H > 1/2$), on the other hand, everything else remaining the same, the rate of decay of the optimal bandwidth to zero is $n^{(H-1)/(3-H)}$. This rate is slower under long-memory than under independence or zero correlations (see Beran 1994, Wand and Jones 1995, and Ghosh 2001). For consistency properties of an approximate maximum likelihood estimate of H for discrete time processes under non-Gaussianity, see Giraitis and Taqqu (1999).

3.3 Rapid change points

Estimation of points of rapid change in the mean function $m(t)$ in a nonparametric regression model with time series observation are of interest, for instance, in palaeo climate research where changes in the past environment are of interest. Here we address estimation of the time points where such changes take place and present some results when there are long memory correlations in the regression errors. Long-memory correlations in geophysical records have been cited several times in the literature and Beran (1994) and Beran et al. (2013) provide background information; also see references therein.

Because of the typical application areas, as for instance in the palaeo sciences, we focus on time series data that are irregularly spaced in time. This was also the topic of interest in the previous section. In the discussion here, therefore, we will be referring to some of the consistency results presented in that section. Also as in the previous section, an additional issue is that of (locally stationary) Gaussian subordinated regression errors, which provides a flexible framework for handling time-dependent marginal distribution function of the errors.

The approach presented here relies on derivatives of the trend function so that their nonparametric estimation is of interest, which is then followed by estimation of the change points. The definition is based on the first three derivatives of the regression function. Kernel smoothing applied to time series observations in a nonparametric regression model with irregularly spaced time points yields estimates of these derivatives. The regression residuals are assumed to be obtained by time-varying Gaussian subordination with long-range dependence.

In the palaeo sciences, proxy data are used to understand past environment conditions. For instance, measurements of oxygen isotopes present in the Greenland ice sheets (see Greenland Ice Core Project in Johnsen et al. 1997) are used to estimate past temperatures. Such measurements are thus called temperature proxies. Time series such as these may often show fast changes. In Figure 3.4, a fast change between the Holocene

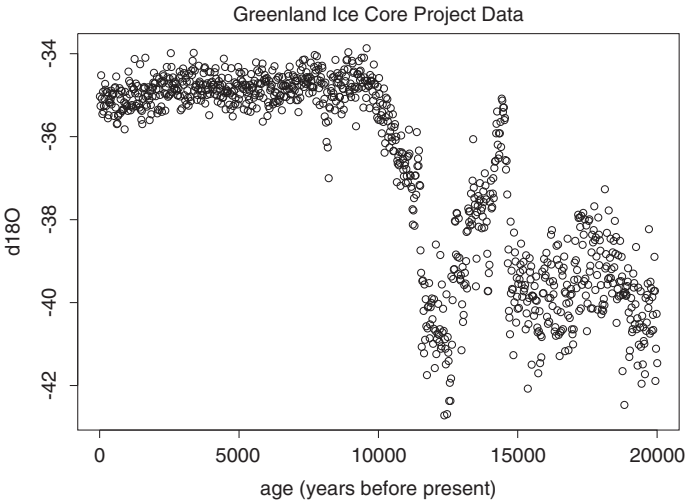

Figure 3.4 Measurements of oxygen isotopes in Greenland ice sheets in the last 20 000 years. The data show temperature shifts and occasional rapid changes. The Holocene (ca. last 11 500 years) was warmer than the Younger Dryas (ca. 11 500-12 700 years before the present). The Younger Dryas experienced severely cold temperatures with extremely abrupt changes marking its boundaries. Quantitative estimates of points of rapid change are of particular interest when identifying periods where abrupt climate changes took place and high environmental variability occurred. *Source:* NASA.

(approximately the last 11 500 years) and the Younger Dryas (about 11 500–12 700 years before present) can be seen. For several reasons, it is important to accurately estimate when such changes took place. From a statistical point of view this falls within the topic of change point estimation.

In addressing rapid changes in the trend function in time series data of the type mentioned above, we make a note of the following: (i) the change is fast but smooth, i.e., not discontinuous changes; (ii) the errors (detrended series) exhibit long-range dependence; (iii) the marginal distributions may change smoothly in the course of time; and (iv) the time points where time series observations are available need not be equidistant.

The change point literature is vast. Some references are: Hinkley (1970), who considers parametric change point estimation for independent random variables with density function $f(x;\theta)$ and a change in θ; Picard (1985) considers tests for time series with changes in the spectrum; an interesting article is due to Müller (1992), who considers estimating the location and size of discontinuities in derivatives of the trend function in nonparametric regression (also see Loader 1996, Gijbels et al. 1999, and Gijbels and Goderniaux 2004a,2004b); Horvath and Kokoszka (2002) test the null hypothesis that the pth derivative of the trend function is continuous against the presence of a discontinuity; Ghosh (2006) considers change point estimation in the context of synchronizing two isotope series; Inclan and Tiao (1994) study the problem of change points in the variance. Chopin (2006) considers Bayesian filtering and smoothing and proposes a state-space representation of change point models. For more reference to change point estimation see, for example, Pons (2003), Koosorok and Song (2007), and Lan et al (2009). In the context of long-memory processes, Giraitis and Leipus (1992) study detection of changes in the spectral distribution, and Beran and Terrin (1994, 1996) study the problem of finding a change in the long-memory parameter (also see Horvath and Shao 1999). Further references are, for instance, Csörgő and Horvath (1997), Horvath and Kokoszka (1997), Kuan and Hsu (1998), and Kokoszka and Leipus (2003). Locally stationary long-memory processes are considered in Jensen and Whitcher (2000) and Beran (2009). Also see Ghosh (2014), Menéndez (2010, 2013), and Ghosh and Draghicescu (2002a,2002b). General references to long-memory processes are, for instance, Granger and Joyeux

(1980), Hosking (1981), Beran (1994), and Beran et al. (2013). For kernel smoothing of time series with long-memory see Hall and Hart (1990), Csörgő and Mielniczuk (1995), Ray and Tsay (1997), Beran and Feng (2002), and Beran et al. (2002). Basic information on bandwidth selection for kernel estimates in the iid context can be found in Gasser and Müller (1984).

3.3.1 Model and definition of rapid change

We address an estimation of rapid change points in the context of a nonparametric regression problem with Gaussian subordinated errors. Consider the latent process $Z(u), u \in \mathbb{R}$, that is a continuous-time stationary Gaussian process with

$$\mathbb{E}(Z(u)) = 0, \mathbb{V}ar(Z) = 1 \tag{3.134}$$

and

$$\gamma_Z(v) = \mathbb{C}ov(Z(u), Z(u+v)) \sim C_Z v^{2H-2} \tag{3.135}$$

as $v \to \infty$ where $H \in (0,1)$ and for two functions a and b, $a(v) \sim b(v)$ implies that $a(v)/b(v)$ tends to one.

Now consider the nonparametric regression model

$$y_i = y(T_i) = m(t_i) + u_i \tag{3.136}$$

where $u_i = G(Z(T_i), t_i)$, $T_i \in \mathbb{R}_+$, $T_1 \le T_2 \le \cdots \le T_n$, $t_i = T_i/T_n \in [0,1]$ and m is a smooth function. For fixed $t \in [0,1]$, $G(\cdot, t)$ is an L^2 function on $\mathbb{R}$ with

$$\mathbb{E}(G(Z)) = (2\pi)^{-1/2} \int G(z) \exp(-z^2/2) dz = 0, \tag{3.137}$$

$$\mathbb{E}(G^2(Z)) < \infty. \tag{3.138}$$

We may then expand G using the Hermite polynomial expansion

$$G(Z_i, t_i) = \sum_{k=q}^{\infty} \frac{c_k(t_i)}{k!} H_k(Z_i). \tag{3.139}$$

Here H_k are Hermite polynomials and $q \ge 1$ is the Hermite rank, the smallest positive integer for which the Hermite coefficient c_q is not equal to zero. This formulation of the regression error is convenient for analyzing time series, in particular arising in geophysics and elsewhere where time-dependent changes

in the marginal distribution function can be envisaged. In particular, the marginal distribuion of G may be non-Gaussian; see Ghosh et al. 1997 and Ghosh and Draghicescu 2002a,2002b) for further discussions and examples. In addition, we let the time points $T_1, \dots, T_n$, where the observations on the response variable, i.e., $y_1, \dots, y_n$, are available, be arbitrary, i.e. they may be non-equidistant. When these points are equidistant, we have $T_i = i, i = 1, 2, \dots, n$. In such a case, the rescaled times are $t_i = i/n$.

Let $m^{(\nu)}(t)$ be the νth derivative of m with respect to t. We define change points in terms of an exceedance threshold applied to the first derivative of the trend function. In other words, rapid change occurs if $|m^{(1)}(t)| > \eta$ where $\eta > 0$ is a given threshold. This was also considered in Müller and Wang (1990) in the context of hazard function estimation.

Definition: Let $\eta > 0$ be given. Then $\{\tau_1, \tau_2, \dots, \tau_p\}, (\tau_i \in [0, 1])$, are points of rapid change of $m(\cdot)$ if

$$m^{(2)}(\tau_i) = \frac{\partial^2 m(\tau_i)}{\partial \tau_i^2} = 0, i = 1, \dots, p, \tag{3.140}$$

$$0 < |m^{(3)}(\tau_i)| < \infty, \tag{3.141}$$

and

$$|m^{(1)}(\tau_1)| \geq |m^{(1)}(\tau_2)| \geq \cdots \geq |m^{(1)}(\tau_p)| \geq \eta. \tag{3.142}$$

3.3.2 Estimation and asymptotics

To estimate the rapid change points $\tau_1, \dots, \tau_p$, we simply follow the definition. In other words, find $\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_p \in [0, 1]$ such that $\widehat{m}^{(2)}(\tau_i) = 0$ and $|\widehat{m}^{(1)}(\widehat{\tau}_1)| \geq |\widehat{m}^{(1)}(\widehat{\tau}_2) \geq \cdots \geq |\widehat{m}^{(1)}(\widehat{\tau}_p)| \geq \eta$, where $\eta > 0$ is a given threshold.

The first step is thus to define an estimator for m and its derivatives. In the previous section, such an estimator $\widehat{m}^{(\nu)}(t)$ was defined. Moreover, for a finite sample size n and a given threshold η, the number of points where $\widehat{m}^{(2)}(t)$ equals zero is random. However, if m is at least twice continuously differentiable, then consistency of $\widehat{m}^{(2)}$ implies that p is also estimated consistently. Consistency of $\widehat{m}^{(\nu)}(t)$ was also addressed in the

previous section in the context of trend estimation, where we derive the rates for the asymptotic bias and variance and show that the mean squared error converges to zero.

In addition to consistency, for developing confidence intervals for the rapid change points, their asymptotic distribution is needed. For this, a multivariate central limit theorem is given below. Note that, since the rapid change points are defined in terms of the second derivative of m, the kernel estimator of this trend derivative becomes relevant.

Note that asymptotically the distribution of

$$\Delta_n = (T_n b)^{(1-H)q}\{\widehat{m}^{(2)}(\tau_i) - E[\widehat{m}^{(2)}(\tau_i)]\} \tag{3.143}$$

(for $q = 1$) is equivalent to the asymptotic distribution of

$$\tilde{\Delta}_n = (T_n b)^{(1-H)q}\frac{(-1)^\nu}{nb^{\nu+1}}\sum_{j=1}^{n} K^{(\nu)}\left(\frac{t_j - \tau_i}{b}\right)\frac{c_q(\tau_i)}{q!}H_q(Z_j) \tag{3.144}$$

$$= (T_n b)^{1-H}\frac{(-1)^\nu}{nb^{\nu+1}}\sum_{j=1}^{n} K^{(\nu)}\left(\frac{t_j - \tau_i}{b}\right)c_1(\tau_1)Z_j \tag{3.145}$$

which is a sequence of normal variables. Asymptotic independence of $\widehat{m}^{(\nu)}(t)$ and $\widehat{m}^{(\nu)}(s)$ for $t \neq s$ follows by analogous arguments as in the previous section; also see Csörgő and Mielniczuk (1995). We may thus summarize as follows:

CLT: $t_1, t_2, \ldots, t_k$ **fixed**
Suppose that the Hermite rank q of G is one. Let

$$\mathbf{t} = (t_1, \ldots t_k)', \tag{3.146}$$

$$\widehat{\mathbf{m}}^{(2)}(\mathbf{t}) = \left(\widehat{m}^{(2)}(t_1), \ldots, \widehat{m}^{(2)}(t_k)\right)', \tag{3.147}$$

and define the $k \times k$ diagonal matrix

$$\mathbf{D} = diag\left\{\sqrt{I_1(t_1)}, \ldots, \sqrt{I_1(t_k)}\right\}. \tag{3.148}$$

where $I_q, q = 1, 2, \ldots$, is defined in (3.75). Then as $n \to \infty$,

$$b^\nu (T_n b)^{1-H} D^{-1}\{\widehat{\mathbf{m}}^{(2)}(\mathbf{t}) - E[\widehat{\mathbf{m}}^{(2)}(\mathbf{t})]\} \underset{d}{\to} (\zeta_1, \ldots, \zeta_k)', \tag{3.149}$$

where $\zeta_i \sim iidN(0, 1)$ variables.

To construct confidence intervals, one is concerned with both bias and the variance of the estimator. If bias dominates, then no reasonable confidence interval can be given. Therefore, either one ensures that the bias is more negligible than the variance or that both have similar orders of contribution to the mean squared error; i.e., the rate of convergence to zero of the squared bias equals that for the variance, provided that all else remain fixed. These two cases are therefore

$$b^{2k} \sim C \cdot (T_n b)^{(2H-2)q}, \tag{3.150}$$

where the contribution of the squared bias and of the variance to the mse is of the same order, and

$$b^{2k} = o\big((T_n b)^{(2H-2)q}\big), \tag{3.151}$$

where the contribution of the bias is asymptotically negligible. In these two cases, if the Hermite rank of G equals unity, then $\widehat{\tau}_n$, properly centered and scaled, is asymptotic normal. To see this, one uses a Taylor series expansion,

$$\widehat{\tau}_{n:i} - E(\widehat{\tau}_{n:i}) = -\widehat{m}^{(2)}(\tau_i)[m^{(3)}(\tau_i)]^{-1} + o_p((T_n b)^{H-1}), \tag{3.152}$$

and makes uses of consistency of the denominator, along with CLT for the derivative estimator in the numerator and Slutsky's lemma. We can summarize these ideas as follows.

Define the vector

$$\tau = (\tau_1, \tau_2, \ldots, \tau_p)'. \tag{3.153}$$

Thus the elements of τ are the points of rapid change of m. Now consider the estimates

$$\widehat{\tau}_n = \big(\widehat{\tau}_{n;1,}\widehat{\tau}_{n;2}, \ldots, \widehat{\tau}_{n;p}\big)'. \tag{3.154}$$

Moreover, define the $p \times p$ diagonal matrix

$$\tilde{\mathbf{D}} = diag\left(\sqrt{I_1(\tau_1)}/|m^{(3)}(\tau_1)|, \ldots, \sqrt{I_1(\tau_p)}/|m^{(3)}(\tau_p)|\right), \tag{3.155}$$

where $I_q, q = 1, 2, \ldots$, is defined in (3.75).

Then under the conditions stated above, the asymptotic distribution of $\widehat{\tau}_n$ is

$$\text{(i) } b^{2k} = o\left((T_n b)^{2H-2}\right) \Longrightarrow (T_n b)^{1-H}\tilde{\mathbf{D}}^{-1}(\widehat{\tau}_n - \tau) \underset{d}{\to} (\zeta_1, \ldots, \zeta_p)' \tag{3.156}$$

where ζ_i are iid $N(0,1)$ variables. Moreover,

$$\text{(ii) } b^{2k} \sim C \cdot (T_n b)^{2H-2} \Longrightarrow (T_n b)^{1-H}\tilde{\mathbf{D}}^{-1}\left(\widehat{\tau}_n - \tau\right) \underset{d}{\to} (\mu_1 + \zeta_1, \ldots, \mu_p + \zeta_p)', \tag{3.157}$$

where ζ_i are as in (i) and

$$\mu_i = \left[\frac{m^{(k)}(\tau_i)}{k!}\int_{-1}^{1} K^{(\nu)}(u)u^{k-\nu}du\right] / m^{(3)}(\tau_i). \tag{3.158}$$

Note that when $q > 1$, non-Gaussian limit theorems can be derived. Also, it can be shown that the number of zeroes of $\widehat{m}^{(2)}$ converges to p in probability. Therefore, if n is large enough, one may assume that p is estimable with arbitrary precision so that these asymptotic distributions also hold when p is unknown and estimated.

As an example, consider a data set from the Greenland Ice Core Project (Johnsen et al. 1997). These data from natural archives reveal major fluctuations in the past temperature. Oxygen isotope measurements are used as a temperature proxy to reconstruct the temperatures in the ancient history of the earth. The Holocene (ca. last 11 500 years) is warmer than the Younger Dryas (ca. 11 500–12 700 years before present), which was remarkably cooler. Moreover, the transition from the Younger Dryas to the Holocene was rather abrupt. Younger Dryas is named after the Alpine wildflower *Dryas octopetala*, which left its mark on a number of fossil records, such as palaeo pollen samples or in ice cores. The Younger Dryas finally ended some ten to twelve thousand years ago, and the milder and relatively stable Holocene epoch started with a rapid increase in the temperature.

Menéndez, Ghosh, and Beran (2010) estimate rapid change points using the data from Figure 3.4 and a threshold $\eta = 100$. They note major rapid change points around the Younger Dryas. Specifically, the rapid change is estimated to have occurred around 11 560 and 14 658 years before present (1997). Taking

a bandwidth that is smaller than the optimal bandwidth, simple approximate 95% confidence intervals ignoring the bias are computed and these are (in years before present) [11 554, 11 566] and [14 646, 14 670] respectively.

For nonparametric curve estimation, bandwidth selection is an important issue. In case of long-memory correlations in the errors, bandwidth selection under monotone Gaussian subordination, irregularly spaced data, and long-memory are given in Menéndez et al. (2013). Due to monotonicity, estimation of the latent Gaussian process can be facilitated by using the regression residuals to estimate the underlying marginal distribution of the errors. For discrete time processes and with evenly spaced time series observations, Ghosh and Draghicescu (2002a) propose to directly estimate the variance of the Priestley–Chao estimator. When the marginal distributions are stationary, Ray and Tsay (1997) and Beran and Feng (2002) propose bandwidth selection methods; also see references therein. For additional information of bandwidth selection see Herrmann et al. (1992).

3.4 Nonparametric *M*-estimation of a trend function

3.4.1 Kernel-based *M*-estimation

We consider the regression model

$$y_i = m(t_i) + u_i, \tag{3.159}$$

where $t_i = i/n$, $m \in C^3[0,1]$, and the regression errors are stationary; however, they are Gaussian subordinated, i.e.,

$$u_i = G(Z_i). \tag{3.160}$$

Here Z_i is a zero mean stationary latent Gaussian process with $var(Z_i) = 1$. The transformation G is such that $\mathbb{E}[G(Z)] = 0$ and $\mathbb{E}[G^2(Z)] < \infty$. The autocovariance function and spectral density of Z_i will be denoted respectively by $\gamma_Z(k)$ and

$$f_Z(\lambda) = \frac{1}{2\pi} \sum \gamma_Z(k) \exp(\imath k \lambda), \tag{3.161}$$

where $\imath = \sqrt{-1}$.

We are interested in a nonparametric estimation of $m(t)$ where $t \in (0,1)$ is rescaled time. Let K be a non-negative symmetric kernel with support $[-1,1]$ with $\int K(x)dx = 1$. Given a bandwidth $b > 0$, the Nadaraya–Watson estimator of the trend function m is obtained by minimizing the weighted quadratic form

$$Q(\theta) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right)(y_i - \theta)^2 \tag{3.162}$$

with respect to θ, and setting

$$\widehat{m}_{NW}(t) = \hat{\theta} = \operatorname{argmin} Q(\theta) \tag{3.163}$$

or by solving

$$Q^{(1)}(\hat{\theta}) = -\frac{2}{nb} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right)(y_i - \hat{\theta}) = 0. \tag{3.164}$$

Since $\widehat{m}_{NW}(t)$ is a local least squares estimator, one may argue that it is not robust against outliers. For a general theory of robustness see, for example, Hampel et al. (1986), Huber (1981), and Huber and Ronchetti (2009). However, extending Equation (3.164), one can define more general estimators as solutions of

$$\frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right)\psi(y_i - \hat{\theta}) = 0, \tag{3.165}$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is a function such that

$$\mathbb{E}[\psi(G(Z))] = 0. \tag{3.166}$$

Setting $\psi(x) = x$, the Nadaraya–Watson estimator is obtained. If, however, ψ is bounded, then $\hat{\theta}$ is robust in the sense that the influence function is bounded (see, for example, Hampel et al. 1986 for the definition of influence functions). A standard example of a bounded ψ-function is the Huber-function

$$\psi_{Huber}(x) = \min(c, \max(x, -c)), 0 < c < \infty. \tag{3.167}$$

Robust kernel estimators as defined in (3.165) are considered for instance in Robinson (1984) and Härdle (1989) when the regression errors are iid, and in Boente and Fraiman (1989) under the assumption of strong mixing. In both cases, choosing robust (i.e., bounded) ψ-functions generally leads to a loss in asymptotic efficiency compared to the least squares solution. The situation is

different under long-memory. In what follows, we examine this special case.

The following notations will be used:

$$R(g^{(2)}) = \int_{\Delta}^{1-\Delta} [g^{(2)}(t)]^2 dt, 0 < \Delta < \frac{1}{2}, \tag{3.168}$$

$$\mu_2(K) = \int_{-1}^{1} x^2 K(x) dx, \tag{3.169}$$

$$J(K,d) = \int_0^1 \int_0^1 K(x)K(y)|x-y|^{2d-1} dx\, dy, \tag{3.170}$$

$$V_{\psi}(d) = \frac{1}{\mathbb{E}^2[\psi'(u)]} J(K,d). \tag{3.171}$$

To derive asymptotic properties of $\hat{m}(t) = \hat{\theta}$ defined in (3.165), Beran et al. (2003) use the following assumptions:

$$\gamma_Z(k) \underset{k\to\infty}{\sim} C_Z|k|^{2d-1} \text{ for some } d \in \left(0, \frac{1}{2}\right), \tag{3.172}$$

$$b \to 0, nb \to \infty. \tag{3.173}$$

The function ψ is assumed to be differentiable almost everywhere with respect to the Lebesgue measure,

$$\mathbb{E}[\psi(u_i)] = 0, \mathbb{E}[\psi^2(u_i)] < \infty, \mathbb{E}[\psi'(u_i)] \neq 0, \tag{3.174}$$

$$h_{\delta}(y) = \sup_{x\leq\delta} |\psi(y+x) - \psi(y)| \leq c \tag{3.175}$$

for some $\delta > 0$ and $0 < c < \infty$, and, for almost all y, we have

$$\lim_{\delta\to 0} h_{\delta}(y) = 0. \tag{3.176}$$

Moreover, it is assumed that $\psi(G(Z))$ has Hermite rank l ($l \geq 1$) with lth Hermite coefficient c_l, and there exist measurable functions M_2 and M_3 such that

$$|\psi''(x)| < M_2(x), \mathbb{E}[M_2(u_i)] < \infty \tag{3.177}$$

and

$$|\psi'''(x)| < M_3(x), \mathbb{E}[M_3(u_i)] < \infty. \tag{3.178}$$

Given these conditions we have the following asymptotic results.

Bias

In a first step, uniform consistency and an asymptotic formula for the bias can be shown by adapting standard techniques for kernel and M-estimators (see, for example, Beran 1991, Hall and Hart 1991, and Huber 1967). The following holds uniformly in $t \in (\Delta, 1-\Delta)$:

$$\mathbb{E}[\hat{m}_n(t) - m(t)] = \tfrac{1}{2} b^2 m''(t) I(K) + o(b^2) \tag{3.179}$$

Variance

On the other hand, the asymptotic formula for the variance is as follows. Let $\frac{1}{2}(1-l^{-1}) < d < \frac{1}{2}$. Then for uniformly in $t \in (\Delta, 1-\Delta)$,

$$(nb)^{1-2d} \mathbb{V}ar(\hat{m}_n(t)) = c_l^2 C_Z^l V_\psi(d). \tag{3.180}$$

To see how to derive the formula for the variance, one may start with the estimating equation

$$\frac{1}{nb} \sum \psi(y_i - \hat{g}(t)) K\left(\frac{t_i - t}{b}\right) = 0. \tag{3.181}$$

Then by Taylor expansion it can be shown that

$$\hat{m}_n(t) - m(t) = \mathbb{E}^{-1}[\psi(u)] \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) \psi(G(Z_i)) + R_n \tag{3.182}$$

with an asymptotically negligible remainder term R_n. The result then follows by applying the Hermite polynomial expansion of $\psi(G(Z))$ and showing that the first term in the expansion dominates asymptotically.

This leads to the following formulas for the asymptotic integrated mean squared error (imse) and the asymptotically optimal bandwidth. If $\frac{1}{2}(1-l^{-1}) < d < \frac{1}{2}$, then

$$\begin{aligned} \mathrm{IMSE}_\Delta &= \int_\Delta^{1-\Delta} \mathbb{E}\{[\hat{m}_n(t) - m(t)]^2\} dt \\ &= b^4 \frac{[m''(t) I(K)]^2}{4} + (nb)^{2d-1} c_l^2 C_Z^l V_\psi(d) \\ &\quad + o(\max(b^4, (nb)^{2d-1})) \end{aligned}$$

with

$$C_1 = \frac{[m''(t)I(K)]^2}{4}, C_2 = (nb)^{2d-1}c_l^2 C_Z^l V_\psi(d) \tag{3.183}$$

The asymptotically optimal bandwidth is of the form

$$b_{opt} = C_{opt} n^{(2d-1)/(5-2d)} \tag{3.184}$$

where

$$C_{opt} = \left[\frac{(1-2d)C_2}{4C_1}\right]^{1/(5-2d)}. \tag{3.185}$$

The asymptotic distribution of $\hat{m}_n$ is Gaussian only if $l = 1$. More specifically, we have the following asymptotic result. Let $t \in (0,1)$ and $\frac{1}{2}(1-l^{-1}) < d < \frac{1}{2}$. Then

$$(nb)^{1/2-d}\frac{\hat{m}_n(t) - m(t)}{\sigma_l} \underset{d}{\to} Z_l \tag{3.186}$$

where

$$\sigma_l = c_l\sqrt{C_Z^l V_\psi} \tag{3.187}$$

and Z_l is a Hermite process (see Rosenblatt 1984 and Taqqu 1975) of order l at time 1.

The most important case for practical applications is $l = 1$. Thus, suppose that $G(x) = x$. Then

$$c_1 = \mathbb{E}[Z\psi(Z)] = \mathbb{E}[\psi'(Z)] \tag{3.188}$$

This leads to

$$V_\psi(d) = c_1^{-2}J(K,d) \tag{3.189}$$

and

$$\sigma_1 = c_1\sqrt{C_Z V_\psi} = \sqrt{J(K,d)} \tag{3.190}$$

which is does not depend on ψ. We thus have the following result. Suppose that $G(x) = x$, and ψ has Hermite rank 1. Let σ_1 be given by (3.190). Then

$$(nb)^{1/2-d}\frac{\hat{m}_n(t) - m(t)}{\sigma_1} \underset{d}{\to} Z_1 \sim N(0,1). \tag{3.191}$$

The essential point of this result is that neither the standardization and nor the asymptotic distribution of $\hat{m}_n(t)$ depend on the function ψ. This means that all kernel M-estimators are asymptotically equivalent. In contrast to the case of independent or weakly dependent residuals, under long-range dependence using robust M-estimators does not lead to a loss of efficiency. This phenomenon has been observed first in Beran (1991) in the context of location estimation for stationary Gaussian subordination processes with long-memory (also see Giraitis et al. 1996 and Sibbertsen 1999 for extensions to linear regression).

The asymptotic results above are derived in Beran et al. (2003). The authors also discuss an application to local location estimation for wind speed data. Typically extremely strong wind speed measurements have a short duration, but tend to affect the Nadaraya–Watson estimator. To separate the effect of such extremes from "normal" wind speeds, it is therefore preferable to use local robust location estimation.

3.4.2 Local polynomial *M*-estimation

We again consider model (3.159) and (3.160). An alternative to kernel M-estimation is local polynomial M-estimation. Standard local polynomial estimation of $m(t)$ is obtained by solving

$$\frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right)\left(y_i - x'\hat{\beta}\right)x_j = 0 \; (j = 0, 1, \ldots, p), \quad (3.192)$$

where $p \in \mathbb{N}$, $x = (1, t, t^2, \ldots, t^p)'$ and $\hat{\beta} = \hat{\beta}(t) \in \mathbb{R}^p$, and setting $\hat{m}(t) = x'\hat{\beta}(t)$. In anology to kernel M-estimation, (3.192) can be generalized to

$$\frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right)\psi\left(y_i - x'\hat{\beta}\right)x_j = 0 \; (j = 0, 1, \ldots, p). \quad (3.193)$$

The asymptotic distribution of $\hat{\beta}$ is derived in Beran et al. (2002). In the following the same notation and assumptions as in the previous section are used. In particular, l denotes the Hermite rank of $\psi(G(Z))$, c_l is its lth Hermite coefficient, and the spectral

distribution f_Z of the latent Gaussian process Z_i is assumed to be such that

$$f_Z(\lambda) \underset{\lambda \to 0}{\sim} c_{f,Z} |\lambda|^{-2d} \tag{3.194}$$

with $0 < c_{f,Z} < \infty$ and

$$\frac{1}{2} - \frac{1}{2l} < d < \frac{1}{2}. \tag{3.195}$$

This implies that the spectral density f_ψ of $\psi(G(Z_i))$ has a pole at zero of the form $c_{f,\psi} |\lambda|^{-2d_l}$ with

$$0 < d_l = \frac{1}{2} + l\left(d - \frac{1}{2}\right) < \frac{1}{2}. \tag{3.196}$$

We define $g_{ij} = g_{ij}(t) = cov(\hat{\beta}_{i-1}, \hat{\beta}_{j-1})$ as

$$G_n = (g_{ij})_{i,j=1,\dots,p+1}. \tag{3.197}$$

Furthermore, let

$$p_{ij} = \frac{\sqrt{(2j-1)(2l-1)}}{j+l-1} \; (i,j = 1, \dots, p+1), \tag{3.198}$$

if $i + j$ is even, and otherwise set $p_{ij} = 0$, and define

$$P = (p_{ij})_{i,j=1,\dots,p+1}, \tag{3.199}$$

$$\kappa_{ij}(d_l) = \frac{\sqrt{(2i-1)\,(2l-1)}\Gamma(1-2d_l)}{\Gamma(d_l)\Gamma(1-d_l)} \tag{3.200}$$

and

$$Q = (q_{ij})_{i,j=1,\dots,p+1} \tag{3.201}$$

with

$$q_{ij} = \kappa_{ij}(d_l) \int_{-1}^{1} \int_{-1}^{1} x^{i-1} y^{j-1} |x-y|^{2d_l - 1} dx\, dy. \tag{3.202}$$

Finally, we define

$$D_n = (d_{ij})_{i,j=1,\dots,p+1} \tag{3.203}$$

where $d_{ii} = 0$ and

$$d_{jj} = 2\frac{(nb)^j}{2j-1} \; (j = 1, \dots, p+1). \tag{3.204}$$

To simplify presentation the results below consider the rectangular kernel

$$K(v) = \frac{1}{2}1\{-1 \leq v \leq 1\}. \tag{3.205}$$

The asymptotic covariance matrix of $\hat{\beta}$ is given below. Let $\hat{\beta} = \hat{\beta}(t)$ be the solution of (3.193). Then, for b such that $b \to 0$ and $nb \to \infty$, we have

$$\lim_{n\to\infty}(2nb)^{-2d_l}D_nG_nD_n = \frac{2\pi c_{f,Z}^l c_l^2}{l!\mathbb{E}^2[\psi'(G(Z))]}P^{-1}QP^{-1}. \tag{3.206}$$

For the most important case, for $m = 1$ and $u_i = Z_i$ we have $c_1 = \mathbb{E}[\psi'(G(Z))]$ so that (3.206) simplifies to

$$\lim_{n\to\infty}(2nb)^{-2d_l}D_nG_nD_n = \frac{2\pi c_{f,Z}^l}{l!}P^{-1}QP^{-1} \tag{3.207}$$

which does not depend on the function ψ. For $\hat{m}_n(t)$ we may thus formulate the following result.

Let $\hat{m}_n(t) = x'\hat{\beta}(t)$ where $\hat{\beta} = \hat{\beta}(t)$ is the solution of (3.193). Assume, furthermore, that the Hermite rank l of $\psi(G(.))$ is one. Then, for b such that $b \to 0$ and $nb \to \infty$,

$$\lim_{n\to\infty}(nb)^{1-2d}\mathbb{V}ar(\hat{m}_n(t)) = v(t), \tag{3.208}$$

where $0 < v(t) < \infty$ does not depend on ψ.

For an explicit expression for $v(t)$ see Beran and Feng (2002) and Ghosh (2001). As for kernel M-estimation, the essence of the result that under long-memory and Gaussian subordination of Hermite rank one, robust local polynomial estimators of trend functions are asymptotically equivalent to the standard non-robust local polynomial estimator. This result can also be generalized to derivatives of m.

4

Semiparametric Regression

4.1 Partial linear models with constant slope

A partial linear model is a regression model containing a smooth nonparametric component and a linear parametric regression component. It is thus a semiparametric model, where the nonparametric component is unspecified except for some regularity conditions such as continuity, differentiability, etc. Below is an example of a partial linear model:

$$y_i = \mathbf{x}_i'\beta + m(t_i) + u_i \tag{4.1}$$

where $\mathbf{x}_i \in \mathbb{R}^p$ is a column vector of explanatory variables and $\beta \in \mathbb{R}^p$ is a column vector of regression parameters, defined as

$$\mathbf{x}_i' = (x_{1,i}, x_{2,i}, \ldots, x_{p,i}), p \geq 1, \tag{4.2}$$

$$\beta' = (\beta_1, \beta_2, \ldots, \beta_p). \tag{4.3}$$

The nonparametric component m is a smooth function to be estimated in $\mathbb{C}^2[0,1]$. The u_i are regression errors with zero mean and constant variance. We consider the case when u_i is a stationary long-memory process with a covariance function γ_u and a spectral density ϕ_u:

$$\gamma_u(k) = \mathbb{C}ov(u_i, u_{i+k}) = \int exp(\imath k\lambda)\phi_u(\lambda)d\lambda, \imath = \sqrt{-1}, \tag{4.4}$$

$$\phi_u(\lambda) \sim c_u|\lambda|^{-\alpha_u} \text{ as } \lambda \to 0, \tag{4.5}$$

where for two functions $a(\nu)$ and $b(\nu)$, $a(\nu) \sim b(\nu)$ implies $a(\nu)/b(\nu)$ converges to a constant, $c_u > 0$ is a constant, and $0 \leq \alpha_u < 1$.

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

When β and m are unknown but the errors are uncorrelated is addressed in Speckman (1988). This author suggests a $\sqrt{n}$ consistent estimator for β when the explanatory variables contain a rough component. Beran and Ghosh (1998) generalize Speckman's result to the case when the regression errors have long-memory. These authors show that even under long-memory, a $\sqrt{n}$ rate of convergence for the estimated slope holds. For related ideas in the context of errors in covariates, see Carroll et al. (1999), Hastie and Tibshirani (1990), among others. Also see Ruppert et al. (2009).

To see how the slope parameter in the partial linear model above is estimated, we start with the data at hand: thus we have observations $(\mathbf{x}_i', y_i)$ at time points $i = 1, 2, \ldots, n$. As usual $t_i = i/n$ will denore rescaled times. Define new notations

$$\mathbf{x}_i' = (x_{1,i}, x_{2,i}, \ldots x_{p,i}), i = 1, 2, \ldots, n, \tag{4.6}$$

$$\mathbf{y}' = (y_1, y_2, \ldots, y_n), \tag{4.7}$$

$$\mathbf{m}' = (m(t_1), m(t_2), \ldots, m(t_n)), t_i = i/n, \tag{4.8}$$

$$\mathbf{u}' = (u_1, u_2, \ldots, u_n). \tag{4.9}$$

Next define the $n \times p$ design matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{M} + \eta \tag{4.10}$$

where M is a deterministic matrix of order $n \times p$ and η is a random matrix of the same order. The elements of η are zero mean random variables. The ith row of $\mathbf{X}$ is $\mathbf{x}_i'$ defined above and the columns of $\mathbf{M}$ are $(\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_p)$, defined as

$$\mathbf{m}_j' = (m_j(t_1), m_j(t_2), \ldots, m_j(t_n)),\ j = 1, 2, \ldots, p, \tag{4.11}$$

whereas the ith row of $\mathbf{M}$ is

$$(m_1(t_i), m_2(t_i), \ldots, m_p(t_i)), i = 1, 2, \ldots, n. \tag{4.12}$$

The random matrix η has the columns

$$\eta = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p) \tag{4.13}$$

where

$$\mathbf{e}_j' = (e_{j,1}, e_{j,2}, \ldots, e_{j,n}),\ j = 1, 2, , \ldots, p, \tag{4.14}$$

and rows

$$\mathbf{e}_i' = (e_{1,i}, e_{2,i}, \ldots, e_{p,i}). \tag{4.15}$$

Then the partial linear model can be rewritten as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{m} + \mathbf{u} = (\mathbf{M}\beta + \mathbf{m}) + (\eta\beta + \mathbf{u}). \tag{4.16}$$

Note that in the last formula, $\mathbf{M}\beta + \mathbf{m}$ is deterministic whereas $\eta\beta + \mathbf{u}$ is stochastic. The expected value of $\mathbf{y}$ is $\mathbf{M}\beta + \mathbf{m}$ so that smoothing $\mathbf{y}$ leads to estimation of the deterministic component, namely its expectation. At the next step one obtains the regression residuals in the model (4.16. On the other hand, from (4.10) the residuals in $\mathbf{X}$ can be estimated by "detrending" the data series containing the values of the explanatory variables. Finally, the residuals are used in a regression through the origin, to estimate β. The logic behind this method of course stems from the fact that when computing correlation between two random variables, the means have to be estimated correctly, because otherwise a bias would result. A simulation study done in Beran and Ghosh (1998) illustrates this point.

Specifically, consider the Nadaraya–Watson kernel (see Gasser et al. 1985)

$$K(t_i, t_j, n, b) = \frac{w\left(\frac{t_i - t_j}{b}\right)}{n^{-1}\sum_{i=1}^{n} w\left(\frac{t_l}{b}\right)}. \tag{4.17}$$

Define the kernel matrix

$$\mathbf{K} = [K(t_i, t_j, n, b)]_{i,j=1,2,\ldots,n} \tag{4.18}$$

where w is a univariate kernel; in particular, it is a bounded, nonnegative, symmetric and piecewise continuous function with support $[-1, 1]$ such that $\int_{-1}^{1} w(s)ds = 1$. Define the regression residuals

$$\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{K})\mathbf{X}, \tag{4.19}$$

$$\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{K})\mathbf{y}, \tag{4.20}$$

and the semiparametric regression estimate of the slope parameter β as

$$\widehat{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{4.21}$$

Beran and Ghosh (1998) prove consistency of the above estimator and illustrate the constant slope model with an application. In what follows, we digress from the constant slope model and

let the slope be time-dependent. This is a slight generalization of the above partial linear model and allows us to see if the linear dependence (i.e., the slope β) may change over time. This is discussed in the next section.

4.2 Partial linear models with time-varying slope

In the previous section, we considered a partial linear model where there is a linear dependence of the response variable Y on the explanatory variable X and the slope parameter β is a constant. In the literature, the constant slope case with stationary short-memory errors is considered in Speckmann (1988) whereas the long-memory case is considered among others in Beran and Ghosh (1998) and Aneiros-Pérez et al. (2004); also see Robinson (1988) and González-Manteiga and Aneiros-Pérez (2003) and references therein for further background information. In some situations, however, one may postulate that, over time, the strength of this linear dependence may change smoothly.

Consider, for instance, the time series in Figure (4.1). The observations are land and sea surface temperatures made available through the homepage of the Met Office, UK:

`http://www.metoffice.gov.uk/research`

The time series are global temperature mean annual anomalies, during the years 1856 through 2014. We seek to estimate the slope function $\beta(t)$ taking the ocean temperatures as y and the land temperatures as x. Optimal bandwidth selection as well as appropriate hypothesis testing will have to be carried out in further detailed analysis, so that the results of this analysis are to be treated as being exploratory. Using an arbitrary bandwidth $b = 0.2$ there is an indication (right panel, Figure 4.4) of a change (increase) in the slope parameter over time until around the year 1900, after which it seems to reach a relatively constant level. In contrast, the plot in the left panel in Figure (4.4) was obtained by fitting the constant slope model as in the previous section, and this model indicates a positive value of the (constant) slope parameter for the entire duration of 1856:2014. The two plots

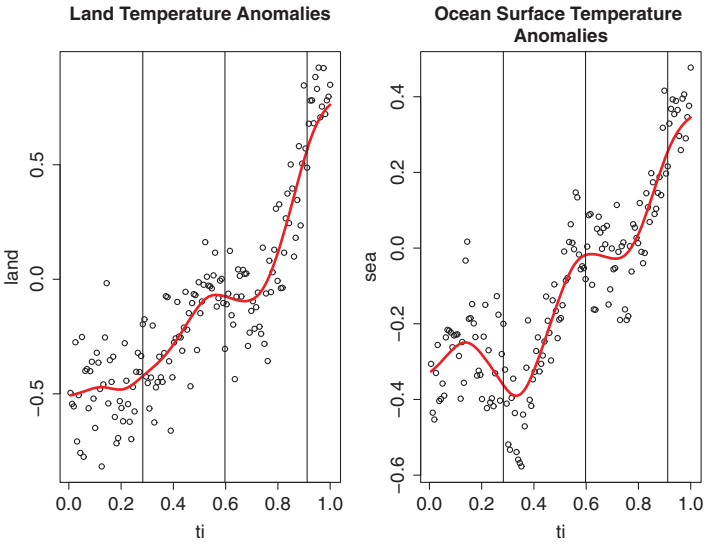

Figure 4.1 Global temperature series 1856:2014: mean annual anomalies, land and ocean. *Source:* Data from Met Office, UK.

in Figure (4.2) show the residuals (left: constant slope model, right: time-varying slope model) and Figure (4.3) shows the normal probability plots for the two fits. Another generalization that we consider here is, unlike in the previous section, where the regression errors are realizations of a stationary process, we let the regression errors be locally stationary, being time-dependent transformations of a latent Gaussian process. The resulting class of marginal error distributions is then vast, consisting of distributions that may change with time, assuming arbitrary shapes, and in particular may be non-Gaussian. The normal distribution is a member of this class.

In addition, we also derive some results under two different correlation types, namely short-memory and long-memory correlations.

As we shall see, uniform consistency of the trend estimates becomes a useful property for estimating the slope function. Following the idea of Parzen (1962), using a characteristic function based approach, we provide a simple proof of weak consistency of the trend estimates as well as of the estimate of the slope function.

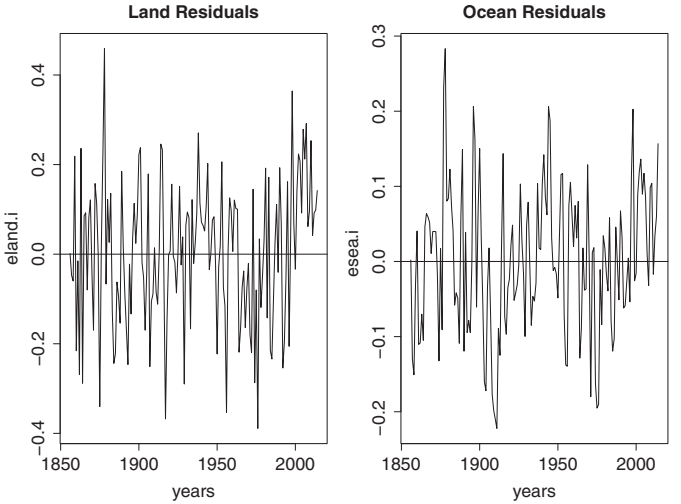

Figure 4.2 Global temperature 1856:2014: land and ocean. The figures show residuals after fitting a partial linear model. Left: constant slope model; Right: time-varying slope model. *Source:* Data from Met Office, UK.

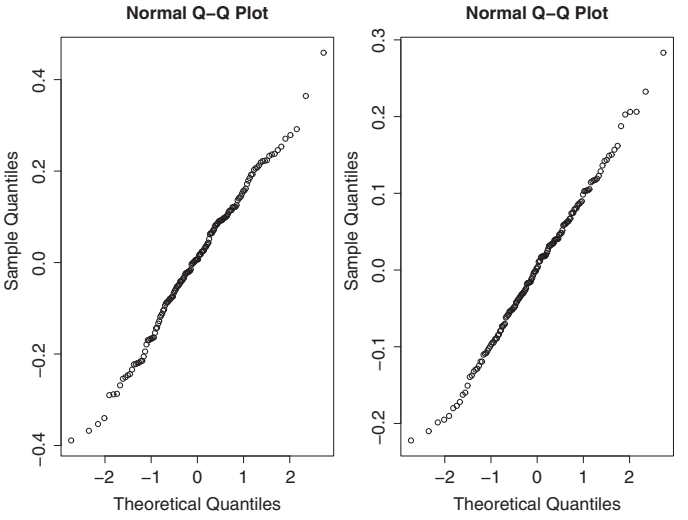

Figure 4.3 Global temperature 1856:2014: land and ocean. The figures show normal probability plots of the residuals after fitting a partial linear model. Left: constant slope model; Right: time-varying slope model. *Source:* Data from Met Office, UK.

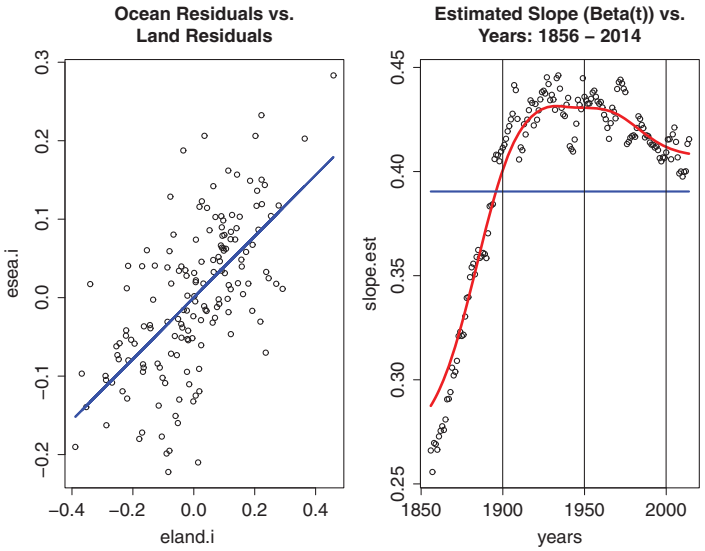

Figure 4.4 Global temperature 1856:2014: land and ocean. The figures show estimates of β after fitting a partial linear model. Left: constant slope model; Right: time-varying slope model. *Source:* Data from Met Office, UK.

We start with a continuous index bivariate process $\{x(T), y(T)\}, T \in \mathbb{R}_+$. Let the observations be available at the discrete time points $T_i = i \in \{1, 2, \ldots, n\}$. Let $x(T_i) = x_i$ and $Y(T_i) = y_i$, $t_i = T_i/n = i/n$ denoting rescaled times. It should be noted that the results of this section can also be generalized to the case when the continuous index bivariate process $\{(x(T), y(T)), T \in \mathbb{R}_+\}$ is observed at irregularly spaced time points. For simplicity of presentation, we let our observations be evenly spaced.

Consider the partial linear model with a smooth slope as follows:

$$y_i = m(t_i) + \beta(t_i) \cdot x_i + u_i \tag{4.22}$$

and

$$x_i = h(t_i) + v_i. \tag{4.23}$$

Here β, m, and h are continuous functions on $[0, 1]$, and $u(T_i) = u_i$ and $v(T_i) = v_i$ are zero mean errors with finite fourth moments.

Equations (4.22) and (4.23) define a partial linear model and the problem is estimation of $\beta(t)$ at $t \in (0, 1)$. Following Ghosh (2014), here we address estimation of this time-varying slope function.

We impose further assumptions on the errors. Let u_i and v_i be time-dependent one-dimensional transformations of some latent stationary Gaussian processes. The transformation, however, is unknown and may be nonlinear. In particular, due to the transformation the errors may have marginal distributions that may change with time, be non-Gaussian, and assume arbitrary shapes. This model for the errors is a slight generalization of Taqqu (1975), where stationarity of the latent Gaussian process is inherited by the subordinated process. Here we let this transformation be time-dependent. This has the advantage of having a flexible marginal distribution function that may change over time. A related statistical problem is a nonparametric prediction of the marginal function at a future time point (Ghosh and Draghicescu 2002b); also see Beran and Ocker 1999). For relevant background information on empirical processes arising from nonlinear functionals of Gaussian processes see Breuer and Major (1983), Csörgő and Mielniczuk (1996), Dehling and Taqqu (1989), Dobrushin and Major (1979), Giraitis and Surgailis (1985), Major (1981), and Taqqu (1975, 1979).

As in the previous section, to estimate the constant slope parameter, a "regression through zero" model is fitted to the regression residuals. However, since the slope parameter is a function of time, to estimate $\beta(t)$, we use a kernel smoothed version of Speckman (1988). We consider both short-memory and long-memory correlations in the latent Gaussian processes and address consistency of the nonparametric curve estimates.

Background information on nonparametric curve estimates under long-memory and related references are given in Beran and Feng (2002), Csörgő and Mielniczuk (1995), Ghosh (2001), Giraitis and Koul (1997), Giraitis et al. (2012, Chapter 12), Guo and Koul (2007), Hall and Hart (1990), and Robinson and Hidalgo (1997).

To prove uniform consistency of the nonparametric curve estimates, we use a kernel that has an absolutely integrable characteristic function. This simple idea leads to a very simple proof of uniform consistency; see Parzen (1962) and Bierens (1983).

Other important work on this topic include Hall and Hart (1990), Mack and Silverman (1982), Nadaraya (1965), Schuster (1970), and Silverman (1978); also see references therein.

The partial linear model considered in this section as well as in the previous section is in fact a special case of a random design regression model. For research on this topic under long-range dependence see Csörgő and Mielniczuk (1999). For a general background on kernel smoothing see Silverman (1986) and Wand and Jones (1995). General reviews of long-memory processes, statistical applications, and theoretical backgrounds can be found in Beran (1994), Beran et al. (2013), Giraitis et al. (2012), and Leonenko (1999).

4.2.1 Estimation

From the partial linear model, by rewriting we have

$$y_i = m(t_i) + \beta(t_i)(h(t_i) + \nu_i) + u_i \tag{4.24}$$

$$= g(t_i) + \epsilon_i, \tag{4.25}$$

where

$$g(t_i) = m(t_i) + \beta(t_i) \cdot h(t_i) \tag{4.26}$$

$$\epsilon_i = u_i + \beta(t_i) \cdot \nu_i \tag{4.27}$$

so that

$$\widehat{g}(t) = \frac{1}{nb_g} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b_g}\right) y_i, \tag{4.28}$$

$$\widehat{h}(t) = \frac{1}{nb_h} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b_h}\right) x_i. \tag{4.29}$$

Consider the regression residuals

$$\widehat{\epsilon}_i = y_i - \widehat{g}(t_i), \tag{4.30}$$

$$\widehat{\nu}_i = x_i - \widehat{h}(t_i). \tag{4.31}$$

Then the slope estimator is

$$\widehat{\beta}(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) \widehat{\nu}_i \widehat{\epsilon}_i}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) \widehat{\nu}_i^2} \tag{4.32}$$

where as $n \to \infty$, $b \to 0$ and $nb \to \infty$ and K is defined below.

The kernel K and the bandwidths b_g and b_h satisfy the following conditions:

Kernel: $K(s) \geq 0$ if $-1 < s < 1$, and $K(s) = 0$ otherwise, $\int_{-1}^{1} K(s)ds = 1$ and $K(-s) = K(s)$, for all s.

Kernel characteristic function: K is a symmetric probability density function on $\mathbb{R}$ with a characteristic function that is absolutely integrable on the whole real line.

Bandwidths: As $n \to \infty$, $b_g \to 0$, $b_h \to 0$, $nb_g \to \infty$, and $nb_h \to \infty$.

Next, we impose further conditions on the model parameters and other quantities.

4.2.2 Assumptions

As usual, for two sequences a_n and b_n, $a_n \sim b_n$ will imply that a_n/b_n converges to a constant as $n \to \infty$. We make the following assumptions:

Trend and slope: The trend functions $m(t)$ and $h(t)$ as well as the slope function $\beta(t)$ where $t \in [0, 1]$ are in $\mathbb{C}^2[0, 1]$.

Errors: The errors $u(T)$ and $v(T)$ are independent. Let $u(T_i) = u_i$ and $v(T_i) = v_i$, where $T_i = i = 1, 2, \ldots, n$:

$$\mathbb{E}(u_i) = 0, \mathbb{E}(v_i) = 0, \tag{4.33}$$

$$\mathbb{V}ar(u_i) < \infty, \mathbb{V}ar(v_i) < \infty, \tag{4.34}$$

$$\mathbb{E}\left(v_i^4\right) < \infty. \tag{4.35}$$

Gaussian subordination: Let $Z(T)$ and $W(T)$ for $T \in \mathbb{R}_+$ be mutually independent, zero mean, unit variance, continuous time stationary latent Gaussian processes. Let $Z_i = Z_i$ and $W(T_i) = W_i$ where $T_i = i = 1, 2, \ldots, n$ and $t_i = i/n$. We assume

$$u_i = G_u(Z_i, t_i), \tag{4.36}$$

$$v_i = G_v(W_i, t_i), \tag{4.37}$$

where $G_u : \mathbb{R} \times [0, 1] \to \mathbb{R}$ and $G_v : \mathbb{R} \times [0, 1] \to \mathbb{R}$ are unknown & square integrable with respect to the standard normal density.

Hermite polynomial expansions: G_u and G_v allow for Hermite polynomial expansions (e.g., Szegő 1975)

$$u_i = G_u(Z_i, t_i) = \sum_{l=r_u}^{\infty} \frac{c_l(t_i)}{l!} H_l(Z_i), \tag{4.38}$$

$$v_i = G_v(W_i, t_i) = \sum_{l=r_v}^{\infty} \frac{d_l(t_i)}{l!} H_l(W_i). \tag{4.39}$$

Here $c_l, d_l \in \mathbb{C}^2[0,1]$ are time-dependent Hermite coefficients, H_l is the Hermite polynomial of degree $l \geq 1$, $r_u \geq 1$ and $r_v \geq 1$ are Hermite ranks, i.e., r_u and r_v are the smallest positive integers such that c_{r_u} and d_{r_v} are not equal to zero.

Covariance functions: $Z(T)$ and $W(T)$ with $T \in \mathbb{R}_+$ have covariances

$$\mathbb{C}ov(Z(T), Z(T+h)) = \gamma_Z(|h|), \tag{4.40}$$

$$\mathbb{C}ov(W(T), W(T+h)) = \gamma_W(|h|), \tag{4.41}$$

where $h \in \mathbb{R}$.

Correlations: short-memory in the latent Gaussian processes: The integrals $\int_{-\infty}^{\infty} \gamma_Z(h)dh$ and $\int_{-\infty}^{\infty} \gamma_W(h)dh$ converge to positive constants. The infinite sums of the autocorrelations converge; e.g., $\sum_{k=-\infty}^{\infty} \gamma_Z(|k|) < \infty$ and $\sum_{k=-\infty}^{\infty} \gamma_W(|k|) < \infty$.

Correlations: long-memory in the latent Gaussian processes: The integrals $\int_{-\infty}^{\infty} \gamma_Z(h)dh$ and $\int_{-\infty}^{\infty} \gamma_W(h)dh$ diverge. The autocorrelations in Z and W are expressed as hyperbolically decaying functions of their lags, i.e., when the Hurst parameters H_u and H_v are such that $1/2 < H_u, H_v < 1$, then for $h \in \mathbb{R}$,

$$\gamma_Z(|h|) \sim C_Z|h|^{2H_u-2}, \text{as } |h| \to \infty \tag{4.42}$$

and

$$\gamma_W(|h|) \sim C_W|h|^{2H_v-2}, \text{as } |h| \to \infty. \tag{4.43}$$

For the lags $k \in \mathbb{N}$, $\sum_{k=-\infty}^{\infty} \gamma_Z(|k|) = \infty$ and $\sum_{k=-\infty}^{\infty} \gamma_W(|k|) = \infty$.

Covariance functions: short-memory: As $n \to \infty$ (short-memory in $Z(T)$ and $W(T)$ respectively), for $x \in \mathbb{R}$, r_u, $r_v \in \mathbb{N}_+$,

$$\sum_{l=r_u}^{\infty} \int_{-1}^{1} \int_{-1}^{1} \left| \gamma_Z \left(nb \left| s_1 - s_2 + \frac{x}{b} \right| \right) \right|^l ds_1 ds_2 = O((nb)^{-1}) \tag{4.44}$$

and

$$\sum_{l=r_u}^{\infty} \int_{-1}^{1} \int_{-1}^{1} \left| \gamma_W \left(nb \left| s_1 - s_2 + \frac{x}{b} \right| \right) \right|^l ds_1 ds_2 = O((nb)^{-1}), \tag{4.45}$$

where as $n \to \infty$, $b \to 0$ and $nb \to \infty$.

Covariance functions: long-memory: As $n \to \infty$ (long-memory in $Z(T)$ and $W(T)$ respectively), for $x \in \mathbb{R}$, $l \in \mathbb{N}_+$

$$\int_{-1}^{1} \int_{-1}^{1} \left| \gamma_Z \left(nb \left| s_1 - s_2 + \frac{x}{b} \right| \right) \right|^l ds_1 ds_2 = O\big((nb)^{l(2H_u-2)}\big) \tag{4.46}$$

and

$$\int_{-1}^{1} \int_{-1}^{1} \left| \gamma_W \left(nb \left| s_1 - s_2 + \frac{x}{b} \right| \right) \right|^l ds_1 ds_2 = O\big((nb)^{l(2H_v-2)}\big) \tag{4.47}$$

where $0.5 < H_u, H_v < 1$ and as $n \to \infty$, $b \to 0$ and $nb \to \infty$.

Further assumptions on bandwidths: short-memory: As $n \to \infty$, $n^{-1/2} b_g^{-1} \to 0$ and $n^{-1/2} b_h^{-1} \to 0$.

Further assumptions on bandwidths: long-memory: As $n \to \infty$, $n^{r_u(H_u-1)} b_g^{-1} \to 0$, $n^{r_v(H_v-1)} b_g^{-1} \to 0$, and $n^{r_u(H_u-1)} b_h^{-1} \to 0$ where $r_u, r_v \in \mathbb{N}_+$ and $1/2 < H_u, H_v < 1$.

Further remarks on the Hermite polynomials:

$$H_l(z) = (-1)^l e^{z^2/2} \frac{d^l}{dz^l} e^{-z^2/2}, l \in \mathbb{N}_+, z \in \mathbb{R} \tag{4.48}$$

$\forall\, i,\ j, l, k \in \mathbb{N}_+$, satisfy the following:

$$\mathbb{E} H_l(Z_i) = \mathbb{E} H_l(W_i) = 0, \tag{4.49}$$

$$\mathbb{V}ar\{H_l(Z_i)\} = \mathbb{V}ar\{H_l(W_i)\} = l!, \tag{4.50}$$

$$\mathbb{C}ov\{H_l(Z_i), H_k(Z_j)\} = \mathbb{C}ov\{H_l(W_i), H_k(W_j)\} = 0, \quad l \neq k, \tag{4.51}$$

$$\mathbb{C}ov\{H_l(Z_i), H_l(Z_j)\} = l!\gamma_Z^l(|i-j|), \tag{4.52}$$

$$\mathbb{C}ov\{H_l(W_i), H_l(W_j)\} = l!\gamma_W^l(|i-j|), \tag{4.53}$$

$$\mathbb{C}ov\{H_l(Z_i), H_k(W_j)\} = 0. \tag{4.54}$$

Miscellaneous facts: Since u_i and v_i have finite variances,

$$\mathbb{V}ar(u_i) = \mu_{2u}(t_i) = \sum_{l=r_u}^{\infty} \frac{c_l^2(t_i)}{l!} < \infty, \tag{4.55}$$

$$\mathbb{V}ar(v_i) = \mu_{2v}(t_i) = \sum_{l=r_v}^{\infty} \frac{d_l^2(t_i)}{l!} < \infty \tag{4.56}$$

and their covariances are

$$\mathbb{C}ov(u_i, u_j) = \gamma_u(i, j; t_i, t_j) = \sum_{l=r_u}^{\infty} \frac{c_l(t_i)c_l(t_j)}{l!}\{\gamma_Z(|T_i - T_j|)\}^l, \tag{4.57}$$

$$\mathbb{C}ov(v_i, v_j) = \gamma_v(i, j; t_i, t_j) = \sum_{l=r_v}^{\infty} \frac{d_l(t_i)d_l(t_j)}{l!}\{\gamma_W(|T_i - T_j|)\}^l. \tag{4.58}$$

Note that the squared error v_i^2 is Gaussian subordinated and due to the finite fourth moment assumption, v_i^2 with its mean subtracted allows for a Hermite polynomial expansion

$$v_i^2 - \mu_{2v}(t_i) = \sum_{l=q_v}^{\infty} \frac{p_l(t_i)}{l!} H_l(W_i). \tag{4.59}$$

In the above expansion, the Hermite coefficients p_l are in $\mathbb{C}^2[0, 1]$ and q_v is Hermite rank of the centered v_i^2 process. The variance of v_i^2 is

$$\mu_{4v}(t_i) = \mathbb{E}\left(v_i^2 - \mu_{2v}(t_i)\right)^2 = \sum_{l=q_v}^{\infty} \frac{p_l^2(t_i)}{l!} < \infty. \tag{4.60}$$

Also the combined error $\epsilon_i = u_i + \beta(t_i)v_i$ has finite variance and since Z_i and W_i are independent,

$$\mathbb{V}ar(\epsilon_i) = \mu_{2\epsilon}(t_i) = \mu_{2u}(t_i) + \beta^2(t_i)\mu_{2v}(t_i) \tag{4.61}$$

and

$$\begin{aligned}\mathbb{C}ov\{\epsilon_i, \epsilon_j\} &= \gamma_\epsilon(i,j;t_i,t_j) \\ &= \gamma_u(i,j;t_i,t_j) + \beta(t_i)\cdot\beta(t_j)\cdot\gamma_v(i,j;t_i,t_j). \end{aligned} \tag{4.62}$$

Now if $|T_i - T_j| \to \infty$ but $t_i \to t$ and $t_j \to t$, then since $1/2 < H_u < 1$,

$$\gamma_u(i,j;t_i,t_j) \sim C_Z^{r_u}\cdot[c_{r_u}(t)]^2\cdot|T_i - T_j|^{(2H_u-2)r_u}. \tag{4.63}$$

This means that u_i will have long-memory if and only if

$$1 - \frac{1}{2r_u} < H_u. \tag{4.64}$$

Similarly, v_i will have long-memory if and only if

$$1 - \frac{1}{2r_v} < H_v. \tag{4.65}$$

If u_i has long-memory, then ϵ_i will be long-range dependent. If both u_i and v_i have long-memory and $\beta(t_i) \neq 0$, then the stronger long-memory parameter will dominate. Let $|T_i - T_j| \to \infty$, $t_i, t_j \to t$. Then if $1 - 1/2r_u < H_u$ and $1 - 1/2r_v < H_v$,

$$\begin{aligned}\gamma_\epsilon(i,j;t_i,t_j) \sim\ & C_Z^{r_u}\cdot[c_{r_u}(t)]^2\cdot|T_i - T_j|^{(2H_u-2)r_u} \\ & + \beta^2(t)\cdot C_W^{r_v}\cdot[d_{r_v}(t)]^2\cdot|T_i - T_j|^{(2H_v-2)r_v}. \end{aligned} \tag{4.66}$$

Let $a_u = (2H_u - 2)r_u$ and $a_v = (2H_v - 2)r_v$. This means that if $a_u > a_v$ or if $\beta(t) = 0$, then

$$\gamma_\epsilon(i,j;t_i,t_j) \sim C_Z^{r_u}\cdot[c_{r_u}(t)]^2\cdot|T_i - T_j|^{(2H_u-2)r_u} \tag{4.67}$$

and if $a_v > a_u$ and $\beta(t) \neq 0$, then

$$\gamma_\epsilon(i,j;t_i,t_j) \sim \beta^2(t)\cdot C_W^{r_v}\cdot[d_{r_v}(t)]^2\cdot|T_i - T_j|^{(2H_v-2)r_v}. \tag{4.68}$$

If $\beta(t_i) = 0$, $\epsilon_i = u_i$, i.e., then the error ϵ_i will have a long-memory ($H_u > 1 - 1/(2r_u)$) or short-memory property as the latent Gaussian process Z_i. When $\beta(t)$ is not zero, ϵ_i will be long-range dependent unless both u_i and v_i have short-memory correlations.

Due to the transformations, the marginal distributions of u_i and v_i may be non-Gaussian and change with time, i.e.,

$$P(u_i \le s) = F_u(s,t_i), \tag{4.69}$$

$$P(v_i \le s) = F_v(s,t_i), \tag{4.70}$$

where $F_u : \mathbb{R} \times [0,1] \to [0,1]$ and $F_v : \mathbb{R} \times [0,1] \to [0,1]$ satisfy some differentiability conditions.

4.2.3 Asymptotics

4.2.3.1 Pointwise weak consistency

First of all note that the trend estimates are consistent. We focus on the estimator for m. The same arguments may be used for h. Note that first of all, since $b \to 0$ and $nb \to \infty$ as $n \to \infty$, for $r \in \mathbb{N}_+$,

$$\left| \frac{1}{nb} \sum_{i=1}^{n} \left(\frac{t_i - t}{b} \right)^r K\left(\frac{t_i - t}{b} \right) - \int_{-1}^{1} s^r K(s) ds \right| = O\left(\frac{1}{nb} \right) \tag{4.71}$$

By taking expectations,

$$\mathbb{E}(\widehat{g}(t)) = \sum_{i=1}^{n} K((t_i - t)/b) g(t_i). \tag{4.72}$$

Now using Taylor series expansion of $g(t_i)$ around t,

$$\widehat{g}(t) = g(t) + \left\{ \frac{b_g^2}{2} \cdot \frac{d^2}{dt^2} g(t) \cdot \int_{-1}^{1} s^2 K(s) ds \right\} + o\left(b_g^2\right) + O\left(\frac{1}{nb_g} \right) + \frac{1}{nb_g} \sum_{i=1}^{n} K\left(\frac{t_i - t}{b_g} \right) \epsilon_i. \tag{4.73}$$

We can absorb $O(1/(nb_g))$ into $o(b_g^2)$ if we let $nb_g^3 \to \infty$ as $n \to \infty$. Since $\mathbb{E}\epsilon_i = 0$,

$$Bias(\widehat{g}(t)) = O\left(b_g^2\right). \tag{4.74}$$

Similarly,

$$Bias(\widehat{h}(t)) = O\left(b_h^2\right). \tag{4.75}$$

To derive the covariance at t and s (the expression for the variance follows by substituting $t = s$), recall that

$$\gamma_\epsilon(i, j; t_i, t_j) = \gamma_u(i, j; t_i, t_j) + \beta(t_i)\beta(t_j)\gamma_v(i, j; t_i, t_j). \tag{4.76}$$

Since $K(s) = 0$ when $|s| > 1$,

$$\mathbb{C}ov(\widehat{g}(t), \widehat{g}(s)) = \frac{1}{(nb_g)^2} \sum_{i=-n(t+b_g)}^{n(t-b_g)} \sum_{j=-n(s+b_g)}^{n(s-b_g)} \left[K\left(\frac{t_i - t}{b_g}\right) \times K\left(\frac{t_j - s}{b_g}\right) \gamma_\epsilon(i, j; t_i, t_j) \right] \tag{4.77}$$

Now, if $\beta(t) \neq 0$, ϵ_i will have short-memory if both u_i and v_i have short-memory. This means that $\beta(t) < \infty$ implies

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \gamma_\epsilon(i, j; t_i, t_j) < \infty. \tag{4.78}$$

Due to the finite variance assumptions on the regression errors,

$$\sum_{l=r_u}^{\infty} c_l^2(t)/l! < \infty, \tag{4.79}$$

$$\sum_{l=r_v}^{\infty} d_l^2(t)/l! < \infty, \tag{4.80}$$

and also by the Cauchy–Schwarz inequality,

$$\sum_{l=r_u}^{\infty} c_l(t)c_l(s)/l! < \infty, \tag{4.81}$$

$$\sum_{l=r_v}^{\infty} d_l(t)d_l(s)/l! < \infty, \tag{4.82}$$

for $t, s \in [-1, 1]$. Now we write

$$|i - j| = nb_g \left| \left(\frac{t_i - t}{b_g} - \frac{t_j - s}{b_g} \right) + \frac{t - s}{b} \right|. \tag{4.83}$$

This means that

$$\mathbb{C}ov(\widehat{g}(t), \widehat{g}(s)) = D_n(t, s) + o(D_n(t, s)) \tag{4.84}$$

where

$$D_n(t, s) = \sum_{l=r_u}^{\infty} \frac{c_l(t)c_l(s)}{l!} \int_{-1}^{1} \int_{-1}^{1} \left[K(s_1)K(s_2)\gamma_Z^l \times (nb_g|s_1 - s_2 + (t - s)/b_g|) \right] ds_1 ds_2 + \beta(t)\beta(s)$$

$$\times \sum_{l=r_v}^{\infty} \frac{d_l(t)d_l(s)}{l!} \int_{-1}^{1}\int_{-1}^{1} \big[K(s_1)K(s_2)$$
$$\times \gamma_W^l(nb_g|s_1 - s_2 + (t-s)/b_g|)\big] ds_1 ds_2$$
$$= O\left(\frac{1}{(nb_g)}\right) \tag{4.85}$$

In the case of long-memory, $\gamma_Z(|k|)$ and $\gamma_W(|k|)$ decay hyperbolically with increasing lags k, and the infinite sums of the auto covariances are non-summable. Consider $\beta(t) \neq 0$ (the case $\beta(t) = 0$ can be derived similarly). Standard arguments involving Taylor series expansions and approximation of Riemann sums by double integrals can be used to derive the following. We have

$$\begin{aligned}
\mathbb{V}ar(\widehat{g}(t)) &= \frac{1}{n^2 b_g^2} \sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{t_i - t}{b_g}\right) K\left(\frac{t_j - t}{b_g}\right) \gamma_\epsilon(i,j;t_i,t_j) \\
&\sim \frac{1}{n^2 b_g^2} \sum_{l=r_u}^{\infty} \left[C_Z^l \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ c_l(t_i)c_l(t_j) K\left(\frac{t_i - t}{b_g}\right) \right.\right. \\
&\quad \left.\left. \times K\left(\frac{t_j - t}{b_g}\right) |i-j|^{l(2H_u-2)} \right\} \right] + \frac{1}{n^2 b_g^2} \sum_{l=r_v}^{\infty} \left[C_W^l \right. \\
&\quad \times \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ \beta(t_i)\beta(t_j) d_l(t_i) d_l(t_j) K\left(\frac{t_i - t}{b_g}\right) \right. \\
&\quad \left.\left. \times K\left(\frac{t_j - t}{b_g}\right) |i-j|^{l(2H_v-2)} \right\} \right] \\
&= A_n(t) + o(A_n(t))
\end{aligned} \tag{4.86}$$

as $n \to \infty$. Now, if $a_u > a_v$, where $a_u = r_u(2H_u - 2)$ and $a_v = r_v(2H_v - 2)$,

$$\begin{aligned}
A_n(t) &= (nb_g)^{r_u(2H_u-2)} \frac{c_{r_u}^2(t)}{r_u!} C_Z^{r_u} \\
&\quad \times \int_{-1}^{1}\int_{-1}^{1} K(s_1)K(s_2)|s_1 - s_2|^{r_u(2H_u-2)} ds_1 ds_2 \\
&= O\big((nb_g)^{a_u}\big)
\end{aligned} \tag{4.87}$$

and if $a_u < a_v$,

$$\begin{aligned} A_n(t) &= \beta^2(t)(nb_g)^{r_v(2H_v-2)} \frac{d^2_{r_v}(t)}{r_v!} C_W^{r_v} \\ &\quad \times \int_{-1}^{1}\int_{-1}^{1} K(s_1)K(s_2)|s_1 - s_2|^{r_v(2H_v-2)} ds_1 ds_2 \\ &= O\big((nb_g)^{a_v}\big). \end{aligned} \tag{4.88}$$

If $a_u = a_v = a_e$,

$$A_n(t) = O\big((nb_g)^{a_e}\big). \tag{4.89}$$

Similarly, the expression for the leading term in covariance between $\widehat{g}(t)$ and $\widehat{g}(s)$ can be derived. Note that

$$\begin{aligned} \mathbb{C}ov(\widehat{g}(t), \widehat{g}(s)) &\sim \frac{1}{n^2 b_g^2} \sum_{l=r_u}^{\infty} \left[C_Z^l \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ c_l(t_i)c_l(t_j) \right.\right. \\ &\quad \left.\left. \times K\left(\frac{t_i - t}{b_g}\right) K\left(\frac{t_j - s}{b_g}\right) |i-j|^{l(2H_u-2)} \right\} \right] \\ &\quad + \frac{1}{n^2 b_g^2} \sum_{l=r_v}^{\infty} \left[C_W^l \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ \beta(t_i)\beta(t_j) d_l(t_i) d_l(t_j) \right.\right. \\ &\quad \left.\left. \times K\left(\frac{t_i - t}{b_g}\right) K\left(\frac{t_j - s}{b_g}\right) |i-j|^{l(2H_v-2)} \right\} \right] \\ &= C_n(t,s) + o(C_n(t,s)) \end{aligned} \tag{4.90}$$

where, if $a_u > a_v$,

$$\begin{aligned} C_n(t,s) &= (nb_g)^{r_u(2H_u-2)} \frac{c_{r_u}(t)c_{r_u}(s)}{r_u!} C_Z^{r_u} \\ &\quad \times \int_{-1}^{1}\int_{-1}^{1} K(s_1)K(s_2)|s_1 - s_2 + (t-s)/b|^{r_u(2H_u-2)} ds_1 ds_2 \\ &= O\big((nb_g)^{r_u(2H_u-2)}\big) \end{aligned} \tag{4.91}$$

whereas if $a_u < a_v$,

$$C_n(t,s) = \beta(t)\beta(s) \times (nb_g)^{r_v(2H_v-2)} \frac{d_{r_v}(t)d_{r_v}(s)}{sr_v!} C_W^{r_v} \times \int_{-1}^{1}\int_{-1}^{1} K(s_1)K(s_2)|s_1 - s_2 + (t-s)/b|^{r_v(2H_v-2)} ds_1 ds_2 = O\left((nb_g)^{r_v(2H_v-2)}\right) \quad (4.92)$$

We may thus summarize the above facts concerning the bias and the variance (covariance) of the estimated trend curve g as follows. First of all, let $\delta = -1$ in the case of short-memory and $\delta = max(r_u(2H_u - 2), r_v(2H_v - 2))$ in the case of long-memory. Specifically for $\widehat{g}(t)$ we have the following (similar results can also be proved for $\widehat{h}(t)$).

Let the assumptions mentioned above hold and also let $nb_g^3 \to \infty$ as $n \to \infty$; then for every $t, s \in (0,1)$,

$$Bias(\widehat{g}(t)) = \left\{ \frac{b_g^2}{2} \cdot \frac{d^2}{dt^2} g(t) \cdot \int_{-1}^{1} s^2 K(s) ds \right\} + o\left(b_g^2\right), \quad (4.93)$$

$$\mathbb{C}ov(\widehat{g}(t), \widehat{g}(s)) = O\left((nb_g)^{\delta}\right). \quad (4.94)$$

The optimal bandwidth b_g^{opt} can then be derived by differentiating the leading term of the mean squared error and is of the order $O(n^{\delta/(4-\delta)})$. If $\delta = -1$ we arrive at the familiar rate of $n^{-1/5}$ for the short-memory, uncorrelated or independent case (also see Herrmann et al. 1992).

4.2.3.2 Uniform consistency

We will argue that, as $n \to \infty$, $\widehat{\beta}(t)$ converges uniformly to $\beta(t)$ in probability. For this, we will make use of the characteristic function of the kernel (see Parzen 1962). Eventually, we also make use of the fact that c_l, d_l, and β are bounded functions and

$$\sum_{l=r}^{\infty} \frac{1}{l!} < e, r \geq 1. \quad (4.95)$$

Let $\psi(s), -\infty < s < \infty$, be the characteristic function of K. Then, due to the inversion theorem,

$$K(w) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isw} \psi(s) ds. \quad (4.96)$$

Now define the kernel smoothed (true) errors

$$S_n(t) = \frac{1}{nb_g}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b_g}\right)\epsilon_i, \tag{4.97}$$

$$Q_n(t) = \frac{1}{nb_h}\sum_{i=1}^{n} K\left(\frac{t_i - t}{b_g}\right)v_i. \tag{4.98}$$

Then we can rewrite

$$S_n(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j e^{-iwt_j}\right)e^{iwt}\psi(wb_g)dw, \tag{4.99}$$

implying

$$\begin{aligned}\mathbb{E}\left\{\sup_t |S_n(t)|\right\} &\leq \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n}\epsilon_j e^{-iwt_j}\right|\cdot|\psi(wb_g)|dw \\ &\leq \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\left\{\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j cos(t_j w)\right)\right.\right. \\ &\quad \left.\left. + \mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j sin(t_j w)\right)\right\}^{1/2}|\psi(b_g w)|\right]dw\end{aligned} \tag{4.100}$$

However,

$$\begin{aligned}&\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j cos(t_j w)\right) + \mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j sin(t_j w)\right) \\ &= \frac{1}{n^2}\sum_{j,k=1}^{n} cos(w(t_j - t_k))\gamma_u(j,k;t_j,t_k) \\ &\quad + \frac{1}{n^2}\sum_{j,k=1}^{n} cos(w(t_j - t_k))\beta(t_j)\beta(t_k)\gamma_v(j,k;t_j,t_k) \\ &= \frac{1}{n^2}\sum_{l=r_u}^{\infty}\sum_{j,k=1}^{n} cos(w(t_j - t_k))\frac{c_l(t_j)c_l(t_k)}{l!}\gamma_Z^l(|j-k|) \\ &\quad + \frac{1}{n^2}\sum_{l=r_v}^{\infty}\sum_{j,k=1}^{n} cos(w(t_j - t_k))\beta(t_j)\beta(t_k)\frac{d_l(t_j)d_l(t_k)}{l!}\gamma_W^l(|j-k|) \\ &= V_n + o(V_n).\end{aligned} \tag{4.101}$$

In the case of short-memory,

$$\begin{aligned} V_n &\sim \sum_{l=r_u}^{\infty} \frac{1}{l!} \int_0^1 \int_0^1 c_l(s_1)c_l(s_2)cos(w(s_1-s_2))\gamma_Z^l(n|s_1-s_2|)ds_1ds_2 \\ &\leq \sum_{l=r_u}^{\infty} \frac{1}{l!} \int_0^1 \int_0^1 |c_l(s_1)c_l(s_2)||\gamma_Z(n|s_1-s_2|)|^l ds_1ds_2 \\ &+ \sum_{l=r_v}^{\infty} \frac{1}{l!} \int_0^1 \int_0^1 |d_l(s_1)d_l(s_2)\beta(s_1)\beta(s_2)||\gamma_W(n|s_1-s_2|)|^l ds_1ds_2 \\ &= O\left(\frac{1}{n}\right) \end{aligned} \tag{4.102}$$

as $n \to \infty$ and

$$\begin{aligned} \mathbb{E}\left\{ \sup_t |S_n(t)| \right\} &\leq const \cdot \frac{1}{\sqrt{n}} \cdot \int_{-\infty}^{\infty} |\psi(wb_g)|dw \\ &= O\left(\frac{1}{\sqrt{n}b_g}\right) \end{aligned} \tag{4.103}$$

which converges to zero as $n \to \infty$.

Under long-memory,

$$\begin{aligned} V_n &\sim \frac{C_Z^{r_u}}{r_u!} \int_0^1 \int_0^1 [c_{r_u}(s_1)c_{r_u}(s_2)cos(w(s_1-s_2)) \\ &\times (n|s_1-s_2|)^{r_u(2H_u-2)}]ds_1ds_2 + \frac{C_W^{r_v}}{r_v!} \\ &\times \int_0^1 \int_0^1 [d_{r_v}(s_1)d_{r_v}(s_2)\beta(s_1)\beta(s_2)cos(w(s_1-s_2)) \\ &\times (n|s_1-s_2|)^{r_v(2H_v-2)}]ds_1ds_2 \\ &= O\left(n^{r_u(2H_u-2)}\right) + O\left(n^{r_v(2H_v-2)}\right) \end{aligned} \tag{4.104}$$

so that

$$\mathbb{E}\left\{ \sup_t |S_n(t)| \right\} = O\left(n^{r_u(H_u-1)}b_g^{-1}\right) + O\left(n^{r_v(H_v-1)}b_g^{-1}\right) \tag{4.105}$$

which converges to zero as $n \to \infty$.

We may then summarize some preliminary facts, namely that, as $n \to \infty$, $S_n(t)$ and $Q_n(t)$ converge to zero uniformly for all $t \in (0, 1)$ in probability.

In addition, due to the regularity conditions on $g(t)$, $h(t)$, and their derivatives, and the conditions on K, b_g, and b_h,

$$\widehat{g}(t) = g(t) + r_{1,n}(t) \tag{4.106}$$

and

$$\widehat{h}(t) = h(t) + r_{2,n}(t) \tag{4.107}$$

as $n \to \infty$, where $r_{1,n}(t)$ and $r_{2,n}(t)$ converge to zero uniformly in probability.

The final result of interest is the consistency of the estimated slope function $\beta(t)$, namely that $\widehat{\beta}(t)$ converges uniformly in probability to $\beta(t)$.

To see this consider first the random quantity that mimics the formula for the slope estimator in equation (4.32), but defined via the (true) regression errors. Thus let

$$\widehat{\theta}(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) v_i \epsilon_i / (nb)}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) v_i^2 / (nb)}. \tag{4.108}$$

However, $\epsilon_i = u_i + \beta(t_i) \cdot v_i$. This means that

$$\widehat{\theta}(t) = \frac{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) v_i u_i / (nb)}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) v_i^2 / (nb)} + \frac{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) \beta(t_i) v_i^2 / (nb)}{\sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) v_i^2 / (nb)}. \tag{4.109}$$

Consider now the ratio $P_n(t)/Q_n(t)$ where

$$P_n(t) = \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) \beta(t_i) v_i^2 / (nb) = \beta(t) \cdot \mu_{2v}(t) + o_p(1), \tag{4.110}$$

$$Q_n(t) = \sum_{i=1}^{n} K\left(\frac{t_i - t}{b}\right) v_i^2 / (nb) = \mu_{2v}(t) + o_p(1). \tag{4.111}$$

The same argument as above can be used to establish the fact that under suitable regularity conditions

$$\sum_{i=1}^{n} K((t_i - t)/b)v_i^2/(nb) \tag{4.112}$$

is a uniformly consistent estimator of $\mathbb{E}(v_i^2) = \mu_{2v}(t_i)$, where as $n \to \infty$, $b \to 0$, and $nb \to \infty$. Similarly,

$$\sum_{i=1}^{n} K((t_i - t)/b)\beta(t_i)v_i^2/(nb) \tag{4.113}$$

converges uniformly in probability to $\beta(t)\mu_{2v}(t)$. We then have that

$$P_n(t) = \beta(t) \cdot \mu_{2v}(t) + o_p(1), \tag{4.114}$$

$$Q_n(t) = \mu_{2v}(t) + o_p(1), \tag{4.115}$$

so that

$$\frac{P_n(t)}{Q_n(t)} = \beta(t) + o_p(1), \mu_{2v}(t) > 0. \tag{4.116}$$

Due to the regularity conditions on the Hermite coefficients, uniform consistency of the curve estimates, it follows that the regression residuals are such that

$$\widehat{\epsilon}_i = \epsilon_i + a_{n,i} \tag{4.117}$$

$$\widehat{v}_i = v_i + b_{n,i} \tag{4.118}$$

$$\widehat{v}_i^2 = v_i^2 + c_{n,i} \tag{4.119}$$

where $a_{n,i}$, $b_{n,i}$, and $c_{n,i}$ converge to zero uniformly in probability. This means that

$$\widehat{\theta}(t) = \widehat{\beta}(t) + o_p(1) \tag{4.120}$$

uniformly in t as $n \to \infty$.

5

Surface Estimation

5.1 Introduction

The problem of mean surface estimation is common in many large scale investigations. There is in particular a vast literature on geostatistics (Kriging). Cressie (1993), Cressie and Huang (1999), Diggle and Ribeiro 2007), Gelfand et al. (2010), Isaaks and Srivastava (1989), Ripley (1981), and Opsomer et al. (1999) are some of the references where background information can be found on this topic. In the literature, of typical interest has been situations where the observations (after removing any spatial trend) are either spatially uncorrelated or have stationary covariances. In this chapter, we start with a nonparametric regression model, where the stationarity assumption for the errors need not hold. In particular, there may be substantial heterogeneity in the data with spatial autocorrelations. To introduce the topic, consider some spatial observations on a real-valued random variable of interest. Our primary aim is kernel estimation of the expected value of this random variable. We also consider estimation of non-exceedance probabilities and estimation of the spatial Gini index.

To give some examples of probability estimation for spatial data, consider for instance a forest monitoring data set from Switzerland (Source: Swiss National Forest Inventory) from the regions Jura and the Swiss Plateau. The observations are of the type (x_i, y_i, z_i), $i = 1, 2, \ldots, n$, where x_i and y_i denote respectively the West-East and the South-North coordinates of the centers of n forest plots on a spatial grid and z_i is the sample mean of the

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

DBH (D13 DBH: the diameter of the stem at 1.3 m height) values (cm) from individual trees of a certain species in the ith-plot in the Swiss forests. For further details see Brändli and Speich (2007) and Keller (2011). Consider the problem of estimating the probability that the plot mean DBH (sample mean of DBH values from individual forest plots) will exceed a given threshold. For the illustrations of this chapter, the threshold is taken to be 45 cm. Figures for the tree species Norway Spruce are as follows: Figure 5.1 shows the cloud plot and the histogram of the raw plot means of DBH, Figure 5.2 shows a spatial map of the plot centers where plot locations with plot mean DBH larger than 45 cm are highlighted in green, and Figure 5.3 shows a spatial map (level plot) of kernel estimates of the probability that a plot mean will exceed the threshold of 45 cm. The spatial patterns are somewhat different for the tree species Beech, which are in Figures 5.4, 5.5, and 5.6 respectively. The same cut-off value of 45 cm was used for both tree species. The bandwidths are selected so that for a uniform kernel one may expect approximately 30 observations in a window. A truncated Gaussian kernel is used for kernel smoothing using a product kernel of the type $K_2 = K_1 \times K_1$ where each K_1 is a truncated Gaussian kernel.

Specifically, let $y(\mathbf{s})$ be a continuous index random field, where $\mathbf{s} \in \mathbb{R}^2_+$ denotes a two-dimensional spatial location. As is typically the case, we will have at our disposal observations on y available at a discrete set of locations. In this chapter, we focus on the problem of estimating the mean surface $\mathbb{E}(y(\mathbf{s}))$ when the marginal distribution of the centered observations $u(\mathbf{s}) = y(\mathbf{s}) - \mathbb{E}(y(\mathbf{s}))$ may vary having arbitrary and non-Gaussian shapes over varying $\mathbf{s}$, and there may be a lack of homogeneity in the data. As considered earlier in this book, a simple model that incorporates these properties in the data (heterogeneity, location-dependent distribution) is Gaussian subordination, i.e., the assumption that the observations are one-dimensional transformation of an unobserved Gaussian process. For generalizations to higher dimensional transformation, see, for example, Bardet and Surgailis (2013).

Thus we assume that the centered observations $u(\mathbf{s})$ are subordinated to a zero mean, unit variance, stationary latent Gaussian random field $Z(\mathbf{s})$ (Taqqu 1975) via a location-dependent

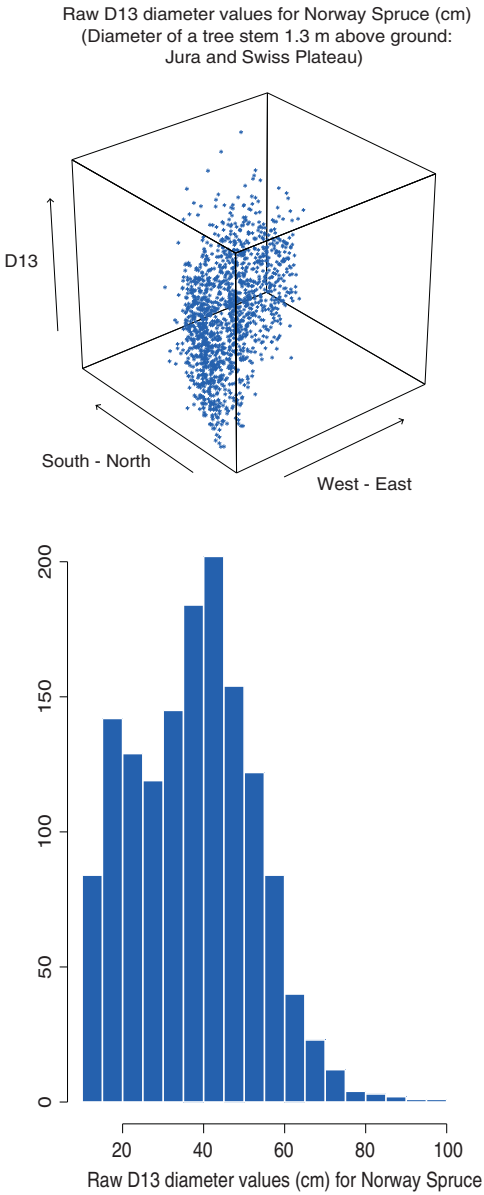

Figure 5.1 Raw D13 diameter values (cm) for Norway Spruce in the Jura and the Swiss Plateau regions: cloud plot and histogram (S-plus) of plot means. *Source:* Data from Swiss National Forest Inventory.

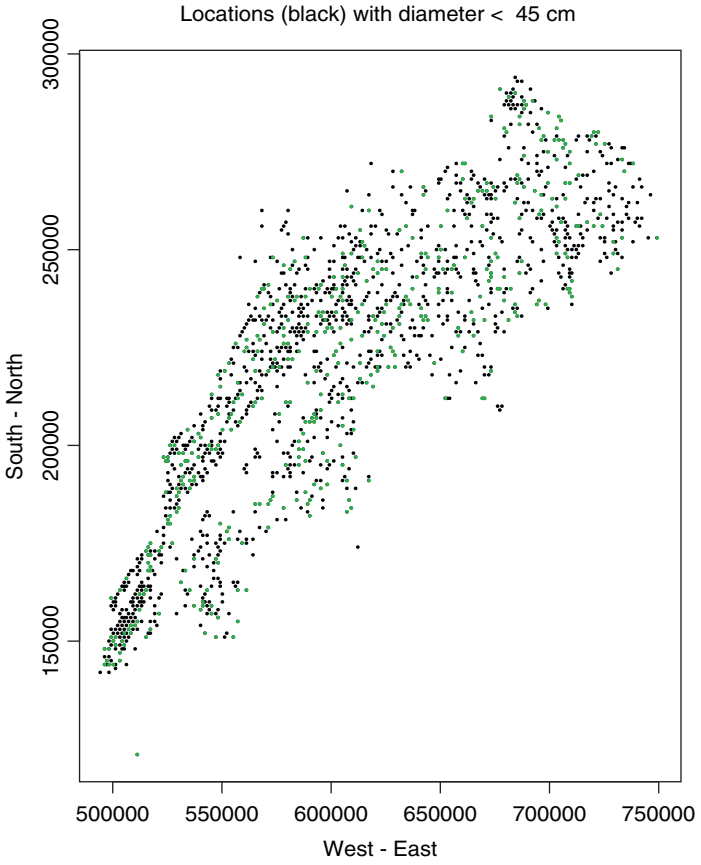

Figure 5.2 Raw D13 diameter values (cm) for Norway Spruce in the Jura and the Swiss Plateau regions: spatial coordinates of the forest plots with mean DBH less than 45 cm are colored black. *Source:* Data from Swiss National Forest Inventory.

transformation. This idea is further explained below. Our main interest lies in nonparametric regression estimation with spatial observations having these properties.

As an illustration consider an excerpt from a global total column ozone data set (Source: NASA), between latitudes 35 and 55 degrees north and longitude values between zero and 20 degrees east. The ozone values are in Dobson Units (DU). The

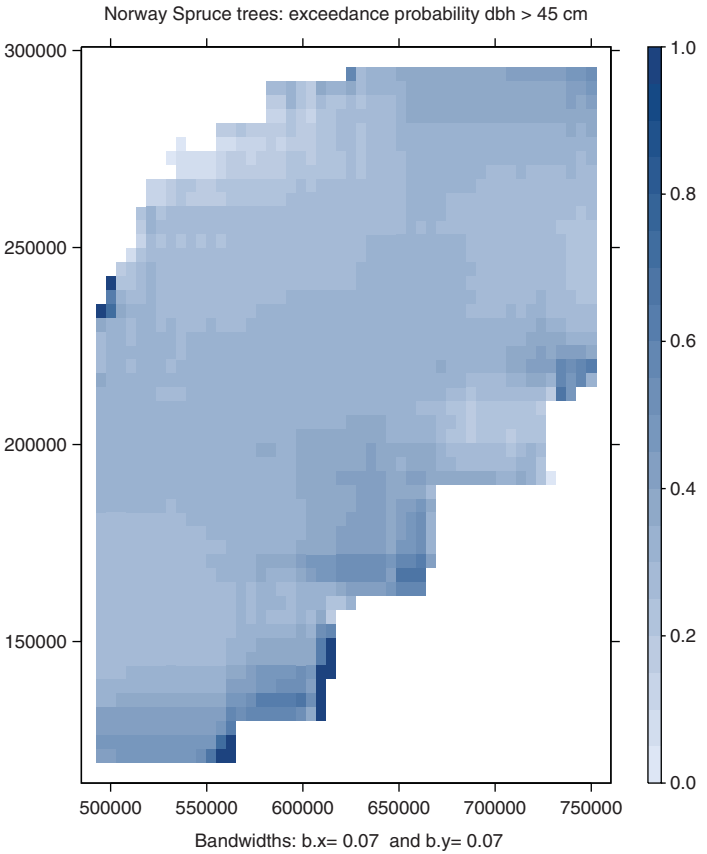

Figure 5.3 Exceedance probability estimates for Norway Spruce in the Jura and the Swiss Plateau regions: level plot of the estimated probabilities $P(DBH > 45\,\text{cm})$, where DBH is a plot mean of single tree diameter values (cm). *Source:* Data from Swiss National Forest Inventory.

raw data and the estimated probability surface maps are shown. Figures 5.7, 5.8, and 5.9 show (a) the raw ozone values, (b) the histogram, and (c) the Priestley–Chao estimate of $P(y(\mathbf{s}) < v)$. Here $v = 332.5$ DU and the sample mean of the ozone values in the selected area is used for illustration.

Whether the problem is estimation of the mean or the marginal distribution function, we first formulate an appropriate

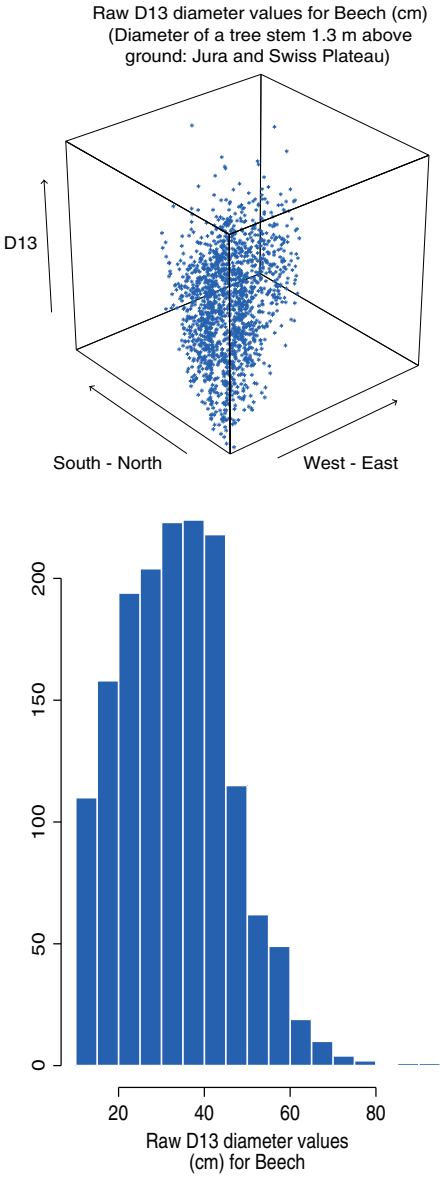

Figure 5.4 Raw D13 diameter values (cm) for Beech in the Jura and the Swiss Plateau regions: cloud plot and histogram (S-plus) of plot means. *Source:* Data from Swiss National Forest Inventory.

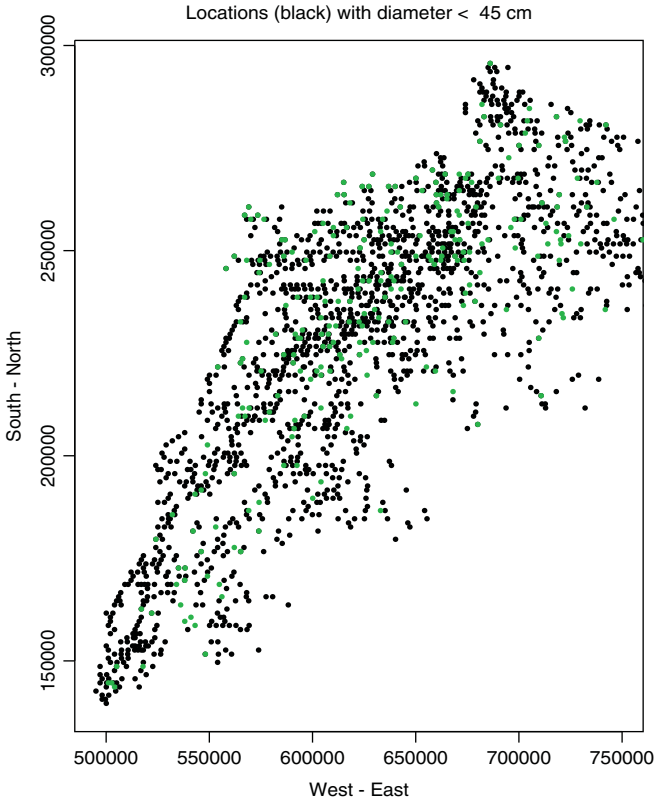

Figure 5.5 Raw D13 diameter values (cm) for Beech in the Jura and the Swiss Plateau regions: spatial coordinates of the forest plots with mean DBH less than 45 cm are colored black. *Source:* Data from Swiss National Forest Inventory.

nonparametric regression model where the deterministic component is to be estimated. We consider Priestley–Chao kernel estimators, but other estimators can also be considered. The advantage of a nonparametric approach is that it allows us to stay fairly flexible as far as the shape of the surface to be estimated is concerned. The kernel estimator is simply a weighted average of the observations, where the weight depends on a bivariate kernel and a bandwidth vector with two elements. We use a product kernel and address a strategy for optimal bandwidth selection.

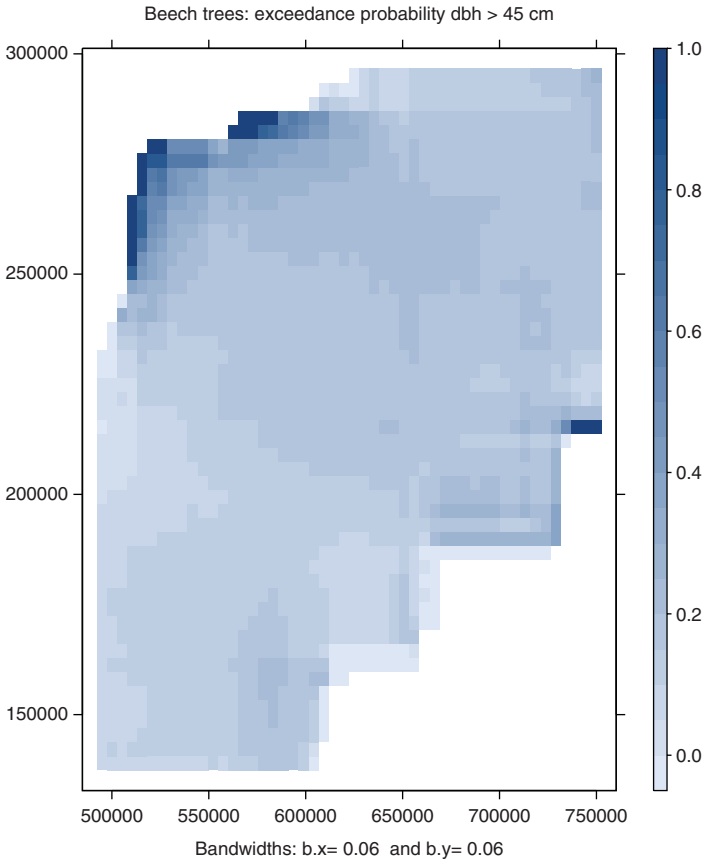

Figure 5.6 Exceedance probability estimates for Beech in the Jura and the Swiss Plateau regions: level plot of the estimated exceedance probability $P(DBH > 45\,\text{cm})$ where DBH is a plot mean of single tree diameter values (cm). *Source:* Data from Swiss National Forest Inventory.

Needless to say, to keep our discussions simple, we consider spatial observations $y(\mathbf{s}) : \mathbb{R}^2_+ \to \mathbb{R}$, where $\mathbf{s}$ denotes a geographical coordinate in $\mathbb{R}^2_+$. However, the methods discussed here generalize easily to situations where $\mathbf{s}$ resides in $\mathbb{R}^d_+$, for instance where $d \geq 2$.

Pointwise consistency of the surface estimator can be established by noting that both bias and variance of the estimator

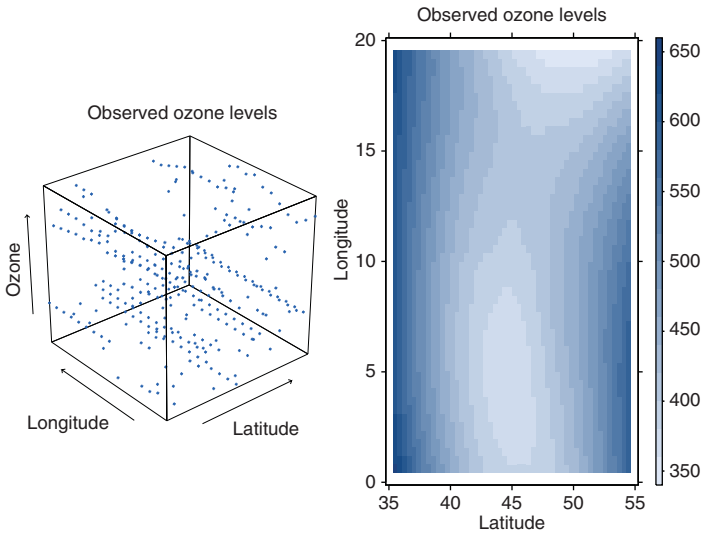

Figure 5.7 Total column ozone maps (ozone values in Dobson units): cloud plot and level plot (S-plus) of raw ozone levels from a section of the ozone field, an excerpt from a global total column ozone data set. The level plot uses a loess smoothing of the data with span = 1 and degree = 2. The coordinates (in decimal degrees) are between latitudes 35 and 55 degree N and longitudes 0 and 20 degrees E. *Source:* NASA.

converge to zero with increasing sample size. In addition, uniform consistency can be established. Uniform consistency of nonparametric curve estimates in one dimension is considered in Parzen (1962), Bierens (1983), Hall and Hart (1990), Mack and Silverman (1982), Schuster (1970), and Silverman (1978), among others. In this chapter, we follow the approach due to Parzen, who considers kernels with an absoluely integrable characteristic function. In this chapter, under Gaussian subordination, Parzen's condition on the kernel, along with some additional regularity conditions, are used to give a simple proof of uniform consistency of the nonparametric surface estimator.

For information on long-memory processes see Beran (1994), Beran et al. (2013), and Giraitis et al. (2012). For long-memory random fields, see in particular Lavancier (2006) and Leonenko (1999). Other relevant references are in Dehling and Taqqu (1989), Breuer and Major (1983), and Giraitis and Surgailis

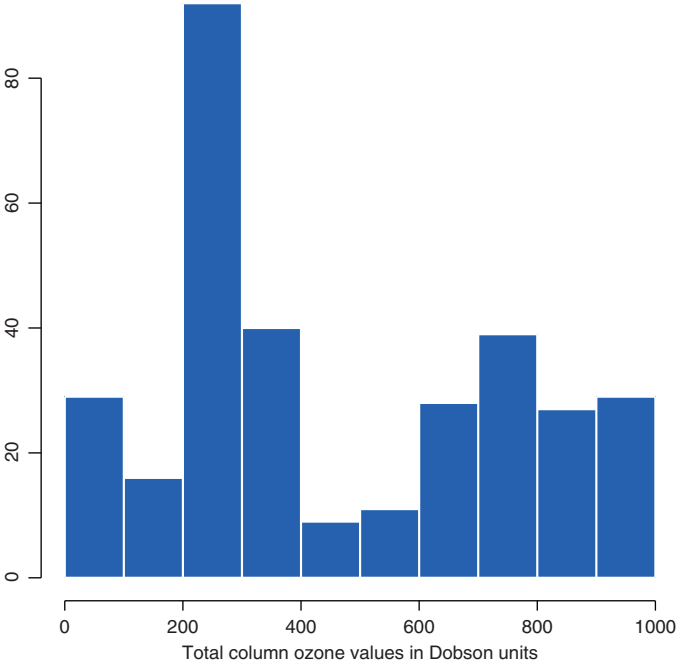

Figure 5.8 Total column ozone (Dobson units): histogram of ozone observations from a section of a global ozone data set (*Source:* NASA). The coordinates (in decimal degrees) are between latitudes 35 and 55 degrees N and longitude values between zero and 20 degrees E.

(1985). Short-memory correlations in $Z(\mathbf{s})$ are discussed in Csörgő and Mielniczuk (1996).

Of special interest are isotropic covariance functions (see Cressie 1993). Our formulation of the covariance function of the long-memory process considered in (5.17) is not strictly isotropic because the function f is present. However, long-memory holds due to the hyperbolic term $|\mathbf{h}|^{-2\alpha}$. Because of this, we use the phrase "isotropically long-range" dependent (see Lavancier 2006). A well-known example of an isotropic long-memory random field is the ising model on a square lattice at a critical temperature (see Cassandro and Jona-Lasinio 1978, Fisher 1964, and Kaufman and Onsager 1949, as well as Cressie 1993, p. 68, for further information).

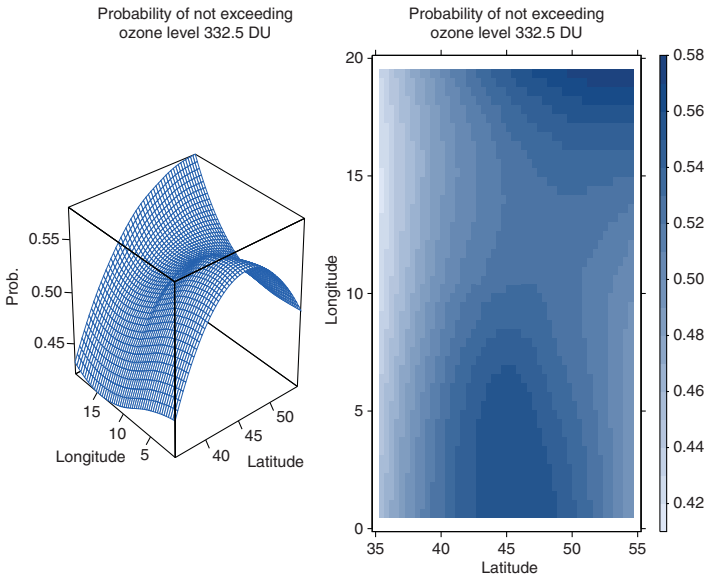

Figure 5.9 Probability surface maps for total column ozone: wireframe plot and level plot (S-plus) of the probability of not exceeding 332.5 DU, between latitudes 35 and 55 degrees N and longitudes 0 and 20 degrees E. The probability surface was estimated using equal bandwidths for both axes and $b_1 = b_2 = 0.15$. The wireframe plot and the level plot use a further loess smoothing of the probability estimates with span = 1 and degree = 2. *Source:* NASA.

For spatial and spatio temporal processes see, among others, Cressie (1993) and Cressie and Huang (1999). See Beran et al. (2006) and references therein for relevant information on estimation for a separable long-memory random field on a lattice. For spatio-temporal separable processes see Fuentes (2006). These separable processes are not considered here.

For kernel smoothing of long-memory time series data, see Beran and Feng (2002), Csörgő and Mielniczuk (1995), Ghosh (2001), Ghosh and Draghicescu (2002), Ghosh et al. (1997), Giraitis et al. (2012, Chapter 12), Guo and Koul (2007), Menéndez et al. (2010, 2013), Ray and Tsay (1997), Robinson (1997), and Robinson and Hidalgo (1997), as well as Chapter 3 of this book on Trend Estimation. For a general background on kernel smoothing see Silverman (1986) and Wand and Jones

(1995), as well as earlier chapters in this book. These methods can then be adapted for the spatial case, some of which are presented here.

The observations that we consider here are available on a grid of known locations $\mathbb{A}$, with Gaussian subordinated errors having spatial auto-correlations that have either short-range or long-range dependence. For such data, Ghosh (2009) addresses a problem in spatial ecology, where the problem is to estimate the number of unseen plant species. In this case, the latent Gaussian process $Z(\mathbf{s})$ serves as a model for a (centered) background process that is decisive of species occurrence. It turns out that in this case, under some additional regularity conditions, a convergent species–area relation can be derived, giving foundation to the well-known and much debated hyperbolic shape of the so-called species–area curves in the ecological literature (Chao 2004). Thus suppose that species j occurs at location $\mathbf{s}$ if and only if $Z(\mathbf{s}) \in A_j$ where A_j is an interval on the real line. Based on a surveyed sample on a regular grid on occurrence of plant species, the problem is to establish the species–area relation, which is the mathematical relation between the expected number of species and area. Ghosh (2009) shows that, if $\zeta_{k,j}$ is the observed number of plots where species j occurs when k sites have been sampled, then for δ, a non-negative value less than or equal to 1, $Var\left(\zeta_{k,j}\right) \propto k^{\delta}$. In particular, if the unknown total number of species is much larger than the number of plots surveyed, then the proportion of unseen species can be approximated by a constant multiple of $k^{\beta-1}$, where $\beta \in [0.5, 1)$, whereas the species proportion increments are asymptotically proportional to $k^{\beta-2}$. In other words, a hyperbolic decay or long-memory in the spatial autocorrelations in the background process $Z(\mathbf{s})$ leads to a fast convergence rate for the number of species with increasing area. This result thus has implications for assessing how many species are present in an area, having applications in nature conservation problems. For theoretical details and some numerical results, the reader is referred to Ghosh (2009). Numerical applications can be found in Ghosh (2009) and Ghosh et al. (1997b), who use a bootstrap based approach for constructing species–area curves.

In the time series context, the Gaussian subordination model has been considered by a large number of authors. For some

recent applications with continuous index processes for analyzing irregularly spaced time series observations see Menéndez et al. (2010) and Menéndez et al. (2012). For additional information on nonlinear transformation of Gaussian processes see Bardet and Surgailis (2013), Breuer and Major (1983), Csörgő and Mielniczuk (1996), Dobrushin and Major (1979), Giraitis and Surgailis (1995), Major (1981), and Taqqu (1975, 1979); also see Beran (1994), Beran et al. (2013), Doukhan et al. (2003), Embrechts and Maejima (2002), Giraitis et al. (2012), and Künsch (1986) among others. Ghosh (2015a, 2015b) considers a Gaussian subordinated spatial process.

5.2 Gaussian subordination

Let the locations where the observations are available be on a square grid,

$$\{(i,j), i,j = 1,2,\dots,n\}.$$

Thus suppose that we have $k = n^2$ observations $y(\mathbf{s_1})$, $y(\mathbf{s_2}), \dots, y(\mathbf{s_k})$ observed at the locations $\mathbf{s_1}, \mathbf{s_2}, \dots, \mathbf{s_k}$. Let the rth location be $\mathbf{s_r} = (s_{1r}, s_{2r}) \in \mathbb{R}^2_+, r = 1,2,\dots,k$, and let

$$\mathbf{t_r} = \mathbf{s_r}/n = (s_{1r}/n, s_{2r}/n) \in [0,1]^2 \tag{5.1}$$

denote the k rescaled locations.

We are interested in estimation of

$$m(\mathbf{t}) = \mathbb{E}(y(\mathbf{s})) \tag{5.2}$$

where the mean surface $m(\mathbf{t})$ for $\mathbf{t} \in [0,1]^2$ is in $\mathbb{C}^3([0,1]^2)$, and $\mathbf{t} = \mathbf{s}/n$ denotes a rescaled location.

We will make the following assumptions concerning the errors or the centered observations. Let

$$u(\mathbf{s}) = y(\mathbf{s}) - \mathbb{E}(y(\mathbf{s})) = y(\mathbf{s}) - m(\mathbf{t}) \tag{5.3}$$

have finite variance and be subordinated to a latent Gaussian random field Z such that

$$u(\mathbf{s}) = G(Z(\mathbf{s}), \mathbf{t}), \mathbf{s} \in \mathbb{R}^2_+, \mathbf{t} \in [0,1]^2, \tag{5.4}$$

for some unknown function $G : \mathbb{R} \times [0,1]^2 \to \mathbb{R}$. We let G be a Lebesgue-measurable L^2 function with respect to the standard

normal density. On the other hand, Z has zero mean and the covariance function

$$\mathbb{C}ov(Z(\mathbf{s_1}), Z(\mathbf{s_2})) = \gamma_Z(|\mathbf{s_1} - \mathbf{s_2}|), \mathbf{s_1}, \mathbf{s_2} \in \mathbb{R}^2_+, \tag{5.5}$$

where $|\cdot|$ denotes the Euclidean norm.

Due to the finite variance assumption, we can write down the Hermite polynomial expansion

$$u(\mathbf{s}) = \sum_{l=q}^{\infty} \frac{c_l(\mathbf{t})}{l!} H_l(Z(\mathbf{s})) \tag{5.6}$$

where $\mathbf{s} \in \mathbb{R}^2_+, \mathbf{t} \in [0,1]^2$. Here $q \geq 1$ is the Hermite rank of G, c_l are Hermite coefficients assumed to be in $\mathbb{C}^3([0,1]^2)$, and $H_l, l \geq 1$, is the Hermite polynomial of degree l.

The Hermite polynomials (e.g. Szegő 1975) satisfy

$$Cov(H_l(Z(\mathbf{s}+\mathbf{h})), H_{l'}(Z(\mathbf{h}))) = 0, \text{ if } l \neq l', \tag{5.7}$$

whereas

$$Cov(H_l(Z(\mathbf{s}+\mathbf{h})), H_l(Z(\mathbf{h}))) = l!\{\gamma_Z(|\mathbf{h}|)\}^l \tag{5.8}$$

and

$$Var(H_l(Z(\mathbf{s})) = l!. \tag{5.9}$$

Using these properties, it is easy to see that the regression error variance is location-dependent as follows:

$$\begin{aligned}\sigma^2(\mathbf{t}) = \mathbb{V}ar(u(\mathbf{s})) &= \mathbb{V}ar\left(\sum_{l=q}^{\infty} \frac{c_l(t)}{l!} H_l(Z(\mathbf{s}))\right) \\ &= \sum_{l_1,l_2=q}^{\infty} \frac{c_{l_1}(\mathbf{t}) c_{l_2}(\mathbf{t})}{l_1! l_2!} \mathbb{C}ov\big(H_{l_1}(Z(\mathbf{s}) H_{l_2}(Z(\mathbf{s}))\big) \\ &= \sum_{l=q}^{\infty} \frac{c_l^2(\mathbf{t})}{l!}\end{aligned} \tag{5.10}$$

where $\mathbf{t} = \mathbf{s}/n$ are rescaled locations. We assume σ to be three times continuously differentiable and uniformly bounded for every $\mathbf{t} \in [0,1]^2$.

If G is the identity function $G(x, \cdot) = x$, then $u(\mathbf{s})$ is Gaussian. However, when this is not the case, the class of marginal probability distributions is broad. This then raises the question of estimation of the marginal probability distribution of u. We consider this in the context of computing the spatial Gini index.

In terms of the spatial grid $(i, j), i, j = 1, 2, \ldots, n$, where observations are available, the rescaled locations being $(i/n, j/n) \in (0, 1]^2$, the nonparametric regression model of interest is

$$y(i, j) = m(i/n, j/n) + u(i, j), \tag{5.11}$$

where the errors u have zero mean and finite variance and due to our assumption of Gaussian subordination, they satisfy

$$u(i, j) = G(Z(i, j), i/n, j/n) = \sum_{l=q}^{\infty} \frac{c_l(i/n, j/n)}{l!} H_l(Z(i, j)), \tag{5.12}$$

where for every fixed (i, j), $Z(i, j)$ has zero mean and unit variance. Also, $q \geq 1$ is the Hermite rank of G, whereas $c_l : [0, 1]^2 \to \mathbb{R}$ are Hermite coefficients, assumed to be in $\mathbb{C}^3([0, 1]^2)$, $l = q, q + 1$, $q + 2, \ldots, H_l(\cdot)$ being the Hermite polynomial of degree l.

5.3 Spatial correlations

The covariance function of $Z(i, j)$ in terms of the discrete locations is

$$Cov(Z(i_1, j_1), Z(i_2, j_2)) = \gamma_Z \left(\sqrt{(i_1 - i_2)^2 + (j_1 - j_2)^2} \right). \tag{5.13}$$

We will consider two types of spatial auto correlations in Z: (a) short-memory and (b) long-memory; see Beran (1994), Lavancier (2006), and Major (1981). In these cases, γ_Z is assumed to satisfy the following properties for any integer $q_0 \geq 1$:

Short-memory

$$\sum_{l=q_0}^{\infty} \int_{[0,1]^2} \int_{[0,1]^2} \left| \gamma_Z(\sqrt{k} |\mathbf{t_1} - \mathbf{t_2}|) \right|^l d\mathbf{t_1} d\mathbf{t_2} = 0\left(\frac{1}{k}\right), k \to \infty \tag{5.14}$$

and in terms of the discrete lags $\mathbf{h} \in \mathbb{N}_+^2$, the spatial autocorrelations are infinitely summable; i.e.,

$$\sum_{\mathbf{h}} |\gamma_Z(\mathbf{h})|^{q_0} < \infty. \tag{5.15}$$

Long-memory

Let H $(0.5 < H < 1)$ be the Hurst coefficient, $0 < \alpha < 1/q_0$ and $H = 1 - \alpha/2$. Then

$$\int_{[0,1]^2} \int_{[0,1]^2} \left|\gamma_Z(\sqrt{k}|\mathbf{t_1} - \mathbf{t_2}|)\right|^{q_0} d\mathbf{t_1} d\mathbf{t_2} = 0\left(k^{q_0(2H-2)}\right), k \to \infty. \tag{5.16}$$

In terms of the discrete lags $\mathbf{h} \in \mathbb{N}_+^2$, the spatial autocorrelations decay hyperbolically with increasing lag:

$$\gamma_Z(\mathbf{h}) \sim C_Z |\mathbf{h}|^{-2\alpha} f\left(\frac{\mathbf{h}}{|\mathbf{h}|}\right), \text{as } |\mathbf{h}| \to \infty. \tag{5.17}$$

Here $C_Z > 0$ and f is a continuous function on $S = \{\mathbf{v} \in \mathbb{R}^2 : |\mathbf{v}| = 1\}$, the unit circle on $\mathbb{R}^2$. C_Z may also be replaced by $L(|\mathbf{h}|)$, where L is a slowly varying function at infinity on $[0, \infty)$ (Dobrushin and Major 1979). In particular in this case, the slow hyperbolic decay causes non-summability of the correlations, i.e.,

$$\sum_{\mathbf{h}} |\gamma_Z(\mathbf{h})|^{q_0} = \infty. \tag{5.18}$$

When $q_0 = q$, where q is the Hermite rank of G, like the long-memory in Z, the errors u will also have long-memory. Therefore, to study this case, we will assume that $q_0 = q$.

Note that for $\mathbf{s} \in \mathbb{R}_+^2$ and $\mathbf{h} \in \mathbb{R}^2$, with $|\mathbf{h}| \to \infty$, and for $\mathbf{v}_1, \mathbf{v}_2, \mathbf{t} \in (0,1)^2$, with $\mathbf{v}_1 \to \mathbf{t}$ and $\mathbf{v}_2 \to \mathbf{t}$,

$$\mathbb{C}ov[u(\mathbf{s}, \mathbf{v}_1), u(\mathbf{s} + \mathbf{h}, \mathbf{v}_2)] \sim C_Z^m \frac{c_q^2(\mathbf{t})}{q!} |\mathbf{h}|^{-2q\alpha} \tag{5.19}$$

where $\sim$ implies that the ratio of the two sides converges to a constant as $|\mathbf{h}| \to \infty$. In this case the data will have long-range dependence if and only if

$$0 < 2q\alpha < 2 \tag{5.20}$$

or, in other words, if and only if $0 < \alpha < 1/q$; cf. Lavancier (2006).

5.4 Estimation of the mean and consistency

The following enumeration of the observations helps to write down the estimator of the mean surface m. As mentioned earlier, there are k locations, and let these be numbered $r = 1, 2, \ldots, k$, where $k = n^2$. Let the rth observation be $y(\mathbf{s}_r)$, where

$$\mathbf{s}_r = (s_{1r}, s_{2r}) \in \mathbb{R}^2_+, \tag{5.21}$$

$$\mathbf{t}_r = (t_{1r}, t_{2r}) = (s_{1r}/n, s_{2r}/n) \in (0, 1]^2. \tag{5.22}$$

Let K be a kernel, which is a symmetric, univariate continuous probability density function on $[-1, 1]$. Let $b_n = b$ be a sequence of bandwidths such that as $n \to \infty$, $b \to 0$, and $nb \to \infty$.

Define the estimator (Priestley and Chao 1972) of the surface m at $\mathbf{t} = (t_1, t_2) \in (0, 1)^2$ by

$$\widehat{m}(\mathbf{t}) = \frac{1}{kb^2} \sum_{r=1}^{k} K\left(\frac{t_{1r} - t_1}{b}\right) K\left(\frac{t_{2r} - t_2}{b}\right) y(\mathbf{s}_r). \tag{5.23}$$

We use a product kernel but this is by no means a restriction. The estimator can also be defined using a general bivariate kernel that is not a product of two univariate kernels.

5.4.1 Asymptotics

Note that as $n \to \infty$,

$$\frac{1}{nb} \sum_{i=1}^{n} \left(\frac{v_i - v}{b}\right)^j K\left(\frac{v_i - v}{b}\right) = \int_{-1}^{1} w^j K(w) dw \left[1 + O\left(\frac{1}{nb}\right)\right] \tag{5.24}$$

where $v_i, v \in (0, 1)$ and $j \geq 0$.

We first examine the bias of the estimator. Taking the expected value,

$$\mathbb{E}(\widehat{m}(\mathbf{t})) = \frac{1}{kb^2} \sum_{r=1}^{k} K\left(\frac{t_{1r} - t_1}{b}\right) K\left(\frac{t_{2r} - t_2}{b}\right) m(\mathbf{t}_r), \mathbf{t}_r = \mathbf{s}_r/n. \tag{5.25}$$

A Taylor series expansion of $m(\mathbf{t}_r)$ around $\mathbf{t}$ reveals that

$$\widehat{m}(\mathbf{t}) = m(\mathbf{t}) + \frac{b^2}{2}\int_{-1}^{1} v^2 K(v)dv \left[\frac{\partial^2}{\partial t_1^2}\{m(\mathbf{t})\} + \frac{\partial^2}{\partial t_2^2}\{m(\mathbf{t})\}\right] + r_n + \frac{1}{kb^2}\sum_{r=1}^{k} K\left(\frac{t_{1r}-t_1}{b}\right) K\left(\frac{t_{2r}-t_2}{b}\right) u(\mathbf{s}_r) \quad (5.26)$$

where $r_n = o(b^2)$ and $nb^3 \to \infty$ as $n \to \infty$.

Since

$$\mathbb{E}(u(\mathbf{s}_r)) = 0, \quad (5.27)$$

the expression for the bias follows. In particular, the leading term in the asymptotic expression of the bias depends on the second partial derivatives of m. As for the variance of the surface estimator,

$$\mathbb{V}ar(\widehat{m}(\mathbf{t})) = \frac{1}{k^2b^4}\sum_{i,j=1}^{k}\left[K\left(\frac{t_{1i}-t_1}{b}\right)K\left(\frac{t_{2i}-t_2}{b}\right) K\left(\frac{t_{1j}-t_1}{b}\right)K\left(\frac{t_{2j}-t_2}{b}\right) \times \mathbb{C}ov(u(\mathbf{s}_i), u(\mathbf{s}_j))\right] \quad (5.28)$$

In case of long-memory, an explicit formula for the leading term of the asymptotic variance can be obtained. We have

$$\mathbb{V}ar(\widehat{m}(\mathbf{t})) = \frac{1}{k^2b^4}\sum_{i,j=1}^{k}\left[K\left(\frac{t_{1i}-t_1}{b}\right)K\left(\frac{t_{2i}-t_2}{b}\right) K\left(\frac{t_{1j}-t_1}{b}\right)K\left(\frac{t_{2j}-t_2}{b}\right) \times \sum_{l=q}^{\infty}\frac{c_l(\mathbf{t}_i)c_l(\mathbf{t}_j)}{l!}\gamma_Z^l(|\mathbf{s}_i-\mathbf{s}_j|)\right] \quad (5.29)$$

$$\sim a(q,\alpha) \times (nb)^{-2q\alpha} c_q^2(\mathbf{t}) C_Z^q \quad (5.30)$$

as $n \to \infty$, where $a(q,\alpha)$ is given by

$$a(q,\alpha) = \int_{-1}^{1}\cdots\int_{-1}^{1}\left\{(v_1-v_2)^2 + (v_3-v_4)^2\right\}^{-q\alpha}\prod_{i=1}^{4}K(v_i)dv_i. \quad (5.31)$$

Here $\sim$ implies that the ratio of the two sides converges to one as $n \to \infty$.

In the case of short-memory, on the other hand, since the auto covariances of the latent Gaussian process are summable and the Hermite coefficients satisfy suitable regularity conditions, the variance of the process $u(\mathbf{s})$ is uniformly bounded.

We may then summarize the above findings as follows. As $n \to \infty$, for fixed $\mathbf{t} \in (0,1)^2$, the bias of $\widehat{m}(\mathbf{t})$, $\mathbf{t} \in (0,1)^2$, is given by

$$\mathbb{E}[\widehat{m}(\mathbf{t})] - m(\mathbf{t}) = \frac{b^2}{2} \int_{-1}^{1} v^2 K(v) dv \times \left[\frac{\partial^2}{\partial t_1^2}\{m(\mathbf{t})\} + \frac{\partial^2}{\partial t_2^2}\{m(\mathbf{t})\} \right] + o(b^2).$$

Moreover, if the latent Gaussian process has short-memory, then

$$\mathbb{V}ar[\widehat{m}(\mathbf{t})] = O((nb)^{-2}). \tag{5.32}$$

On the other hand, under long-memory,

$$\mathbb{V}ar[\widehat{m}(\mathbf{t})] = O((nb)^{-2q\alpha}) \tag{5.33}$$

and $0 < \alpha < 1/q$ where q is the Hermite rank of G.

The above discussion leads to the pointwise consistency of the kernel surface estimator. This follows directly from Chebyshev's inequality.

As for estimation of the mean squared error, using an appropriate higher order kernel (Gasser and Müller 1984), the second partial derivatives of m may be estimated, leading to an estimate of the bias term. For instance, denoting the vth derivative of the kernel K by $K^{(v)}$, to estimate $(\partial^v / \partial u^v m(u,v))$ for $(u,v) \in (0,1)^2$, one may use the kernel estimator (see Gasser and Müller 1984)

$$\widehat{\frac{\partial^v}{\partial u^v} m(u,v)} = \frac{(-1)^{v+1}}{k b_1^{v+1} b_2} \sum_{i=1}^{k} K^{(v)}\left(\frac{t_{1i} - u}{b_1}\right) K\left(\frac{t_{2i} - v}{b_2}\right) y(t_{1i}, t_{2i}),$$

where $b_1, b_2 \to 0$ and in case of short-memory $nb_1^{2v+1}, nb_2 \to \infty$ as $n \to \infty$, with a modified condition for long-memory. This can be plugged into the bias part of the mean squared error. On the other hand, the variance of the estimator involves many unknowns, including the Hermite rank, the Hermite coefficients,

and the entire set of correlations in the data; see, however, a variogram based idea which we discuss below. See Ghosh and Draghicescu (2002) who address this problem for time series data.

As a first step towards estimating the variance of the surface estimator, uniform consistency in probability is a useful result.

We need to establish that

$$S_n(\mathbf{t}) = \frac{1}{kb^2} \sum_{r=1}^{k} K\left(\frac{t_{1r} - t_1}{b}\right) K\left(\frac{t_{2r} - t_2}{b}\right) u(\mathbf{s}_r), \quad (5.34)$$

where $\mathbf{t} = (t_1, t_2) \in (0,1)^2$ and $k = n^2$ converges to zero uniformly for all $\mathbf{t} \in (0,1)^2$ in probability. For this, it is enough to show that

$$\tilde{S}_n = \mathbb{E}\left\{ \underset{t_1,t_2}{\text{Sup}} |S_n(t_1, t_2)| \right\} \quad (5.35)$$

converges to zero as $n \to \infty$. In the case of short-memory, we let as $n \to \infty$, $\sqrt{n}b \to \infty$. In the case of long-memory, let $n^{q(1-H)}b \to \infty$ as $n \to \infty$ where $q \geq 1$ and $0.5 < H < 1$.

We follow an approach due to Parzen (1962), who used a characteristic function based approach; also see Bierens (1983).

Thus assume that the kernel K has a characteristic function $\psi(s), s \in \mathbb{R}$, that is absolutely integrable on the whole real line. Thus,

$$\psi(s) = \int_{-\infty}^{\infty} exp(\imath s x) K(x) dx \quad (5.36)$$

where $\imath = \sqrt{-1}$. Equivalently, the inversion formula states that

$$K(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} exp(-\imath s x) \psi(s) ds. \quad (5.37)$$

This means

$$K\left(\frac{t_{1r} - t_1}{b}\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} exp\left(-\imath s \frac{t_{1r} - t_1}{b}\right) \psi(s) ds, \quad (5.38)$$

$$K\left(\frac{t_{2r} - t_2}{b}\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} exp\left(-\imath s \frac{t_{2r} - t_2}{b}\right) \psi(s) ds. \quad (5.39)$$

Substituting, and writing $u(\mathbf{s}_j) = u_j$,

$$\begin{aligned}
S_n &= \frac{1}{k}\sum_{j=1}^{k} u_j \int_{-\infty}^{\infty} \frac{1}{2\pi b} e^{-\imath s_1 (t_{1j} - t_1/b)} \psi(s_1) ds_1 \\
&\quad \times \int_{-\infty}^{\infty} \frac{1}{2\pi b} e^{-\imath s_2 (t_{2j} - t_2/b)} \psi(s_2) ds_2 \\
&= \frac{1}{k}\sum_{j=1}^{k} u_j \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\imath w_1 t_{1j}} \psi(bw_1) e^{\imath t_1 w_1} dw_1 \cdot \frac{1}{2\pi} \\
&\quad \times \int_{-\infty}^{\infty} e^{-\imath w_2 t_{2j}} \psi(bw_2) e^{\imath t_2 w_2} dw_2 \\
&= \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\left(\frac{1}{k}\sum_{j=1}^{k} u_j e^{-\imath [w_1 t_{1j} + w_2 t_{2j}]} \right) \right. \\
&\qquad \left. e^{\imath [t_1 w_1 + t_2 w_2]} \psi(bw_1)\psi(bw_2) \right] dw_1 dw_2
\end{aligned}$$

Taking expected value,

$$\begin{aligned}
\mathbb{E}\left\{ \sup_{t_1,t_2} |S_n(t_1,t_2)| \right\} \le \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \left[\mathbb{E}\left| \frac{1}{k}\sum_{j=1}^{k} u_j e^{-\imath [w_1 t_{1j} + w_2 t_{2j}]} \right| \cdot \right. \\
\left. |\psi(bw_1)\psi(bw_2)| \right] dw_1 dw_2
\end{aligned}$$

but

$$\begin{aligned}
&\mathbb{E}\left| \frac{1}{k}\sum_{j=1}^{k} u_j e^{-\imath [w_1 t_{1j} + w_2 t_{2j}]} \right| \\
&\le \left\{ \mathbb{V}ar\left(\frac{1}{k}\sum_{j=1}^{k} u_j cos\left([w_1 t_{1j} + w_2 t_{2j}] \right) \right) \right. \\
&\quad \left. + \mathbb{V}ar\left(\frac{1}{k}\sum_{j=1}^{k} u_j sin\left([w_1 t_{1j} + w_2 t_{2j}] \right) \right) \right\}^{1/2} \qquad (5.40) \\
&= \left\{ \frac{1}{k^2}\sum_{j=1}^{k}\sum_{l=1}^{k} cos[w_1(t_{1j} - t_{1l}) + w_2(t_{2j} - t_{2l})] \cdot \mathbb{C}ov(u_j, u_l) \right\}^{1/2} \\
&= W_n^{1/2}, \text{say.}
\end{aligned}$$

Now

$$u_j = G(Z(s_{1j}, s_{2j}), t_{1j}, t_{2j}) = \sum_{r=q}^{\infty} \frac{c_r(t_{1j}, t_{2j})}{r!} H_r(Z(s_{1j}, s_{2j})).$$

Therefore,

$$\mathbb{C}ov(u_j, u_l) = \sum_{r=q}^{\infty} \left[\frac{c_r(t_{1j}, t_{2j}) c_r(t_{1l}, t_{2l})}{r!} \left\{ \gamma_Z \left(\sqrt{(s_{1j} - s_{1l})^2 + (s_{2j} - s_{2l})^2} \right) \right\}^r \right].$$

In the case of short-memory,

$$\begin{aligned} W_n &\sim \sum_{r=q}^{\infty} \frac{1}{r!} \int_0^1 \cdots \int_0^1 \left[c_r(t_1, t_2) c_r(v_1, v_2) \cdot \right. \\ &\quad \left. cos[w_1(t_1 - t_2) + w_2[v_1 - v_2)] \right. \\ &\quad \times \left\{ \gamma_Z \left(n \sqrt{(t_1 - t_2)^2 + (v_1 - v_2)^2} \right) \right\}^r \Big] dt_1 dt_2 dv_1 dv_2 \\ &\leq \sum_{r=q}^{\infty} \frac{1}{r!} \int_0^1 \cdots \int_0^1 \left[\left| c_r(t_1, t_2) c_r(v_1, v_2) \right| \right. \\ &\quad \left. \times \left| \left\{ \gamma_Z \left(n \sqrt{(t_1 - t_2)^2 + (v_1 - v_2)^2} \right) \right\}^r \right| \right] dt_1 dt_2 dv_1 dv_2 \\ &= O\left(\frac{1}{k} \right) \end{aligned}$$

This implies, as $n \to \infty$,

$$\begin{aligned} \mathbb{E} \left\{ \operatorname*{Sup}_{t_1, t_2} |S_n(t_1, t_2)| \right\} &\leq const \cdot \frac{1}{\sqrt{k}} \int_{-\infty}^{\infty} |\psi(w_1 b) \psi(w_2 b)| dw_1 dw_2 \\ &= const \cdot \frac{1}{\sqrt{k}} b^{-2} \int_{-\infty}^{\infty} |\psi(z_1)| dz_1 \\ &\quad \times \int_{-\infty}^{\infty} |\psi(z_2)| dz_2 \\ &= O\left(\frac{1}{(\sqrt{n} b)^2} \right) \end{aligned}$$

which converges to zero as $n \to \infty$ if also $nb^2 \to \infty$.

In the case of long-memory,

$$W_n \sim \frac{C_Z^q}{q!}\int_0^1 \cdots \int_0^1 \left[\{c_q(t_1,t_2)c_q(v_1,v_2)cos[w_1(t_1-t_2) + w_2(v_1-v_2)]\} \times \left\{n\sqrt{(t_1-t_2)^2+(v_1-v_2)^2}\right\}^{q(2H-2)}\right] dt_1 dt_2 dv_1 dv_2$$
$$= O(k^{q(2H-2)}),$$

i.e.,

$$\mathbb{E}\left\{\sup_{t_1,t_2}|S_n(t_1,t_2)|\right\} = O\left(\frac{n^{q(2H-2)}}{b^2}\right) = O\left(\frac{1}{(n^{q(1-H)}b)^2}\right),$$

which converges to zero as $n \to \infty$ if $n^{q(1-H)}b \to \infty$. Here $q \geq 1$ and $0.5 < H < 1$. Hence the result follows for both short-memory and long-memory cases.

Thus $n \to \infty$, $\widehat{m}(\mathbf{t})$ converges uniformly in probability to $m(\mathbf{t})$, where $\mathbf{t} \in (0,1)^2$.

5.5 Variance estimation

For optimal estimation, a good choice of the bandwidth is relevant. Arguments relating the weak consistency of the estimator with Chebyshev's inequality, and the so-called bias–variance trade-off suggest minimization of the mean squared error of the estimator as a function of the bandwidth, as an approach to derive the formula for the optimal bandwidth. However, the asymptotic expression of the mean squared error contains many unknown quantities so that a data-driven algorithm is required to solve the minimization problem. For instance, in a plug-in approach, the asymptotic leading term of the mean squared error, which is the sum of the squared bias and the variance, can be estimated and minimized as a function of the bandwidth.

The leading term in the asymptotic expression of the bias involves the second partial derivatives of the mean surface. These derivatives may be estimated using higher order kernels; see Gasser and Müller (1984). However, variance estimation is

also difficult due to the presence of many unknown parameters and functions. This problem becomes even harder when very large spatial data sets are involved having substantial heterogeneity. It may, however, be possible to develop a direct variance estimation algorithm that uses smoothed variograms, thus avoiding estimation of these unknown parameters. This requires, in the first place, establishing the uniform convergence in probability of the surface estimator to the true function. This is addressed in this chapter using an idea described in Parzen (1962). Specifically, we use kernels with absolutely integrable characteristic functions for constructing the nonparametric surface estimator. Recall that our nonparametric regression model involves spatial observations that are nonlinear transformations of a latent Gaussian random field. Under appropriate regularity conditions, a local-stationarity type property emerges, which can be exploited to suggest a direct estimator of the variance of the Priestley–Chao kernel estimator. This approach to variance estimation avoids estimation of the various nuisance parameters.

Since the Hermite coefficients c_l are continuously differentiable functions, for $\mathbf{t_i}, \mathbf{t_j} \to \mathbf{t}$,

$$\mathbb{C}ov(u(\mathbf{s_i}), u(\mathbf{s_j})) \sim \sum_{l=q}^{\infty} \frac{1}{l!} c_l^2(\mathbf{t})[\gamma_Z(|\mathbf{s_i} - \mathbf{s_j}|)]^l = g(|\mathbf{s_i} - \mathbf{s_j}|, \mathbf{t}) \tag{5.41}$$

for an appropriately defined covariance function g. In other words, a local-stationarity type property emerges (also see Dahlhaus 1997). As usual, $\sim$ indicates that the ratio of the two sides converges to one as $n \to \infty$.

This means that we can suggest the following approximation to the variance of the surface estimator. Let K be a symmetric continuous probability density function with its support on $[-1, 1]$ and vanishing outside this interval, and let $b_n = b$ be a sequence of bandwidths such that $b \to 0$ and $nb \to \infty$ as $n \to \infty$. Also let g be as in (5.41). Then

$$\mathbb{V}ar(\widehat{m}(\mathbf{t})) = v_n(\mathbf{t}) + o(v_n(\mathbf{t})) \tag{5.42}$$

where

$$v_n(\mathbf{t}) = \frac{1}{k^2 b^4} \sum_{i=1}^{k} \sum_{j=1}^{k} P_i(\mathbf{t}) P_j(\mathbf{t}) g(|\mathbf{s_i} - \mathbf{s_j}|, \mathbf{t}) \tag{5.43}$$

where $\mathbf{t} = (t_1, t_2) \in (0,1)^2$, and P_i and P_j are defined as

$$P_i(\mathbf{t}) = K\left(\frac{t_{1i} - t_1}{b}\right) K\left(\frac{t_{2i} - t_2}{b}\right), \tag{5.44}$$

$$P_j(\mathbf{t}) = K\left(\frac{t_{1j} - t_1}{b}\right) K\left(\frac{t_{2j} - t_2}{b}\right). \tag{5.45}$$

The question is then how to estimate the function g. Once this function has been estimated, we may substitute this estimate in the above formula for the variance. One option is to confine this to a window of vanishing size.

Let $\mathbf{d}_n \in [0,1]^2$ be a vector such that as $n \to \infty$, it converges to the null vector of dimention 2. Then due to continuity, both $m(\mathbf{u} + \mathbf{d_n}) - m(\mathbf{t})$ and $\sigma^2(\mathbf{u} + \mathbf{d_n}) - \sigma^2(\mathbf{t})$ converge to zero as $n \to \infty$. Now

$$\mathbb{E}[u(\mathbf{s_i}) - u(\mathbf{s_j})]^2 = \mathbb{V}ar(u(\mathbf{s_i})) + \mathbb{V}ar(u(\mathbf{s_j})) - 2\mathbb{C}ov(u(\mathbf{s_i}), u(\mathbf{s_j})). \tag{5.46}$$

Consider the regression residuals $\widehat{u_n}(\mathbf{s_i})$, namely,

$$\widehat{u_n}(\mathbf{s_i}) = y(\mathbf{s_i}) - \widehat{m}(\mathbf{t_i}), \tag{5.47}$$

where $\mathbf{t_i}$ is the rescaled location corresponding to the spatial coordinate $\mathbf{s_i}$. Since

$$\begin{aligned}\widehat{u_n}(\mathbf{s_i}) - u(\mathbf{s_i}) &= y(\mathbf{s_i}) - \widehat{m}(\mathbf{t_i}) - u(\mathbf{s_i}) \\ &= m(\mathbf{t_i}) - \widehat{m}(\mathbf{t_i}),\end{aligned}$$

we may conclude that, as $n \to \infty$, $\widehat{u_n}(\mathbf{s}) - u(\mathbf{s}), \mathbf{s} \in [0,1]^2$, converges to zero uniformly in probability.

Now define

$$\widehat{\sigma_n^2}(\mathbf{t}) = \frac{1}{kb^2} \sum_{r=1}^{k} K\left(\frac{t_{1r} - t_1}{b}\right) K\left(\frac{t_{2r} - t_2}{b}\right) \widehat{u_n}(\mathbf{s}_r), \tag{5.48}$$

$$\begin{aligned}\widehat{\delta}_n(|\mathbf{h}|, \mathbf{t}) = \frac{1}{kb^2} \sum_{r=1}^{k} \Bigg[& K\left(\frac{t_{1r} - t_1}{b}\right) K\left(\frac{t_{2r} - t_2}{b}\right) \\ & (\widehat{u_n}(\mathbf{s}_r) - \widehat{u_n}(\mathbf{s}_r + \mathbf{h}))^2 \Bigg].\end{aligned} \tag{5.49}$$

Using similar arguments as before, we may summarize as follows: as $n \to \infty$,

1. $\widehat{\sigma_n^2}(\mathbf{t})$ converges uniformly in probability to $\mathbb{V}ar(u(\mathbf{s})) = \sigma^2(\mathbf{t})$ and
2. $\widehat{\delta_n}(|\mathbf{h}|, \mathbf{t})$ converges uniformly in probability to $\mathbb{E}(u(\mathbf{s}) - u(\mathbf{s} + \mathbf{h}))^2$
   where $\mathbf{t} \in (0,1)^2$ is the rescaled coordinate corresponding to $\mathbf{s} = n\mathbf{t}$.

This means that a consistent estimator of the variance of $\widehat{m}(\mathbf{t})$ can be given as

$$\widehat{v}_n(\mathbf{t}) = \frac{1}{k^2 b^4} \sum_{i=1}^{k} \sum_{j=1}^{k} P_i(\mathbf{t}) P_j(\mathbf{t}) \widehat{g_n}(|\mathbf{s_i} - \mathbf{s_j}|, \mathbf{t}) \tag{5.50}$$

where

$$\widehat{g_n}(|\mathbf{s_i} - \mathbf{s_j}|, \mathbf{t}) = \frac{1}{2} \left(\widehat{\sigma_n^2}(\mathbf{t_i}) + \widehat{\sigma_n^2}(\mathbf{t_j}) - \widehat{\delta_n}(|\mathbf{s_i} - \mathbf{s_j}|, \mathbf{t}) \right). \tag{5.51}$$

The above discussions lead to a strategy for the bandwidth selection method. In a plug-in bandwidth selection approach, one may then minimize the asymptotic (leading term) expression of the mean squared error of $\widehat{m}(\mathbf{t}), \mathbf{t} \in (0,1)^2$, with respect to the bandwidth b, where the bias and the variance terms have been estimated using the above suggestions. Now, in the discussion above, we have let the bandwidths for the two spatial coordinates be equal. This restriction is not required. See Ghosh (2015a) for further details.

5.6 Distribution function and spatial Gini index

An interesting application of the surface estimation problem is computation of the spatial Gini index or Gini coefficient, which we denote by Γ. This index was introduced by C. Gini in 1912; see Gastwirth (1972) for a detailed account of this and other measures of inequality. The Gini index is a popular tool in economics. It is used to compare income inequality in different countries. Consider a non-negative continuous random variable with mean $\mu > 0$ and cumulative probability distribution F.

This Gini index is related to the Lorenz curve (Lorenz, 1905), which is defined via the distribution function as

$$L(p) = \frac{1}{\mu} \int_0^p F^{-1}(x)dx \tag{5.52}$$

where $0 < p < 1$ and $F^{-1}(\alpha)$ is the α-quantile $0 < \alpha < 1$ of F. The Lorenz curve has interesting properties. First of all, it takes values between 0 and 1. Also, it is convex and its first derivative is

$$L'(p) = \frac{d}{dp}L(p) = \frac{1}{\mu}F^{-1}(p) \tag{5.53}$$

so that $L'(p) = 1$ if and only if $p = F(\mu)$. The Gini index Γ is the area between $L(p), 0 < p < 1$, and the 45° line $x = y, x, y \in [0, 1]$. It takes values on the interval zero to one, a high value indicating a more uneven distribution. Note that this is the area under the curve $g(p) = p - L(p), p \in [0, 1]$. Since $L(p)$ is convex, $g(p)$ is concave, which means that there is a point p_m where $g(p)$ reaches its maximum, the point being $p_m = F(\mu)$. This p_m is thus the *maximum discrepancy* between the *line of equality* and the Lorenz curve, and in case of income distribution, p_m is the fraction of the population that receives less than the average income in the population.

A formula that we use in our kernel based calculations for the spatial Gini index is based on the mean absolute difference (Kendall and Stuart 1963):

$$\begin{aligned}\Delta &= \int_0^\infty \int_0^\infty |x - y| dF(x)dF(y) \\ &= 2 \int_0^\infty F(y)(1 - F(y))dy\end{aligned} \tag{5.54}$$

so that the Gini index is

$$\Gamma = \Delta/(2\mu). \tag{5.55}$$

In order to estimate the Gini index as a function of the spatial coordinates, we start with a nonparametric regression model with a non-negative response variable. We make a note that the starting point of considering a (nonparametric) regression model allows one to compare the Gini index based on all sorts of factors that could affect income inequality. We assume that the

centered observations, i.e., the regression errors have finite variance, may be spatial correlated, and, in particular, their marginal probability distribution may be non-Gaussian and vary as a function of location.

As a model for these errors, we assume Gaussian subordination, i.e., let the errors (or the observations) be the result of transforming an unobserved Gaussian process, the transformation itself being unknown. For background information for such subordinated processes, see the previous two sections in this chapter. In this section we address how to estimate the Gini index using such spatial observations.

Thus the response variable of interest is continuous, nonnegative, and is denoted by y. We assume that y has finite variance and $k = n^2, n \in \mathbb{N}_+$ observations are available on y at k spatial locations on a square grid. Let $\mathbf{s} \in \mathbb{R}^2_+$ denote a spatial location. We are interested in computing the Gini index for these observations.

We make the folowing additional assumptions. At the location $\mathbf{s}$, suppose that $y(\mathbf{s})$ is subordinated to a latent zero mean, unit variance Gaussian process $Z(\mathbf{s})$ with an isotropic covariance function γ_Z through an unknown transformation G. Specifically, let the function

$$G : \mathbb{R} \times (0,1)^2 \to \mathbb{R} \tag{5.56}$$

be such that $G(z, \cdot)$, where $z \in \mathbb{R}$, is square integrable with respect to the standard normal density and

$$y(\mathbf{s}) - \mathbb{E}(y(\mathbf{s})) = u(\mathbf{s}) = G(Z(\mathbf{s}), \mathbf{t}) \tag{5.57}$$

where $\mathbf{t} = \mathbf{s}/n$ denotes a rescaled location. One important consequence of the above transformation is that the marginal distribution F_u of u, i.e.,

$$F_u(\mathbf{t}, v) = P\{u(\mathbf{s}) \leq v\}, v \in \mathbb{R}, \tag{5.58}$$

may be location-dependent, and in particular F_u may be non-Gaussian. In other words, this formulation allows us to have a flexible model for a nonstationary marginal distribution function of u. Here we consider Z to be univariate and the above transformation to be one-dimensional, although generalizations are possible.

Let the mean of y be

$$m(\mathbf{t}) = \mathbb{E}\{y(\mathbf{s})\} \tag{5.59}$$

and let the marginal distribution of y be

$$F_y(\mathbf{t}, v) = P\{y(\mathbf{s}) \leq v\}, v \in \mathbb{R}. \tag{5.60}$$

The spatial Gini index Γ can be expressed in terms of F_y and m as follows:

$$\begin{aligned}\Gamma(\mathbf{t}) &= \frac{1}{m(\mathbf{t})} \int_0^\infty F_y(\mathbf{t}, v)(1 - F_y(\mathbf{t}, v))dv \\ &= 1 - \frac{1}{m(\mathbf{t})} \int_0^\infty (1 - F_y(\mathbf{t}, v))^2 dv. \end{aligned} \tag{5.61}$$

This is the distribution function based formula for the Gini index which we extend to incorporate spatial locations. For background information see Kendall and Stuart (1963), Gastwirth (1972), and Lerman and Yitzhaki (1984).

To compute, we start with spatial observations $y(\mathbf{s_i}) > 0$ on the non-negative random variable y, where $\mathbf{s_i} \in \mathbb{N}_+^2, i = 1, 2, \ldots, k$, are locations where the observations are avilable. For simplicity of discussion, let the data be available on a square grid

$$\{1, 2, \ldots, n\}^2 \subset \mathbb{N}_+^2 \tag{5.62}$$

with $k = n^2$ observations, for some integer $n \to \infty$. Let the ith location be $\mathbf{s}_i = (s_{1i}, s_{2i})$, with $s_{1i}, s_{2i} \in \mathbb{N}_+, i = 1, 2, \ldots, k$. Also suppose that $\mathbf{t}_i = (t_{1i}, t_{2i})$, with $t_{1i} = s_{1i}/n$ and $t_{2i} = s_{2i}/n$ be the ith rescaled location, $i = 1, 2, \ldots, k$.

To estimate the spatial Gini index, one uses a plug-in approach (see Ghosh 2015b). Estimates of the functions m and F_y are plugged into (5.61). The resulting estimator is consistent due to consistency properties of these curve estimates $\widehat{m}$ and $\widehat{F}_y$.

The nonparametric estimate of the Gini index at the rescaled location $\mathbf{t}$ is thus given by

$$\begin{aligned}\widehat{\Gamma}(\mathbf{t}) &= \frac{1}{\widehat{m}(\mathbf{t})} \int_0^\infty \widehat{F}_y(\mathbf{t}, v)(1 - \widehat{F}_y(\mathbf{t}, v))dv \\ &= 1 - \frac{1}{\widehat{m}(\mathbf{t})} \int_0^\infty (1 - \widehat{F}_y(\mathbf{t}, v))^2 dv. \end{aligned} \tag{5.63}$$

Here $\widehat{m}(\mathbf{t})$ and $\widehat{F}_y(\mathbf{t}, \cdot)$ are respectively nonparametrically estimated mean and marginal distribution function of y at the rescaled location $\mathbf{t} \in (0,1)^2$, i.e.,

$$\widehat{m}(\mathbf{t}) = \frac{1}{kb_1b_2} \sum_{i=1}^{k} K\left(\frac{t_{1i} - t_1}{b_1}\right) K\left(\frac{t_{2i} - t_2}{b_2}\right) y(\mathbf{s}_i) \quad (5.64)$$

is the estimate of the mean and

$$\widehat{F}_y(\mathbf{t}, v) = \frac{1}{kh_1h_2} \sum_{i=1}^{k} K\left(\frac{t_{1i} - t_1}{h_1}\right) K\left(\frac{t_{2i} - t_2}{h_2}\right) w(\mathbf{s}_i, v) \quad (5.65)$$

is the estimate of the marginal distribution function or the nonexceedance probability F_y. Here, w is an indicator function such that, for a given threshold $v \in \mathbb{R}$,

$$w(\mathbf{s}, v) = 1, \text{if } y(\mathbf{s}) \leq v \quad (5.66)$$

and $\quad w(\mathbf{s}, v) = 0, \text{otherwise},$

so that

$$\mathbb{E}(w(\mathbf{s}, v)) = P(y(\mathbf{s}) \leq v) = F_y(\mathbf{t}, v), \quad (5.67)$$

justifying the use of a kernel regression approach as above to estimate F_y. We assume that $F_y(\mathbf{t}, \cdot)$ has finite and continuous partial derivatives up to order three. The kernel K is a univariate continuous symmetric probability density function on $[-1, 1]$. The sequence of bandwidths $b_i, h_i \to 0$ and $nb_i, nh_i \to \infty$ as $n \to \infty$ where $i = 1, 2$.

For illustration, we consider the same ozone data example as in the previous section. These data are obtained as an excerpt from a global total column ozone data set (Source: NASA), between latitudes 35 and 55 degrees north and longitude values between zero and 20 degrees east. The raw data and histogram of the ozone observations are in Figures 5.7 and 5.8. Variations in ozone levels may be due to a number of factors including changes in the balance of chemical production (Fahey and Hegglin 2011). To understand local variations, spatial Gini index Γ may be computed using the kernel based formula above.

Figure 5.10 shows a level-plot (S-plus) of spatial Gini index computed for this data set. The map indicates a locally more

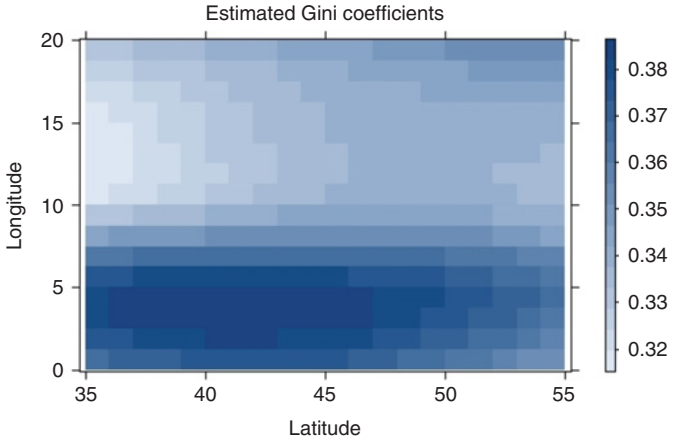

Figure 5.10 Spatial Gini index map: level plot (S-plus) of nonparametrically estimated spatial Gini index for total column ozone values (*Source:* NASA) at coordinates (in decimal degrees) between latitudes 35 and 55 degrees N and longitudes 0 and 20 degrees E.

uneven distribution of the ozone values along the longitudinal gradient expressed by large values (darker color) of the Gini index indicating a substantial change in the Gini index along the longitudinal gradient for the region considered. Since $y(\mathbf{s})$ is nonnegative, the regression surface $m(\mathbf{t})$ can also be estimated from

$$\widehat{m}(\mathbf{t}) = \int_0^\infty \left\{1 - \widehat{F}_y(\mathbf{t}, v)\right\} dv. \tag{5.68}$$

The two formulas (5.64) and (5.68) for estimating m are of course equivalent since

$$\int_0^\infty (1 - w(\mathbf{s}_i, v))dv = y(\mathbf{s}_i), \tag{5.69}$$

which follows by splitting the integral in (5.68) into $\int_{t \le y}$ and $\int_{t>y}$.

It is clear that estimation of the location-dependent probability distribution function F_y by smoothing the $0-1$ values of $w(\mathbf{s}_i, v)$, $i = 1, 2, \ldots, k$, is a special case of the nonparametric surface estimation problem. Moreover, the indicator values w are also Gaussian subordinated, as is y itself. As in the previous section, this fact can be further exploited to prove consistency of the spatial Gini index estimator. For related background information

on estimation of the cumulative distribution function for nonlinear transformation of Gaussian random fields, see Dehling and Taqqu (1989), Breuer and Major (1983), and Giraitis and Surgailis (1985); also see Menéndez et al. (2010, 2012) and Ghosh and Draghicescu (2002) for some statistical applications.

The formula for estimating the spatial Gini index Γ as given in (5.63) is not computationally convenient because it requires integration over the positive side of the real line. However, a closer look reveals that this formula can be written as a double-sum as follows.

First of all, for any integer $p \geq 1$,

$$\int_0^\infty \prod_{j=1}^{p} \{1 - w(\mathbf{s}_j, v)\} dv = \int_0^{J_p} dv = J_p$$

where

$$J_p = min\{y(\mathbf{s}_j), j = 1, 2, \ldots, p\}.$$

This can be seen by noting that

$$\prod_{j=1}^{p} \{1 - w(\mathbf{s}_j, v)\} = 0 \tag{5.70}$$

unless

$$w(\mathbf{s}_j, v) = 0, \forall j = 1, 2, \ldots, p \tag{5.71}$$

or, equivalently, unless

$$min\left\{y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_p)\right\} < v. \tag{5.72}$$

This means that, taking $p = 2$, the integral $\int_0^\infty (1 - \widehat{F}(\mathbf{t}, v)^2 dv$ is equal to ($h_1 = h_2 = h$ is used for simplicity)

$$\begin{aligned}
&\frac{1}{k^2 h^4} \sum_{i_1, i_2 = 1}^{k} \tilde{K}_{i_1}(\mathbf{t}) \tilde{K}_{i_2}(\mathbf{t}) \int_0^\infty (1 - w(\mathbf{s}_{i_1}, v))(1 - w(\mathbf{s}_{i_2}, v)) dv \\
&= \frac{1}{k^2 h^4} \sum_{i_1, i_2 = 1}^{k} \tilde{K}_{i_1}(\mathbf{t}) \tilde{K}_{i_2}(\mathbf{t}) \int_0^{min(y(\mathbf{s}_{i_1}), y(\mathbf{s}_{i_2}))} dv \\
&= \frac{1}{k^2 h^4} \sum_{i_1, i_2 = 1}^{k} \tilde{K}_{i_1}(\mathbf{t}) \tilde{K}_{i_2}(\mathbf{t}) min(y(\mathbf{s}_{i_1}), y(\mathbf{s}_{i_2})),
\end{aligned}$$

where $\tilde{K}_j(\mathbf{t}) = K((s_{1j} - t_1)/h)K((s_{2j} - t_2)/h), j = 1, 2, \ldots, k,$ is a bivariate kernel and K is a symmetric univariate kernel that has been defined earlier. In particular, the kernel estimate of the spatial Gini index $\widehat{\Gamma}(\mathbf{t})$ can be written as

$$\widehat{\Gamma}(\mathbf{t}) = 1 - \frac{\widehat{\eta}(\mathbf{t})}{\widehat{m}(\mathbf{t})} \tag{5.73}$$

where

$$\widehat{\eta}(\mathbf{t}) = \frac{1}{k^2h^4} \sum_{i=1}^{k} \sum_{j=1}^{k} K_i(\mathbf{t})K_j(\mathbf{t}) min \{y(\mathbf{s}_i), y(\mathbf{s}_j)\} \tag{5.74}$$

is the kernel estimate of the marginal mean of the bivariate spatial minima of y at the rescaled location $\mathbf{t} = (t_1, t_2)$ defined earlier.

5.6.1 Asymptotics

As in the previous sections in this chapter, the latent Gaussian process $Z(\mathbf{s})$ may have short-memory or long-memory correlations implying respectively convergence or divergence of the infinite sum of its auto correlations over all distances (lags). The covariance function at the two-dimensional lag $\mathbf{r} = (r_1, r_2)$ is

$$Cov(Z(\mathbf{s}), Z(\mathbf{s} + \mathbf{r})) = \gamma_Z(\mathbf{r}), \tag{5.75}$$

where we let the distance be defined as $|\mathbf{r}| = \sqrt{r_1^2 + r_2^2}$, the Euclidean norm.

Two important correlation types that we consider are (see Beran 1994, Lavancier 2006, and Major 1981 for an overview)

$$\textit{Short-memory}: \sum_{\mathbf{r}} |\gamma_Z(\mathbf{r})|^l < \infty, \tag{5.76}$$

$$\textit{Long-memory}: \gamma_Z(\mathbf{r}) \sim C_Z|\mathbf{r}|^{-2\alpha} f\left(\frac{\mathbf{r}}{|\mathbf{r}|}\right), \text{as } |\mathbf{r}| \to \infty, \tag{5.77}$$

where $0 < \alpha < 1/l$ for some positive integer l. In particular, in the case of long-memory, $\sum_{\mathbf{r}} |\gamma_Z(\mathbf{r})|^l = \infty$. For additional details see the first section of this chapter. As usual, here also $\sim$ indicates that the ratio of the two sides converges to one as $|r| \to \infty$. Also, $C_Z > 0$ and f is a continuous function on $S = \{\mathbf{y} \in \mathbb{R}^2 : |\mathbf{y}| = 1\}$, the unit circle on $\mathbb{R}^2$. Note that C_Z may also be replaced by

a slowly varying function at infinity on $[0, \infty)$ (Dobrushin and Major 1979).

Since the centered observations $u(\mathbf{s})$ are Gaussian subordinated, we may use the argument of the previous section to prove consistency of the surface estimator $\widehat{m}(\mathbf{t})$. As for the probability function estimator $\widehat{F}(\mathbf{t}, v)$, we note that due to the assumption on the regression errors u, the indicator function w is also Gaussian subordinated. Thus,

$$w(\mathbf{s}, v) - F(\mathbf{t}, v) = \tilde{G}(Z(\mathbf{s}), \mathbf{t}, v) \tag{5.78}$$

for some appropriately defined function $\tilde{G}$. Since w is a zero-one function, it is well-defined and it allows for a Hermite polynomial expansion as

$$w(\mathbf{s}, v) - F(\mathbf{t}, v) = \sum_{l=\tilde{q}}^{\infty} \frac{\tilde{c}_l(\mathbf{t}, v)}{l!} H_l[Z(\mathbf{s})], \tag{5.79}$$

where $\tilde{c}_l(\mathbf{t}, v)$ are Hermite coefficients, $\tilde{q}$ is the Hermite rank, assumed to be a constant, and H_l are Hermite polynomials. Since the $H_l(Z)$ where $Z \sim N(0, 1)$ are orthogonal polynomials (hence uncorrelated) and have $\mathbb{V}ar(H_l(Z) = l!$, due to the finite variance of $w(\mathbf{s}, v)$, for $\mathbf{t} \in (0, 1)^2$,

$$\mathbb{V}ar[w(\mathbf{s}, v)] = \sum_{l=\tilde{q}}^{\infty} \frac{\tilde{c}_l^2(\mathbf{t}, v)}{l!} < \infty, \tag{5.80}$$

which is assumed to be continuously differentiable and uniformly bounded for every $\mathbf{t} \in (0, 1)^2$. The Hermite coefficients

$$c_l(\mathbf{t}, v) = \mathbb{E}[G(Z, \mathbf{t}, v) H_l(Z)], l = \tilde{q}, \tilde{q} + 1, \ldots \tag{5.81}$$

are assumed to be continuously differentiable functions so that the above holds where $\tilde{q}$ is the Hermite rank of G.

We also have

$$\mathbb{C}ov(w(\mathbf{s}, v), w(\mathbf{s} + \mathbf{r}, v)) = \gamma_w(\mathbf{r}, v) \sim \frac{\tilde{c}_{\tilde{q}}^2(\mathbf{t}, v)}{\tilde{q}!} C_Z^{\tilde{q}} |\mathbf{r}|^{-2\tilde{q}\alpha} \tag{5.82}$$

so that, for fixed v, the infinite sum of $\gamma_w(\mathbf{r}, v)$ over all $\mathbf{r} \in \mathbb{Z}^2$ diverges if and only if $\alpha < 1/\tilde{q}$.

Following exactly the same line of argument as for proving consistency of $\widehat{m}(\mathbf{t})$, we may summarize the following. In particular, as $n \to \infty$, the leading terms in the asymptotic expressions for the bias and the variance of $\widehat{F}(\mathbf{t}, v)$, where for simplicity we take $h_1 = h_2 = h$, are:

Bias:

$$\mathbb{E}[\widehat{F}(\mathbf{t}, v)] - F(\mathbf{t}, v) \approx \frac{h^2}{2}\mu_2(K)\left[\frac{\partial^2}{\partial t_1^2}\{F(\mathbf{t}, v)\} + \frac{\partial^2}{\partial t_2^2}\{F(\mathbf{t}, v)\}\right].$$

Variance (short-memory):

$$\mathbb{V}ar[\widehat{F}(\mathbf{t}, v)] = O((nh)^{-2}). \tag{5.83}$$

Variance (long-memory):

$$\mathbb{V}ar[\widehat{F}(\mathbf{t}, v)] = O((nh)^{-2\tilde{q}\alpha}). \tag{5.84}$$

In fact, one may extend previous arguments and prove uniform consistency of these estimators $\widehat{m}$ and $\widehat{F}$ under the assumption that the Hermite coefficients of w and u are uniformly continuous and differentiable functions of their arguments $\mathbf{t} \in (0, 1)^2$ and $v \in \mathbb{R}$. Following the same line of argument (*reduction principle*) as in Dehling and Taqqu (1989), uniform consistency of $\widehat{F}(\mathbf{t}, v)$ in v can also be established. These results combined in particular imply consistency of

$$\widehat{\eta}(t) = \int_0^\infty \widehat{F}(\mathbf{t}, v)(1 - \widehat{F}(\mathbf{t}, v))dv. \tag{5.85}$$

Due to Slustky's lemma, the above result in conjunction with the weak consistency of $\widehat{m}$ implies consistency of the kernel estimate $\widehat{\Gamma}(\mathbf{t})$. In other words, as $n \to \infty$, the kernel estimator $\widehat{\Gamma}(\mathbf{t})$ of the spatial Gini index converges in probability to the true Gini index $\Gamma(\mathbf{t})$ at $\mathbf{t} \in (0, 1)^2$.

References

Akaike, H. (1954) An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, **6**, 127–132.

Altman, N.S. (1990) Smoothing of data with correlated errors. *Journal of the American Statistical Association*, **85**, 749–759.

Aneiros-Pérez, G., González-Manteiga, W., Vieu, P. (2004) Estimation and testing in a partial linear regression model under long-memory dependence. *Bernoulli*, **10**, 49–78.

Azzalini, A., Bowman, A.W. (1990) A look at some data on the Old Faithful Geyser. *Applied Statistics*, **39**, 357–365.

Bardet, J.-M., Surgailis, D. (2013) Moment bounds and central limit theorems for Gaussian subordinated arrays. *Journal of Multivariate Analysis*, **114**, 457–473.

Bartlett, M.S. (1963) Statistical estimation of density functions. *Sankhya, The Indian Journal of Statistics, Ser. A*, **25**, 245–254.

Bartlett, M.S., Medhi, J. (1955) On the efficiency of procedures for smoothing periodograms from time series with continuous spectra. *Biometrika*, **42**, 143–150.

Benedetti, J.K. (1977) On the nonparametric estimation of regression functions. *Journal of the Royal Statistical Society, Ser. B*, **39**, 248–253.

Beran, J. (1991) *M*-estimators of location for Gaussian and related processes with slowly decaying serial correlations. *Journal of the American Statistical Association*, **86**, 704–707.

Beran, J. (1992) Statistical methods for data with long-range dependence. *Statistical Science*, **7**, 404–427.

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

Beran, J. (1994) *Statistics for Long-Memory Processes*. Chapman & Hall, New York.

Beran, J. (2009) On parametric estimation for locally stationary long-memory processes. *Journal of Statistical Planning and Inference*, **139**, 900–915.

Beran, J., Feng, Y. (2001a) Local polynomial estimation with a FARIMA-GARCH error process. *Bernoulli*, **7**, 733–750.

Beran, J., Feng, Y. (2001b) Semiparametric fractional autoregressive models. *Statistical Review*, **II**, 125–128.

Beran, J., Feng, Y. (2002a) SEMIFAR models – a semiparametric framework for modelling trends, long-range dependence and nonstationarity. *Computational Statistics and Data Analysis*, **40**, 393–419.

Beran, J., Feng, Y. (2002b) Local polynomial fitting with long-memory, short-memory and antipersistent errors. *The Annals of the Institute of Statistical Mathematics*, **54**, 291–311.

Beran, J., Feng, Y. (2002c) Iterative plug-in algorithms for SEMIFAR models – definition, convergence and asymptotic properties. *Journal of Computational and Graphical Statistics*, **11**, 690–713.

Beran, J., Feng, Y. (2007) Weighted averages and local polynomial estimation for fractional linear ARCH processes. *Journal of Statistical Theory and Practice*, **1**, 149–166.

Beran, J., Feng, Y., Ghosh, S. (2015) Modelling long-range dependence and trends in duration series: an approach based on EFARIMA and ESEMIFAR models. *Statistical Papers*, **56**, 431–451.

Beran, J., Feng, Y., Ghosh, S., Kulik, R. (2013) *Long Memory Processes – Probabilistic Properties and Statistical Models*. Springer-Verlag, Heidelberg.

Beran, J., Feng, Y., Ghosh, S., Sibbertsen, P. (2002) On robust local polynomial estimation with long-memory errors. *International Journal of Forecasting*, **18**, 227–241.

Beran, J., Ghosh, S. (1991) Slowly decaying correlations, Testing normality, nuisance parameters. *Journal of the American Statistical Association*, **86**, 785–791.

Beran, J., Ghosh, S. (1998) Root-n-consistent estimation in partial linear models with long-memory errors. *Scandinavian Journal of Statistics*, **25**, 345–357.

Beran, J., Ghosh, S., Schell, D. (2009) Least square estimation for stationary lattice processes with long-memory. *Journal of Multivariate Analysis*, **100**, 2178–2194.

Beran, J., Ghosh, S, Sibbertsen, P. (2003) Nonparametric M-estimation with long-memory errors. *Journal of Statistical Planning and Inference*, **117**, 199–205.

Beran, J., Ocker, D. (1999) SEMIFAR forecasts, with applications to foreign exchange rates. *Journal of Statistical Planning and Inference*, **80**, 137–153.

Beran, J., Terrin, N. (1994) Estimation of the long-memory parameter based on a multivariate central limit theorem. *Journal of Time Series Analysis*, **15**, 269–278.

Beran, J., Terrin, N. (1996) Testing for a change of the long-memory parameter. *Biometrika*, **83**, 627–638.

Bickel, P., Li, B. (2007) Local polynomial regression on unknown manifolds. *IMS Lecture Notes Monograph Series. Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. Institute of Mathematical Statistics*, **54**, 177–186.

Bickel, P., Ritov, Y. (1988) Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhya, The Indian Journal of Statistics, Ser. A*, **50**, 381–393.

Bickel, P., Rosenblatt, M. (1973) On some global measures of the deviations of density function estimates. *Annals of Statistics*, **1**, 1071–1095.

Bierens, H.J. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of American Statistical Association*, **77**, 699–707.

Bierens, H.J. (1987) Kernel estimators of regression functions. In *Advances in Econometrics: Fifth World Congress*, Vol. 1, T.F. Bewley (Ed.). Cambridge University Press, Cambridge, pp. 99–144.

Billingsley, P. (1968) *Convergence of Probability Measures*. John Wiley & Sons Inc., New York.

Birgé, L., Massart, P. (1995) Estimation of integral functionals of a density. *Annals of Statistics*, **23**, 11–29.

Bochner, S. (1955) *Harmonic Analysis and the Theory of Probability*. University of California Press, Berkeley and Los Angeles.

Boente, G., Fraiman, R. (1989) Robust nonparametric regression estimation for dependent observations. *The Annals of Statistics*, **17**, 1242–1256.

Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.

Bowman, A.W., Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis. The kernel approach with S-Plus illustrations*. Clarenden Press, Oxford.

Bowman, A.W., Hall, P., Titterington, D.M. (1984) Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, **71**, 341–351.

Bradley, R.C. (1983) Asymptotic normality of some kernel-type estimators of probability density. *Statistics and Probability Letters*, **1**, 295–300.

Brändli, U.-B., Speich, S. (2007) *Swiss NFI Glossary and Dictionary*. Available from the World Wide Web http://www.lfi.ch/glossar, Swiss Federal Research Institute WSL, Birmensdorf.

Breidt, F.J., Opsomer, J.D. (2000) Local polynomial regresssion estimators in survey sampling. *Annals of Statisics*, **28**, 1026–1053.

Breuer, P., Major, P. (1983) Central limit theorems for nonlinear functionals of Gaussian fields. *Journal of Multivariate Analysis*, **13**, 425–441.

Brillinger, D.R. (1993) The digital rainbow: some history and applications of numerical spectrum analysis. *The Canadian Journal of Statistics*, **21**, 1–19.

Cacoullos, T. (1966) Estimation of a multivariate density. *Annals of the Institute of Statitsical Mathematics*, **18**, 179–189.

Cao, R., Lugosi, G. (2005) Goodness-of-fit tests based on kernel density estimator. *Scandinavian Journal of Statistics*, **32**, 599–616.

Carroll, R.J., Maca, J.D., Ruppert, D. (1999) Nonparametric regression with errors in covariates. *Biometrika*, **86**, 541–554.

Cassandro, M., Jona-Lasinio, G. (1978) Critical point behaviour and probability theory. *Advances in Physics*, **27**, 913–941.

Čencov, N.N. (1962) Evaluation of an unknown distribution density from observations. *Soviet Math.*, **3**, 1559–1562.

Chao, A. (2004) Species richness estimation. In *Encyclopedia of Statistical Sciences*, 2nd Edition, N. Balakrishnan, C.B. Read,

and B. Vidakovic (Eds.) John Wiley & Sons Inc., New York, 7909–7916.

Chen, S.X. (2000) Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, **52**, 471–480.

Chen, S.X. (2002) Local linear smoothers using asymmetric kernels. *Annals of the Institute of Statistical Mathematics*, **54**, 312–323.

Cheng, K.F., Lin, P.E. (1981a) Nonparametric estimation of a regression function. *Probability Theory and Related Fields*, **57**, 223–233.

Cheng, K.F., Lin, P.E. (1981b) Nonparametric estimation of a regression function: limiting distribution. *Australian Journal of Statistics*, **23**, 186–195.

Cheng, M.Y., Fan, J., Marron, J.S. (1997) On automatic boundary corrections. *Annals of Statistics*, **25**, 1691–1708.

Chiu, S.T. (1989) Bandwidth selection for kernel estimates with correlated noise. *Statistics and Probability Letters*, **8**, 347–354.

Chopin, N. (2007) Dynamic detection of change points in long time series. *Annals of the Institute of Statistical Mathematics*, **59**, 349–366.

Chow, Y.-S., Geman, S., Wu, L.-D. (1983) Consistent cross-validated density estimation. *Annals of Statitsics*, **11**, 25–38.

Clark, R.M. (1977) Non-parametric estimation of a smooth regression function. *Journal of the Royal Statistical Society, Ser. B*, **39**, 107–113.

Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W.S., Devlin, S. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596–610.

Cline, D.B.H. (1988) Admissible kernel estimators of a multivariate density. *Annals of Statistics*, **16**, 1421–1427.

Cline, D.B.H. (1989) Consistency for least squares regression esimators with infinite variance data. *Journal of Statistical Planning and Inference*, **23**, 163–179.

Coeurjolly, J.F. (2000) Simulation and identification of the fractional Brownian motion: a bibliographical and comparative study. *Journal of Statistical Software*, **5**, 1–53.

Collomb, G. (1981) Estimation non-paramétrique de la régression: revue bibliographique. *International Statisitcal Review*, **49**, 75–93.

Collomb, G. (1985a) Non-parametric time series analysis and prediction: uniform almost sure convergence of the window and K-NN autoregression estimates. *Statistics*, **16**, 297–307.

Collomb, G. (1985b) Nonparametric regression: an up-to-date bibliography. *Statistics*, **16**, 309–324.

Cook, D.R. (1977) Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.

Cook, D.R. (1979) Influential observations in linear regression. *Journal of the American Statistical Association*, **74**, 169–174.

Cox, D.R. (1984) Long-range dependence: a review. In *Statistics: An Appraisal, Proceedings of the 50th Anniversary Conference*, H.A. David and H.T. David (Eds.). Iowa State University Press, Ames, pp. 55–74.

Cramér, H. (1972) Studies in the history of probability and statistics. XXVIII. On the history of certain expansions used in mathematical statistics. *Biometrika*, **59**, 205–207.

Craven, P., Wahba, G. (1979) Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Cressie, N.A.C. (1993) Statistics for Spatial Data. John Wiley & Sons Inc., New York.

Cressie, N., Huang, H.-C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340.

Csörgő, M., Horvath, L. (1997) *Limit Theorems in Change-Point Analysis*. John Wiley & Sons Ltd, Chichester.

Csörgő, S. (1981a) Limit behaviour of the empirical characteristic function. *Annals of Probability*, **9**, 130–144.

Csörgő, S. (1981b) multivariate empirical characteristic functions. *Probability Theory and Related Fields*, **55**, 203–229.

Csörgő, S., Mielniczuk, J. (1995) Nonparametric regression under long-range dependent normal errors. *Annals of Statistics*, **23**, 1000–1014.

Csörgő, S., Mielniczuk, J. (1996) The empirical process of a short-range dependent stationary sequence under Gaussian subordination. *Probability Theory and Related Fields*, **104**, 15–25.

Csörgő, S., Mielniczuk, J. (1999) Random-design regression under long-range dependence. *Bernoulli*, **5**, 209–224.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Annals of Statistics*, **25**, 1–37.

Daniell, P.J. (1946) Discussion on Symposium on autocorrelation in time series. *Supplement to the Journal of the Royal Statistical Society*, **8**, 88–90.

Deheuvels, P. (1977) Estimation non parametrique de la densité par histogrammes generalisés. *Revue de Statistique Appliquée*, **25**, 5–42.

Dehling, H., Taqqu, M.S. (1989) The empirical process of some long-range dependent sequences with an application to U-statistics. *Annals of Statistics*, **17**, 1767–1783.

Devroye, L.P. (1978) The uniform convergence of the Nadaraya–Watson regression function estimate. *The Canadian Journal of Statistics*, **6**, 179–191.

Devroye, L.P., Wise, G.L. (1980) Consistency of a recursive nearest neighborhood regression function estimate. *Journal of Multivariate Analysis*, **10**, 539–550.

Devroye, L. (1987) *A Course in Density Estimation*. Birkhäuser Verlag, Boston.

Diggle, P.J. (1990) *Time Series: A Biostatistical Introduction*. Oxford University Press, Oxford.

Diggle, P.J., Ribeiro P.J. (2007) *Model-Based Geostatistics*. Springer, New York.

Diggle, P.J., Wasel, I.A. (1997) Spectral analysis of replicated biomedical time series. *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)*, **46**, 31–71.

Doane, D.P. (1976) Aesthetic frequency classifications. *The American Statistician*, **30**, 181–183.

Dobrushin, R.L., Major, P. (1979) Non-central limit theorems for non-linear functional of Gaussian fields. *Probability Theory and Related Fields*, **50**, 27–52.

Doukhan, P., Oppenheim, G., Taqqu, M.S. (2003) *Theory and Applications of Long-Range Dependence*. Birkhäuser Verlag, Boston.

Draghicescu, D. (2002) Nonparametric Quantile Estimation for Dependent Data. PhD Thesis, EPFL.

Drygas, H. (1976) Weak and strong consistency of least square estimators in regression models. *Probability Theory and Related Fields*, **34**, 119–127.

Duin, R.P.W. (1976) On the choice of the smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans, Comput*. **C-25**, 1175–1179.

Edgeworth, F.Y. (1905) The law of error. *Transactions of the Cambridge Philosophical Society*, **xx**, 36–65, 113–141.

Eicker, F. (1963) Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics*, **34**, 447–456.

Einstein, A. (1914) Méthode pour la détermination de valeurs statistques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. *Archives des Sciences Physiques et Naturelles*, **37**, 254–256.

Embrechts, P., Maejima, M. (2002) *Selfsimilar Processes*. Princeton University Press, Princeton.

Engle, R., Granger, C., Rice, J., Weiss, A. (1986) Nonparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81**, 310–320.

Epanechnikov, V.A. (1969) Nonparametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, **14**, 153–158.

Efromovich, S., Low, M. (1996) On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *Annals of Statitsics*, **24**, 682–686.

Efromovich, S., Samarov, A. (2000) Adaptive estimation of the integral of squared regression derivatives. *Scandinavian Journal of Statistics*, **27**, 335–351.

Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Eubank, R.L. (2000) Spline regression. In *Smoothing and Regression: approaches, computation and application*, M.G. Schimek (Ed.). John Wiley & Sons Inc., New York.

Fahey, D.W., Hegglin, M.I. (2011) Twenty questions and answers about the ozone layer: 2010 update. In: Scientific assessment of ozone depletion: 2010, vol 52, Global Ozone Research and Monitoring Project Report. World Meteorological Organization, Geneva.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., Engel, J. (1997) Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, **49**, 79–99.

Fan, J., Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Farrell, R.H. (1967) On the lack of uniformly consistent sequence of estimators of a density function. *Annals of Mathematical Statistics*, **38**, 471–475.

Farrell, R.H. (1972) On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Annals of Mathematical Statistics*, **43**, 170–180.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd Edition. Wiley, New York.

Feng, Y. (1999) *Kernel – and Locally Weighted Regression – With Application to Time Series Decomposition*. Verlag fr Wissenschaft und Forschung, Berlin.

Feng, Y. (2004) *Non- and Semiparametric Regression with Fractional Time Series Errors – Theory and Applications to Financial Data. Habilitation Work*. University of Konstanz, Konstanz.

Feng, Y. (2007) On the asymptotic variance in nonparametric regression with fractional time series errors. *Journal of Nonparametric Statistics*, **19**, 63–76.

Feng, Y., Beran, J. (2009) Filtered log-periodogram regression of long memory processes. *Journal of Statistical Theory and Practice*, **3**, 777–793.

Feuerverger, A., Ghosh, S. (1988) An asymptotic Neyman–Pearson type result under symmetry constraints. *Communications in Statistics – Theory and Methods*, **17**, 1557–1564.

Feuerverger, A., McDunnough, P. (1981a) On some Fourier methods for inference. *Journal of the American Statistical Association*, **76**, 379–387.

Feuerverger, A., McDunnough, P. (1981b) On the efficiency of empirical characteristic function procedures. *Journal of the Royal Statistical Society, Ser. B*, **43**, 20–27.

Feuerverger, A., Mureika, R.A. (1977) The empirical characteristic function and its applications. *Annals of Statistics*, **5**, 88–97.

Fisher, R.A. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155–160.

Fisher, M.E. (1964) Correlation functions and the critical region of simple fluids. *Journal of Mathematical Physics*, **5**, 944–962.

Fisher, R.A. (1997) On an absolute criterion for fitting frequency curves. *Statistical Science*, **12**, 39–41.

Fix, E., Hodges, J.L. (1951) *Discriminatory analysis and nonparametric estimation: Consistency properties. Project No. 21-49-004, Report No 4*. USAF School of Avian Medicine, Randolf Field, Texas.

Freedman, D., Diaconis, P. (1981) On the histogram as a density estimator: L_2 theory. *Probability Theory and Related Fields*, **57**, 453–476.

Frei, C., Schär, C. (1998) Centennial variations of intense precipitation in Switzerland. In Proceedings of the European Conference on Applied Climatology, October 1998, Vienna, Austria.

Fryer, M.J. (1976) Some errors associated with the nonparametric estimation of density functions. *Journal of the Institute of Mathematics and Its Applications*, **18**, 371–380.

Fuentes, M. (2006) Testing for separability of spatial temporal covariance functions. *Journal of Statistical Planning and Inference*, **136**, 447–466.

Gasser, T., Kneip, A., Köhler, W. (1991) A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association*, **86**, 643–652.

Gasser, T., Müller, H.G. (1984) Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, **11**, 171–185.

Gasser, T., Müller, H.G. (1986) Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

Gasser, T., Müller, H.G., Mammitsch (1985) Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, **47**, 238–252.

Gastwirth, J.L. (1972) The estimation of Lorenz curve and Gini index. *Review of Economics and Statistics*, **54**, 306–316.

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328.

Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (Eds.) (2010) *Handbook of Spatial Statistics*. CRC Press, New York.

Georgiev, A.A. (1984) Kernel estimates of functions and their derivatives with applications. *Statistics and Probability Letters*, **2**, 45–50.

Geweke, J., Porter-Hudak, S. (1983) The estimation and application of long memory time series models. *Journal of Time Series Analysis*, **4**, 221–237.

Ghosh, S. (1996) A new graphical tool to detect non-normality. *Journal of the Royal Statistical Society B*, **58**, 691–702.

Ghosh, S. (2001) Nonparametric trend estimation in replicated time series. *Journal of Statistical Planning and Inference*, **97**, 263–274.

Ghosh, S. (2003) Estimating the moment generating function of a linear process. *Student*, **4**, 211–218.

Ghosh, S. (2006) Regression-based age estimation of a stratigraphic isotope sequence in Switzerland. *Vegetation History and Archaeobotany*, **15**, 273–278.

Ghosh, S. (2009) The unseen species number revisited. *Sankhya, The Indian Journal of Statistics, Ser. B*, **71**, 137–150.

Ghosh, S. (2013) Normality testing for a long-memory sequence using the empirical moment generating function. *Journal of Statistical Planning and Inference*, **143**, 944–954.

Ghosh, S. (2014) On local slope estimation in partial linear models under Gaussian subordination. *Journal of Statistical Planning and Inference*, **155**, 42–53.

Ghosh, S. (2015a) Surface estimation under local stationarity. *Journal of Nonparametric Statistics*, **27**, 229–240.

Ghosh, S. (2015b) Computation of spatial Gini coefficients. *Communications in Statistics – Theory and Methods*, **44**, 4709–4720.

Ghosh, S. (2017) On estimating the marginal distribution of a detrended series with long-memory. *Communications in Statistics – Theory and Methods*, *accepted*.

Ghosh, S., Beran, J. (2000) Comparing two distributions: the two sample T3 plot. *Journal of Computational and Graphical Statistics*, **9**, 167–179.

Ghosh, S., Beran, J. (2006) On estimating the cumulant generating function for linear processes. *Annals of the Institute of Statistical Mathematics*, **58**, 53–71.

Ghosh, S., Beran, J., Innes, J. (1997) Nonparametric conditional quantile estimation in the presence of long memory. *Student*, **2**, 109–117.

Ghosh, S., Dietrich, M., Scheidegger, C. (1997b) Bootstrap based species–area curve estimation for epiphytic lichens in

Switzerland. In *Häufigkeit, Diversityät, Verbreitung und Dynamik von epiphytischen Flechten in Schweizerischen Mittelland und den Voralpen*. Inaugural dissertation der Philosophisch-naturwissenschaftlichen Fakultät der Universität Bern, Dietrich, M. (1997), University of Bern, Bern.

Ghosh, S., Draghicescu, D. (2002a) An algorithm for optimal bandwidth selection for smooth nonparametric quantiles and distribution functions. In *Statistics in Industry and Technology: Statistical Data Anlaysis Based on the L_1-Norm and Related Methods*, Y. Dodge (Ed.). Birkhäser Verlag, Basel, pp. 161–168.

Ghosh, S., Draghicescu, D. (2002b) Predicting the distribution function for long-memory processes. *International Journal of Forecasting*, **18**, 283–290.

Ghosh, S., Ruymgaart, F. (1992) Applications of empirical characteristic functions in some multivariate problems. *The Canadian Journal of Statistics*, **20**, 429–440.

Gijbels, I., Goderniaux, A.C. (2004a) Bootstrap test for change-points in nonparametric regression. *Journal of Nonparametric Statistics*, **16**, 591–611.

Gijbels, I., Goderniaux, A.C. (2004b) Bandwidth selection for change point estimation in nonparametric regression. *Technometrics*, **46**, 76–86.

Gijbels, I., Hall, P., Kneip, A. (1999) On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics*, **51**, 231–251.

Giraitis, L., Koul, H.L. (1997) Estimation of the dependence parameter in linear regression with long-range-dependent errors. *Stochastic Processes and Their Applications*, **71**, 207–224.

Giraitis, L., Koul, H.L., Surgailis, D. (1996) Asymptotic normality of regression estimators with long memory errors. *Statistics and Probability Letters*, **29**, 317–335.

Giraitis, L., Koul, H.L., Surgailis, D. (2012) *Large Sample Inference for Long Memory Processes*. Imperial College Press, London.

Giraitis, L., Leipus, R. (1992) Testing and estimating in the change-point problem of the spectral function. *Lithuanian Mathematical Journal*, **32**, 20–38.

Giraitis, L., Surgailis, D. (1985) CLT and other limit theorems for functionals of Gaussian processes. *Probability Theory and Related Fields*, **70**, 191–212.

Giraitis, L., Taqqu, M.S. (1999) Whittle estimator for finite-variance non-Gaussian time series with long memory. *Annals of Statistics*, **27**, 178–203.

González-Manteiga, W., Aneiros-Pérez, G. (2003) Testing in partial linear regression models with dependent errors. *Journal of Nonparametric Statistics*, **15**, 93–111.

Granger, C.W.J., Joyeux, R. (1980) An introduction to long-range time series models and fractional differencing. *Journal of Time Series Analysis*, **1**, 15–30.

Greblicki, W., Krzyzak, A. (1980) Asymptotic properties of kernel estimates of a regression function. *Journal of Statistical Planning and Inference*, **4**, 81–90.

Grenander, U. (1954) On the estimation og regresion coefficients in the case of an autocorrelated disturbance. *Annals of Mathematical Statitsics*, **25**, 252–272.

Guo, H., Koul, H.L. (2007) Nonparametric regression with heteroscedastic long memory errors. *Journal of Statistical Planning and Inference*, **137**, 379–404.

Habbema, J.D.F., Hermans, J., van den Broek, K. (1974) A stepwise discriminant analysis program using density estimation. In *Compstat 1974*, G. Bruckmann (Ed.). Physica Verlag, Vienna, pp. 101–110.

Hall, P. (1982) Cross-validation in density estimation. *Biometrika*, **69**, 383–390.

Hall, P. (1983) Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics*, **11**, 1156–1174.

Hall, P. (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimates. *Journal of Multivariate Analysis*, **14**, 1–16.

Hall, P. (1987) On the use of compactly supported density estimates in problems of discrimination. *Journal of Multivariate Analysis*, **23**, 131–158.

Hall, P. (1992a) Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics*, **20**, 675–694.

Hall, P. (1992b) *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hall, P., Hart, J.D. (1990) Nonparametric regression with long-range dependence. *Stochastic Processes and Their Applications*, **36**, 339–351.

Hall, P., Marron, J. (1987a) Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation. *Probability Theory and Related Fields*, **74**, 567–581.

Hall, P., Marron, J. (1987b) Estimation of integrated squared density derivatives. *Statistics and Probability Letters*, **6**, 109–115.

Hall, P., Marron, J.S., Park, B.U. (1992) Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1–20.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986) *Robust Statistics – The Approach Based on Influence Functions*. John Wiley & Sons Inc., New York.

Härdle, W. (1989) Asymptotic maximal deviation of *M*-smoothers. *Journal of Multivariate Analysis*, **29**, 163–179.

Härdle, W. (1990) *Smoothing Techniques with Implementations in S*. Springer-Verlag, New York.

Härdle, W., Hall, P., Marron, J.S. (1992) Regression smoothing parameters that are not far from their optimum. *Journal of American Statistical Association*, **87**, 227–233.

Härdle, W., Marron, J.S. (1985) Optimum bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, **13**, 1465–1481.

Härdle, W., Marron, J.S., Wand, M.P. (1990) Bandwidth choice for density derivatives. *Journal of the Royal Statistical Society, Ser. B*, **52**, 223–232.

Härdle, W., Scott, D.W. (1992) Smoothing in low and high dimensions by weighted averaging using rounded points. *Computational Statistics*, **7**, 97–128.

Hart, J.D. (1990) Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, **18**, 873–890.

Hastie, T.J., Loader, C. (1993) Local regression: automatic kernel carpentry (with discussion). *Statistical Science*, **8**, 120–143.

Hastie, T.J., Tibshirani, R.J. (1990) *Generalized Additive Models*. Chapman & Hall, London.

Haslett, J., Raftery, A.E. (1989) Space–time modelling with long-memory dependence: assessing Ireland's wind power resource. *Applied Statistics*, **38**, 1–50.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S.P., Alllen, J.R.M., Huntley, B., Mitchell, F. (2006)

Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society, Ser. A*, **3**, 395–438.
Heidenreich, N.-B., Schindler, A., Sperlich, S. (2013) Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA: Advances in Statistical Analysis*, **97**, 403–433.
Heiler, S., Feng, Y. (1998) A simple root-n bandwidth selector for nonparametric regression. *Journal of Nonparametric Statistics*, **9**, 1–21.
Herrmann, E. (1997) Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics*, **6**, 35–54.
Herrmann, E. (2000) Variance estimation and bandwidth selection for kernel regression. In *Smoothing and Regression, Approaches, Computation and Application*, M.G. Schimek (Ed.). John Wiley & Sons Inc., New York, 71–108.
Herrmann, E., Gasser, T., Kneip, A. (1992) Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, **79**, 783–795.
Hinkley, D.V. (1970) Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17.
Hodges, J.L. Jr., Lehmann, E.L. (1956) The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, **27**, 324–335.
Horvath, L., Kokoszka, P. (1997) The effect of long-range dependence on change point. *Journal of Statistical Planning and Inference*, **64**, 57–81.
Horvath, L., Kokoszka, P. (2002) Change-point detection with non-parametric regression. *Statistics*, **36**, 9–31.
Horvath, L., Shao, Q.-M. (1999) Limit theorems for quadratic forms with applications to Whittle's estimate. *The Annals of Applied Probability*, **9**, 146–187.
Hosking, J.R.M. (1981) Fractional differencing. *Biometrika*, **68**, 165–176.
Huber, P.J. (1967) The behaviour of maximum likelihood estimators under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, L. Le Cam and J. Neymann (Eds.), pp. 221–233.
Huber, P.J. (1981) *Robust Statistics*. John Wiley & Sons Inc., New York.

Huber, P.J., Ronchetti, E.M. (2009) *Robust Statistics*. John Wiley & Sons Inc., New York.

Inclan, C., Tiao, G.C. (1994) Use of cumulative sums of squares for retrospective detection of changes of variances. *Journal of American Statistical Association*, **89**, 913–923.

Isaaks, E.H., Srivastava, R.M. (1989) *An Introduction to Applied Geostatistics*. Oxford University Press, Oxford.

Jensen, M.J., Whitcher, B. (2000) *Time-Varying Long Memory in Volatility: detection and estimation with wavelets. Technical Report*. EURANDOM, Eindhoven.

Johnsen, S.J., Clausen, H.B., Dansgaard, W., Gundestrup, N.S., Hammer, C.U., Andersen, U., Andersen, K.K., Hvidberg, C.S., Dahl-Jensen, D., Steffensen, J.P., Shoji, H., Sveinbjörnsdóttir, A.E., White, J., Jouzel, J., Fisher, D. (1997) The $\delta^{18}o$ record along the Greenland Ice Core Project deep ice core and the problem of possible Eemian climatic instability. *Journal of Geophysical Research*, **102**, 26397–26410.

Johnson, N.L., Kotz, S. (1994) *Continuous Univariate Distributions*, 2nd Edition. John Wiley & Sons Inc., New York.

Jones, M.C. (1994) On kernel density derivative estimation. *Communications in Statitsics, Theory and Methods*, **23**, 2133–2139.

Jones, M.C. (1993) Simple boundary correction for kernel derivative estimation. *Statist. Computing*, **3**, 135–146.

Jones, M.C., Foster, P.J. (1993) Generalized jackknifing and higher order kernels. *Journal of Nonparametric Statistics*, **3**, 81–94.

Jones, M.C., Marron, J.S., Park, B.U. (1991) A simple root-n bandwidth selector. *Annals of Statistics*, **19**, 1919–1932.

Jones, M.C., Marron, J.S., Sheather, S.J. (1996) A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.

Jones, M.C., Sheather, S.J. (1991) Using nonstochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability Letters*, **11**, 511–514.

Kaffes, D., Rao, M.B. (1982) Weak consistency of least square estimators in linear models. *Journal of Multivariate Analysis*, **12**, 186–198.

Kaufman, B., Onsager, L. (1949) Crystal statistics III: short-range order in a binary ising lattice. *Physical Review*, **76**, 1244–1252.

Keller, M. (2011) *Swiss National Forest Inventory. Manual of the Field Survey 2004–2007*. Swiss Federal Research Institute WSL, Birmensdorf.

Kendall, M.G., Stuart, A. (1963) *The Advanced Theory of Statistics*, Vol. I, 2nd Edition. Charles Griffin & Company, London.

Köhler, M., Schindler, A., Sperlich, S. (2014) A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review*, **82**, 243–274.

Koutrevelis, I.A. (1980) A Taylor series-of-fit test of simple hypotheses based on the empirical characteristic function. *Biometrika*, **67**, 238–240.

Koutrevelis, I.A., Meintanis, S.G. (1999) Testing for stability based on the empirical characteristic funstion with applications to financial data. *Journal of Statistical Computation and Simulation*, **64**, 275–300.

Kronmal, R., Tarter, M.E. (1968) The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association*, **63**, 925–952.

Kuan, C.M., Hsu, C.C. (1998) Change-point estimation of fractionally integrated processes. *Journal of Time Series Analysis*, **19**, 693–708.

Künsch, H. (1986) Statistical aspects of self-similar processes. In *Proceedings of the First World Congress of the Bernoulli Society, Tashkent*, Vol. 1, Yu. Prohorov and V.V. Sazonov (Eds.). VNU Science Press, Utrecht, pp. 67–74.

Lai, T.L., Robbins, H., Wei, C.Z. (1979) Strong consistency of least square estimates in multiple regresison II. *Journal of Multivariate Analysis*, **9**, 343–361.

Lavancier, F. (2006) Long memory random fields. In *Dependence in Probability and Statistics, Lecture Notes in Statistics*, Vol. 187, P. Bertail, P., Doukhan and P., Soulier (Eds.). Springer-Verlag, New York, pp. 195–220.

Leonenko, N. (1999) *Limit Theorems for Random Fields with Singular Spectrum*. Kluwer Academic Publishers, Dordrecht.

Lerman, R.I., Yitzhaki, S. (1984) A note on the calculation and interpretation of the Gini index. *Economic Letters*, **15**, 363–368.

Lévy-Leduc, C., Moulines, E., Roueff, F. (2008) Frequency estimation based on the cumulated Lomb–Scargle periodogram. *Journal of Time Series Analysis*, **29**, 1104–1131.

Li, Y., Ruppert, D. (2008) On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436.

Loader, C.R. (1996) Change point estimation using nonparametric regression. *Annals of Statistics*, **24**, 1667–1678.

Loader, C.R. (1999) Bandwidth selection: classical or plug-in? *Annals of Statistics*, **27**, 415–438.

Loève, M. (1960) *Probability Theory*. Van Nostrand, Princeton.

Lorenz, M.O. (1905) Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, **9**, 209–219.

Mack, Y.P., Rosenblatt, M. (1979) Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, **9**, 1–15.

Mack, Y.P., Silverman, B.W. (1982) Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields*, **61**, 405–415.

Major, P. (1981) Limit theorems for non-linear functionals of Gaussian sequences. *Probability Theory and Related Fields*, **57**, 129–158.

Mandelbrot, B.B., van Ness, J.W. (1968) Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, **10**, 422–437.

Marron, J.S., Nolan, D. (1989) Canonical kernels for density estimation. *Statistics and Probability Letters*, **7**, 195–199.

Marron, J.S., Ruppert, D. (1994) Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society, Ser. B*, **56**, 653–671.

Marron, J.S., Wand, M.P. (1992) Exact mean integrated squared error. *Annals of Statistics*, **20**, 712–736.

Masry, E. (1978) Poisson sampling and spectral estimation of continuous-time processes. *IEEE Transactions on Information Theory*, **24**, 173–183.

Menéndez, P., Ghosh, S., Beran, J. (2010) On rapid change points under long-memory. *Journal of Statistical Planning and Inference*, **40**, 3343–3354.

Menéndez, P., Ghosh, S., Künsch, H., Tinner, W. (2013) On trend estimation under monotone Gaussian subordination with long-memory: application to fossil pollen series. *Journal of Nonparametric Statistics*, **25**, 765–785.

Müller, H.-G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics*, **12**, 766–774.

Müller, H.-G. (1992) Change-points in nonparametric regression analysis. *Annals of Statistics*, **20**, 737–761.

Müller, H.-G. (1993) On the boundary kernel method for nonparametric curve estimation near endpoints. *Scandinavian Journal of Statistics*, **20**, 313–328.

Müller, H.-G. (1997) Density estimation. In *Encyclopedia of Statistical Sciences*, S. Kotz, C.B. Read, and D.L. Banks. John Wiley & Sons Inc., New York, pp. 185–200.

Müller, H.-G., Stadt Müller, U., Schmitt, T. (1987) Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, **74**, 743–749.

Müller, H.-G., Wang, J.L. (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, **50**, 61–76.

Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and Its Applications*, **15**, 134–137.

Nadaraya, E.A. (1965) On non-parametric estimates of density functions and regression curves. *Theory of Probability and Its Applications*, **10**, 186–190.

Nadaraya, E.A. (1970) Remarks on non-parametric estimates for density functions and regression curves. *Theory of Probability and Its Applications*, **15**, 134–137.

Opsomer, J.D., Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **25**, 186–211.

Opsomer, J.D., Ruppert, D., Wand, M.P., Holst, U., Hossjer, O. (1999) Kriging with nonparametric variance function estimation. *Biometrics*, **55**, 704–710.

Padgett, W.J., McNichols, D.T. (1984) Nonparametric density estimation from censored data. *Communications in Statistics, Theory and Methods*, **13**, 1581–1611.

Park, B.U., Marron, J.S. (1990) Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, **85**, 66–72.

Parzen, E. (1957) On consistent estimates of the spectrum of a stationary time series. *Annals of Mathematical Statistics*, **28**, 329–348.

Parzen, E. (1961) Mathematical considerations in the estimation of spectra. *Technometrics*, **3**, 167–190.

Parzen, E. (1962) On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**, 1065–1076.

Parzen, E. (Ed.) (1984) *Time Series Analysis of Irregularly Observed Data: Proceedings of a Symposium held at Texas A&M University, College Station, Texas, February 10–13, 1983*. Springer-Verlag, New York.

Pearson, K. (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London, A*, **186**, 343–414.

Pearson, K. (1902a) On the systematic fitting of curves to observations and measurements. I. *Biometrika*, **2**, 265–303.

Pearson, K. (1902b) On the systematic fitting of curves to observations and measurements. II. *Biometrika*, **2**, 1–23.

Pearson, E.S. (1936) Note on probability levels for $\sqrt{b_1}$. *Biometrika*, **28**, 306.

Percival, D.B., Walden, A.T. (2000) *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, UK.

Picard, D. (1985) Testing and estimating change-points in time series. *Advances in Applied Probability*, **17**, 841–867.

Pons, O. (2003) Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Annals of Statistics*, **31**, 442–463.

Prakasa Rao, B.L.S. (1983) *Nonparametric Functional Estimation*. Academic Press, New York.

Press W., Teukolsky S., Vetterling W., Flannery B. (1992) *Numerical Recipes in C: The Art of Scientific Computing*, 2nd Edition. Cambridge University Press, Cambridge.

Priestley, M.B. (1989) *Spectral Analysis and Time Series*. John Wiley & Sons Inc., New York.

Priestley, M.B., Chao, M.T. (1972) Nonparametric function fitting. *Journal of The Royal Statistical Society, Ser. B*, **34**, 385–392.

Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*, 2nd Edition. John Wiley & Sons Inc., New York.

Ray, B.K., Tsay, R. (1997) Bandwidth selection for kernel regression with long-range dependent errors. *Biometrika*, **84**, 791–802.

Rice, J. (1984) Bandwidth choice for nonparametric regression. *Annals of Statistics*, **12**, 1215–1230.

Rice, J. (1986) Convergence rates for partially splined models. *Statistics and Probability Letters*, **4**, 203–208.

Rice, J., Rosenblatt, M. (1976) Estimation of the log survivor function and hazard function. *Sankhya, The Indian Journal of Statistics, Ser. A*, **38**, 60–78.

Ripley, B.D. (1981) *Spatial Statistics*. John Wiley & Sons Inc., New York.

Rigollet, P., Tsybakov, A. (2007) Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, **16**, 260–280.

Robinson, P.M. (1983) Nonparametric estimators for time series. *Journal of Time Series Analysis*, **4**, 185–207.

Robinson, P.M. (1984) Robust nonparametric regression. In *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics*, **26**, 247–255.

Robinson, P. (1988) Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931–954.

Robinson, P.M. (1995) Log-periodogram regression of time series with long range dependence. *Annals of Statistics*, **23**, 1048–1072.

Robinson, P.M. (1997) Large-sample inference for nonparametric regression with dependent errors. *Annals of Statistics*, **25**, 2054–2083.

Robinson, P.M., Hidalgo, F.J. (1997) Time series regression with long-range dependence. *Annals of Statistics*, **25**, 77–104.

Roeder, C. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832–835.

Rosenblatt, M. (1971) Curve estimates. *Annals of Statistics*, **42**, 1815–1842.

Rosenblatt, M. (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Annals of Statistics*, **3**, 1–14.

Rosenblatt, M. (1976) On the maximal deviation of k-dimensional density estimates. *Annals of Probability*, **4**, 1009–1015.

Rosenblatt, M. (1984) Stochastic processes with short-range and long-range dependence. In *Statistics: An Appraisal, Proceedings 50th Anniversary Conference*, H.A. David and H.T. David (Eds.). The Iowa State University Press, pp. 509–520.

Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65–78.

Ruppert, D. (2002) Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

Ruppert, D., Wand, M.P. (1994) Multivariate weighted least squares regression. *Annals of Statistics*, **22**, 1346–1370.

Ruppert, D., Wand, M.P., Carroll, R.J. (2009) Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, **3**, 1193–1256.

Sacks, J., Ylvisaker, D. (1981) Asymptotically optimum kernels for density estimation at a point. *Annals of Statistics*, **9**, 334–346.

Sansone, G. (2004) *Orthogonal Functions*. Dover Publications, Mineola, New York.

Schucany, W.R. (1989) Locally optimal window widths for kernel density estimation with large samples. *Statistics and Probability Letters*, **7**, 401–405.

Schucany, W.R. (2004) Kernel smoothers: an overview of curve estimators for the first graduate course in nonparametric statistics. *Statistical Science*, **19**, 663–675.

Schuster, E.F. (1969) Estimation of a probability density function and its derivatives. *Annals of Mathematical Statistics*, **40**, 1187–1195.

Schuster, E.F. (1970) Note on uniform convergence of density estimates. *Annals of Mathematical Statistics*, **41**, 1347–1348.

Schuster, E.F. (1972) Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *Annals of Mathematical Statistics*, **43**, 84–88.

Schuster, E.F., Yakowitz, S. (1979) Contribution to the theory of nonparametric regression with application to system identification. *Annals of Statistics*, **7**, 139–149.

Schwander, J., Eicher, U., Ammann, B. (2000) Oxygen isotopes of Lake Marl at Gerzensee and Leysin (Switzerland), covering the Younger Dryas and two minor oscillations, and their correlation to the GRIP ice core. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **159**, 203–214.

Schwartz, S.C. (1967) Estimation of a probability density by an orthogonal series. *Annals of Mathematical Statistics*, **38**, 1262–1265.

Schweder, T. (1975) Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics*, **2**, 113–126.

Scott, D.W. (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

Scott, D.W. (1985) Average shifted histograms: effective nonparametric density estimators in several dimensions. *Annals of Statistics*, **13**, 1024–1040.

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons Inc., New York.

Scott, D.W., Tapia, R.A., Thompson, J.R. (1977) Kernel density estimation revisited. *Nonlinear Analysis: Theory, Methods and Applications*, **1**, 339–372.

Scott, D.W., Terrell, G.R. (1987) Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131–1146.

Scott, D.W., Wand, M.P. (1991) Feasibility of multivariate density estimates. *Biometrika*, **78**, 197–206.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons Inc., New York.

Shapiro, J.S. (1969) *Smoothing and Approximation of Functions*. Van Nostrand-Reinhold, New York.

Sheather, S.J. (1992) The performance of six popular bandwidths selection methods on some real data sets. *Computational Statistics*, **7**, 225–250, 271–281.

Sheather, S.J. (2004) Density estimation. *Statistical Science*, **19**, 588–597.

Sheather, S.J., Jones, M.C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Ser. B*, **53**, 683–690.

Sen, A., Srivastava, M.S. (1990) *Regression Analysis: Theory, Methods and Applications*. Springer-Verlag, New York.

Sibbertsen, P. (1999) Robuste Parameterschätzung im linearen Regressionsmodell bei Fehlertermen mit langem Gedächtnis. Verlag für Wissenschaft und Forschung, Berlin (in German).

Silverman, B.W. (1978) Weak and strong uniform consistency of kernel estimate of a density and its derivative. *Annals of Statistics*, **6**, 177–184.

Silverman, B.W. (1981) Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Ser. B*, **43**, 97–99.

Silverman, B.W. (1984a) A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *Journal of the American Statistical Association*, **79**, 584–589.

Silverman, B.W. (1984b) Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 898–918.

Silverman, B.W. (1986) *Density Estimation*. Chapman & Hall, New York.

Silverman, B.W., Jones, M.C. (1989) E. Fix and J.L. Hodges (1951) Discriminatory analysis and nonparametric estimation: consistency properties. *International Statitsical Review*, **57**, 233–247.

Silvey, S.D. (1975) *Statistical Inference*. Chapman & Hall, New York.

Simonoff, J. (1996) *Smoothing Methods in Statistics*. Springer-Verlag, New York.

Speckman, P. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Ser. B*, **50**, 413–436.

Staudenmayer, J., Ruppert, D., Buonaccorsi, J. (2008) Density estimation in the presence of heteroskedastic measurement error. *Journal of the American Statistical Association*, **103**, 726–736.

Stone, C.J. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **36**, 111–147.

Stone, C.J. (1980) Optimal convergence rates for nonparametric estimators. *Annals of Statistics*, **8**, 1348–1360.

Stone, C.J. (1982) Optimal global rates of convergence of nonparametric regression. *Annals of Statistics*, **10**, 1040–1053.

Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics*, **12**, 1285–1297.

Sturges, H.A. (1926) The choice of a class interval. *Journal of the American Statistical Association*, **21**, 65–66.

Stute, W. (1982) A law of the iterated logarithm for kernel density estimators. *Annals of Probability*, **10**, 414–422.

Stute, W. (1986) Conditional empirical processes. *Annals of Statistics*, **14**, 638–647.

Szegő, G. (2003) Orthogonal polynomials. *Colloquium Publications*, **23**, report 2003, American Mathematical Society, Providence.

Tapia, R.A., Thompson, J.R. (1978) *Nonparametric Probability Density Estimation*. Johns Hopkins University Press, Baltimore.

Taqqu, M.S. (1975) Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Probability Theory and Related Fields*, **31**, 287–302.

Taqqu, M.S. (1979) Convergence of integrated processes of arbitrary Hermite rank. *Probability Theory and Related Fields*, **50**, 53–83.

Terrell, G.R., Scott, D.W. (1992) Variable kernel density estimation. *Annals of Statistics*, **20**, 1236–1265.

Thompson, J.R., Tapia, R.A. (1987) *Nonparametric Function Estimation, Modeling, and Simulation*. Society for Industrial and Applied Mathematics, Philadelphia.

Tinner, W., Lotter, A.F. (2001) Central European vegetation response to abrupt climate change at 8.2 ka. *Geology*, **29**, 551–554.

Titterington, D.M. (1980) A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.

Van Ryzin, J. (1969) On strong consistency of density estimates. *Annals of Mathematical Statistics*, **40**, 1765–1772.

Wahba, G. (1984) Partial spline models for the semiparametric estimation of functions of several variables. In *Analyses for Time Series*, Japan-US Joint Seminar, Institute of Statistical Mathematics, Tokyo, pp. 319–320.

Wahba, G. (1990) *Spline Models for Observational Data*. Society for Industrial Mathematics.

Walter, G., Blum, J. (1979) Probability density estimation using delta sequences. *Annals of Statistics*, **7**, 328–340.

Wand, M.P. (1997) Data-based choice of histogram bin width. *The American Statistician*, **51**, 59–64.

Wand, M.P., Jones, M.C. (1995) *Kernel Smoothing*. Chapman & Hall, London.

Watson, G.S. (1964) Smooth regression analysis. *Sankhya, The Indian Journal of Statisics, Ser. A*, **26**, 359–372.

Watson, G.S. (1969) Density estimation by orthogonal series. *Annals of Mathematical Statistics*, **40**, 1496–1498.

Watson, G.S., Leadbetter, M.R. (1963) On the estimation of the probability density, I. *Annals of Mathematical Statistics*, **34**, 480–491.

Watson, G.S., Leadbetter, M.R. (1964a) Hazard analysis I. *Biometrika*, **41**, 175–184.

Watson, G.S., Leadbetter, M.R. (1964b) Hazard analysis II. *Sankhya, The Indian Journal of Statisics, Ser. A*, **26**, 101–116.

Wegman, E.J. (1982) Density estimation – I. In *Encyclopedia of Statistical Sciences*, Vol. 2, S. Kotz, N.L. Johnson, and C.B. Read (Eds.). John Wiley & Sons Inc., New York, pp. 309–315.

Whittle, P. (1957) Curve and periodogram smoothing. *Journal of the Royal Statistical Society, Ser. B*, **19**, 38–47.

Whittle, P. (1958) On the smoothing of probability density functions. *Journal of the Royal Statistical Society, Ser. B*, **20**, 334–343.

West, M. (1994) Some statistical issues in palaeoclimatology. In *Bayesian Statistics*, Vol. 5, J.O. Berger, J.M. Bernado, A.P. Dawid, and A.F.M. Smith (Eds.). Oxford University Press, Oxford, pp. 461–484.

Woodroofe, M. (1967) On the maximum deviation of the sample density. *Annals of Mathematical Statistics*, **38**, 475–481.

Woodroofe, M. (1970) On choosing a delta-sequence. *Annals of Mathematical Statistics*, **41**, 1665–1671.

Yaglom, A.M. (1987) Einstein's 1914 paper on the theory of irregularly fluctuating series of observations. *IEEE Acoustoustics Speech Signal Process Magazine*, **4**, 7–11.

Yajima, Y. (1991) Asymptotic properties of the LSE in a regression model with long-memory stationary errors. *Annals of Statistics*, **19**, 158–177.

Author Index

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.

d

e

f

Subject Index

Kernel Smoothing: Principles, Methods and Applications, First Edition.
Sucharita Ghosh.
© 2018 John Wiley & Sons Ltd. Published 2018 by John Wiley & Sons Ltd.