

## An Introduction to the Advanced Theory of Nonparametric Econometrics

Interest in nonparametric methodology has grown considerably over the past few decades, stemming in part from vast improvements in computer hardware and the availability of new software that allows practitioners to take full advantage of these numerically intensive methods. This book is written for advanced undergraduate students, intermediate graduate students, and faculty, and provides a complete teaching and learning course at a more accessible level of theoretical rigor than Racine's earlier book co-authored with Qi Li, *Nonparametric Econometrics: Theory and Practice* (2007). The open source R platform for statistical computing and graphics is used throughout in conjunction with the R package `np`. Recent developments in reproducible research is emphasized throughout with appendices devoted to helping the reader get up to speed with R, R Markdown, TeX and Git.

**Jeffrey S. Racine** is Professor in the Department of Economics and Professor in the Graduate Program in Statistics in the Department of Mathematics and Statistics at McMaster University, Canada. He holds the Senator William McMaster Chair in Econometrics and is a Fellow of the Journal of Econometrics. He is co-author of *Nonparametric Econometrics: Theory and Practice* (2007). He has published extensively in his field and has co-authored the R packages `np` and `crs` that are available on the Comprehensive R Archive Network (CRAN).

Cambridge University Press

978-1-108-48340-7 — An Introduction to the Advanced Theory and Practice of Nonparametric Econometrics

Jeffrey S. Racine

Frontmatter

[More Information](#)

---

# An Introduction to the Advanced Theory of Nonparametric Econometrics

A Replicable Approach Using R

JEFFREY S. RACINE

McMaster University, Ontario



CAMBRIDGE  
UNIVERSITY PRESS

**CAMBRIDGE**  
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom  
One Liberty Plaza, 20th Floor, New York, NY 10006, USA  
477 Williamstown Road, Port Melbourne, VIC 3207, Australia  
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India  
79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

[www.cambridge.org](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/9781108483407](http://www.cambridge.org/9781108483407)  
DOI: 10.1017/9781108649841

© Jeffrey S. Racine 2019

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2019

Printed in the United States of America by Sheridan Books, Inc.

*A catalogue record for this publication is available from the British Library.*

ISBN 978-1-108-48340-7 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

# Contents

<b>List of Tables</b>	<b>xii</b>
<b>List of Figures</b>	<b>xiv</b>
<b>Preface</b>	<b>xix</b>
<b>Glossary of Notation</b>	<b>xxv</b>
<b>I Probability Functions, Probability Density Functions, and Their Cumulative Counterparts</b>	<b>1</b>
<b>1 Discrete Probability and Cumulative Probability Functions</b>	<b>3</b>
1.1 Overview	3
1.2 Parametric Probability Function Estimation	5
1.3 Nonsmooth Probability Function Estimation	8
1.4 Smooth Kernel Probability Function Estimation	11
1.4.1 Estimator Properties for Unordered Categorical Variables and Kernels	12
1.4.2 The SMSE-Optimal Smoothing Parameter and Rate of Convergence	16
1.4.3 Asymptotic Normality	18
1.4.4 Kernel Estimation and Shrinkage	18
1.4.5 Estimator Properties for Ordered Categorical Variables and Kernels	19
1.5 Nonsmooth Cumulative Probability Function Estimation	22
1.6 Smooth Kernel Cumulative Probability Function Estimation	25
1.7 The Multivariate Extension	27
1.8 Practitioner’s Corner	29
1.8.1 Estimating Probability Functions in R	29
1.8.2 A Monte Carlo Comparison of Probability Estimators	34
<b>Problem Set</b>	<b>45</b>

<b>2</b>	<b>Continuous Density and Cumulative Distribution Functions</b>	<b>49</b>
2.1	Overview	49
2.2	Parametric Density Function Estimation	50
2.3	Nonsmooth Density Function Estimation	51
2.3.1	The Histogram Density Estimator	51
2.3.2	The Naïve Density Estimator	52
2.4	Smooth Kernel Density Function Estimation	56
2.4.1	Properties of the Rosenblatt-Parzen Kernel Density Estimator	58
2.4.2	The IMSE-Optimal Bandwidth and Rate of Convergence	66
2.4.3	The IMSE-Optimal Kernel Function	67
2.4.4	Asymptotic Normality	69
2.4.5	Bandwidth Selection	71
2.4.6	Bias-Reducing Kernel Functions	75
2.5	Smooth Kernel Cumulative Distribution Function Estimation	77
2.5.1	Properties of the Kernel Cumulative Distribution Function Estimator	77
2.5.2	IMSE-Optimal Bandwidth	80
2.5.3	Asymptotic Normality	81
2.5.4	Bandwidth Selection	81
2.6	Smooth Kernel Quantile Function Estimation	82
2.7	The Multivariate Extension	85
2.7.1	Properties of the Multivariate Kernel Density Estimator	87
2.7.2	Properties of the Multivariate Kernel Cumulative Distribution Function Estimator	88
2.8	Entropy and Information Measures	89
2.8.1	Statistical Mechanics and Information Functions	89
2.8.2	Relative Entropy	91
2.8.3	Joint and Conditional Entropy	93
2.8.4	Mutual Information	93
2.8.5	Entropy and Metricness	94
2.8.6	Entropy and Axiom Systems	94
2.8.7	Entropy, Inference, Robustness, and Consistency	95
2.8.8	Kernel Estimation and Entropy	96
2.9	Practitioner's Corner	97
2.9.1	The Smoothed Bootstrap	103
2.9.2	Testing Univariate Asymmetry	104
2.9.3	Testing Equality of Univariate Densities	106
2.9.4	Testing Nonlinear Pairwise Independence	108
2.9.5	Testing Nonlinear Serial Independence	109

## CONTENTS

vii

2.9.6	Bounded Domains and Boundary Corrections	111
2.9.7	Nonlinear Optimization and Multi-Starting	118
2.9.8	Confidence Bands and Nonparametric Estimation	123
<b>Problem Set</b>		<b>129</b>
<b>3</b>	<b>Mixed-Data Probability Density and Cumulative Distribution Functions</b>	<b>131</b>
3.1	Overview	131
3.2	Smooth Mixed-Data Kernel Density and Cumulative Distribution Function Estimation	132
3.2.1	Properties of the Mixed-Data Smooth Kernel Density Estimator	133
3.2.2	Properties of the Mixed-Data Smooth Kernel Cumulative Distribution Estimator	135
3.3	The Multivariate Extension	135
3.4	Smooth Kernel Copula Function Estimation with Mixed-Data	137
3.4.1	Copulae and Dependence	139
3.5	Practitioner's Corner	142
3.5.1	Testing Equality of Mixed-Data Multivariate Densities	142
3.5.2	Generating Copula Function Contours	143
<b>Problem Set</b>		<b>145</b>
<b>4</b>	<b>Conditional Probability Density and Cumulative Distribution Functions</b>	<b>147</b>
4.1	Overview	147
4.2	Smooth Kernel Conditional Density Function Estimation	148
4.2.1	Bandwidth Selection	149
4.2.2	The Presence of Irrelevant Covariates	150
4.3	Smooth Kernel Conditional Cumulative Distribution Function Estimation	152
4.3.1	Bandwidth Selection	153
4.4	Conditional Quantile Function Estimation	154
4.4.1	Parametric Conditional Quantile Function Estimation	154
4.4.2	Smooth Kernel Conditional Quantile Function Estimation	157
4.5	Binary Choice and Multinomial Choice Models	158
4.5.1	Parametric Binary Choice and Multinomial Choice Models	158
4.5.2	Smooth Kernel Binary Choice and Multinomial Choice Models	159
4.6	Practitioner's Corner	162
4.6.1	Generating Counterfactual Predictions	166

4.6.2	Bootstrapping Counterfactual Predictions	166
4.6.3	The Smoothed Bootstrap	170
4.6.4	Assessing Model Performance	172
4.6.5	Average Treatment Effects and Propensity Score Matching	179
	<b>Problem Set</b>	<b>185</b>
	<b>II Conditional Moment Functions and Related Statistical Objects</b>	<b>187</b>
	<b>5 Conditional Moment Functions</b>	<b>189</b>
5.1	Overview	189
	<b>6 Conditional Mean Function Estimation</b>	<b>193</b>
6.1	Overview	193
6.2	Parametric Conditional Mean Models	195
6.2.1	(Re)-interpretation of Conditional Mean Models	197
6.2.2	Counterfactual Experiments and Conditional Mean Models	199
6.3	Local Constant Kernel Regression	206
6.3.1	Estimator Properties	208
6.3.2	The IMSE-Optimal Bandwidth and Kernel Function	218
6.3.3	Asymptotic Normality	219
6.3.4	Outlier-Resistant Local Constant Kernel Regression	219
6.3.5	Bandwidth Selection	220
6.3.6	A Coefficient of Determination for Nonparametric Regression	222
6.3.7	Local Constant Marginal Effects	223
6.4	Local Polynomial Kernel Regression	226
6.5	The Multivariate Local Polynomial Extension	229
6.6	Local Polynomial Kernel Regression and Shrinkage	232
6.7	Multivariate Mixed-Data Marginal Effects	235
6.7.1	A Consistent Test for Predictor Relevance	236
6.8	Time Series Kernel Regression	240
6.9	Shape Constrained Kernel Regression	245
6.10	Practitioner's Corner	248
6.10.1	Kernel Regression Is Weighted Least Squares Esti- mation	248
6.10.2	Joint Determination of the Polynomial Degree and Bandwidth	249
6.10.3	A Consistent Nonparametric Test for Correct Para- metric Specification	253



## CONTENTS

ix

6.10.4	Shape Constrained Kernel Regression	257
6.10.5	A Multivariate Application of Local Linear Regression	260
6.10.6	Confidence Bands and Nonparametric Estimation	263
6.10.7	Assessing Model Performance	264
6.10.8	Fixed-Effects Panel Data Models	269
	<b>Problem Set</b>	<b>273</b>
<b>7</b>	<b>Conditional Mean Function Estimation with Endogenous Predictors</b>	<b>275</b>
7.1	Overview	275
7.2	Ill-Posed Inverse Problems and Identification	276
7.2.1	Kernel Smoothing and Ill-Posedness	277
7.2.2	Singular Design Matrices and Ill-Posedness	279
7.3	Parametric Instrumental Regression	280
7.4	Nonparametric Instrumental Regression	281
7.5	Practitioner's Corner	285
7.5.1	Estimation of Engel Curves	285
7.5.2	Nonparametric Instrumental Regression with a Linear DGP	285
	<b>Problem Set</b>	<b>289</b>
<b>8</b>	<b>Semiparametric Conditional Mean Function Estimation</b>	<b>291</b>
8.1	Overview	291
8.2	Robinson's Partially Linear Model	291
8.3	Varying Coefficient Models	294
8.4	Semiparametric Single Index Models	296
8.4.1	Ichimura's Method (Continuous $Y$ )	297
8.4.2	Klein and Spady's Method (Binary $Y$ )	298
8.5	Summary	300
8.6	Practitioner's Corner	300
8.6.1	A Specification Test for the Partially Linear Model	300
8.6.2	Assessing Model Performance - Continuous $Y$	301
	<b>Problem Set</b>	<b>307</b>
<b>9</b>	<b>Conditional Variance Function Estimation</b>	<b>309</b>
9.1	Overview	309
9.2	Local Linear Conditional Variance Function Estimation	309
9.3	Practitioner's Corner	311
9.3.1	A Simulated Illustration	311
	<b>Problem Set</b>	<b>313</b>

<b>III</b>	<b>Appendices</b>	<b>315</b>
<b>A</b>	<b>Large and Small Orders of Magnitude and Probability</b>	<b>317</b>
A.1	Big and Small O Notation	317
A.2	Big and Small O in Probability Notation	319
<b>B</b>	<b>R, RStudio, TeX, and Git</b>	<b>323</b>
B.1	Installation of R and RStudio Desktop	323
B.2	What Is R?	323
B.2.1	R in the News	324
B.2.2	Introduction to R	324
B.2.3	Econometrics in R	324
B.3	What Is RStudio Desktop?	325
B.3.1	Introduction to RStudio	325
B.4	Installation of TeX	325
B.5	Installation of Git	325
<b>C</b>	<b>Computational Considerations</b>	<b>327</b>
C.1	Binning Methods	327
C.2	Transforms	328
C.3	Parallelism	328
C.4	Multipole and Tree-Based Methods	328
C.5	Computationally Efficient Kernel Estimation in R	328
<b>D</b>	<b>R Markdown for Assignments</b>	<b>333</b>
D.1	Source Code (R Markdown) for This Document	333
D.2	R, RStudio, TeX, and Git	333
D.3	What Is R Markdown?	333
D.4	Creating a New R Markdown Document in RStudio	334
D.5	Including R Results in Your R Markdown Document	334
D.6	Reading Data from a URL	334
D.7	Including Plots	335
D.8	Including Bulleted and Numbered Lists	336
D.9	Including Tables	337
D.10	Including Verbatim (i.e., Freeform) Text	337
D.11	Typesetting Mathematics	337
D.12	Flexible Document Creation	338
D.13	Knitting Your R Markdown Document	338
D.14	Printing Your Document	338
D.15	Troubleshooting and Tips	339
<b>E</b>	<b>Practicum</b>	<b>343</b>
E.1	Overview	343
E.2	Getting Started with R	343

## CONTENTS

xi

E.2.1	Reading Datasets Created by Other Software Programs	344
E.2.2	Nonparametric Estimation of Density Functions	345
E.3	Introduction to the R Package np: Working with <code>npudens()</code>	346
E.3.1	Introduction to the <code>npksum()</code> Function	348
E.3.2	Applied Nonparametric Density Estimation	349
E.3.3	Introduction to Applied Nonparametric Regression	351
E.3.4	Advanced Use of the <code>npksum()</code> Function	352
E.3.5	Consistent Nonparametric Inference	354
E.3.6	Non-nested Model Comparison	357
E.3.7	Semiparametric Models	359
E.3.8	Nonparametric Discrete Choice Models	360
E.3.9	Shape Constrained Nonparametric Regression	362
	<b>Bibliography</b>	<b>367</b>
	<b>Author Index</b>	<b>391</b>
	<b>Subject Index</b>	<b>397</b>

## List of Tables

1.1	Boy-girl ratio in families. The null probability is $p_0(x) = \binom{8}{x} 0.505^x (1 - 0.505)^{8-x}$ and the expected frequency is $e_x = 1,000 \times p_0(x)$ .	6
1.2	Root SMSE summaries when the parametric model is correctly specified.	37
1.3	Root SMSE summaries when the parametric model is incorrectly specified.	40
2.1	Likelihood objective function value to 6 and 18 digits.	120
3.1	Counts of the number of dependants present in 526 households.	132
4.1	Parametric Logit confusion matrix.	160
4.2	Smooth kernel nonparametric confusion matrix.	160
4.3	Parametric multinomial Logit confusion matrix.	161
4.4	Smooth kernel nonparametric confusion matrix.	162
4.5	Apparent versus expected true model performance (higher values are preferred) and $P$ -values from a test for equality of expected true performance. The cases considered are the kernel versus linear index Logit models, kernel versus linear index with interaction Logit models, and kernel versus BIC-optimal index Logit models. Rejection of the null implies that the kernel-based model has significantly higher mean CCR on independent data.	179
4.6	Parametric confusion matrix.	183
4.7	Nonparametric confusion matrix.	183
6.1	Correctly specified parametric model summary.	196
6.2	Incorrectly specified parametric model summary.	196
6.3	Linear model summary.	201
6.4	Nonlinear (quadratic) model summary.	201

## LIST OF TABLES

xiii

6.5	Counterfactual wage differences for married females due to a one-year increase in potential experience ( $\Delta$ PE) under nonlinear ( $\beta_n(x)$ ) and linear ( $\beta_l(x)$ ) parametric specifications.	203
6.6	Linear parametric model summary.	203
6.7	Nonlinear parametric (quadratic with interactions) model summary.	204
6.8	Apparent versus expected true model performance (lower values are preferred) and $P$ -values from a test for equality of expected true performance based on (i) the kernel versus linear models, (ii) the kernel versus linear with interaction models, and (iii) the kernel versus MW-optimal models (rejection of the null implies the kernel model has significantly lower mean ASPE on independent data).	268
8.1	Parametric Logit model confusion matrix.	299
8.2	Single index model confusion matrix.	300
8.3	Apparent versus expected true model performance (lower values are preferred) and $P$ -values from a test for equality of expected true performance. The cases considered are the kernel versus the parametric model and the kernel versus the semiparametric model. Rejection of the null implies that the kernel model has significantly lower mean ASPE on independent data.	305
D.2	Here's the caption. It, too, may span multiple lines.	337

## List of Figures

1.1	Parametric (binomial) versus nonsmooth nonparametric (sample proportion) probability estimates for the number of boys in families with eight children.	7
1.2	Unordered kernel versus nonsmooth nonparametric (sample proportion) probability estimates for mother's religion taken from the Demographic and Health Survey on childhood nutrition in India for a variety of smoothing parameters. Note how, as $\lambda$ increases, the kernel estimator shrinks towards the discrete uniform distribution $p(x) = 1/c = 0.2$ ( $c = 5$ ).	20
1.3	Ordered kernel versus nonsmooth nonparametric (sample proportion) probability estimates for the boy-girl ratio data using a variety of smoothing parameters. Note how, as $\lambda$ increases, the kernel estimator shrinks towards the discrete uniform distribution $p(x) = 1/c = 0.11$ ( $c = 9$ ).	23
1.4	Ordered kernel versus nonsmooth nonparametric (cumulative sample proportion) cumulative probability estimates for the number of boys in families with eight children for a variety of smoothing parameters.	28
1.5	Boxplots of the root SMSEs when the parametric model is correctly specified.	37
1.6	Probability estimates for parametric and kernel smoothed methods for one draw from the underlying DGP when the parametric model is correctly specified ( $\lambda_{ml} = 0.10$ ).	38
1.7	Boxplots of the root SMSEs when the parametric model is incorrectly specified.	41
1.8	Probability estimates for parametric and kernel smoothed methods for one draw from the underlying DGP when the parametric model is incorrectly specified ( $\lambda_{ml} = 0.75$ ).	42
2.1	The Gaussian parametric density estimate for the eruptions data.	52
2.2	The histogram density estimate for the eruptions data.	53
2.3	The naïve weight functions $w(X_i, x, h)$ and $w(z)$ , $x = 3$ , $h = 2$ .	54
2.4	The nonsmooth naïve density estimate for the eruptions data.	55

## LIST OF FIGURES

xv

- 2.5 The nonsmooth naïve weight function  $w(z)$  versus the smooth kernel function  $K(z)$ . 57
- 2.6 The mechanics of the Rosenblatt-Parzen kernel density estimator,  $\hat{f}(X_1)$ ,  $n = 3$ ,  $h = 0.5$ . 58
- 2.7 The Rosenblatt-Parzen kernel density estimate for the eruptions data. 59
- 2.8 A comparison of the histogram, naïve, and Rosenblatt-Parzen density estimates for the eruptions data. 59
- 2.9 Summary of 1,000 kernel density estimates,  $n = 250$ ,  $\chi^2$  DGP, with three bandwidths (too large [leftmost figures], about right [middle figures], too small [rightmost figures]). The upper figures plot the DGP  $f(x)$  along with the pointwise mean ( $1000^{-1} \sum_{m=1}^{1000} \hat{f}_m(x)$ ), 0.025th pointwise quantile, and 0.975th pointwise quantile. The lower figures present the pointwise squared bias ( $(E \hat{f}(x) - f(x))^2$ ) and pointwise variance. 65
- 2.10 The Rosenblatt-Parzen kernel density estimate for the eruptions data, with a range of ad hoc bandwidths (upper left is undersmoothed, lower right oversmoothed). 72
- 2.11 Data driven bandwidth selection for the eruptions data (Sheather and Jones's plug-in on the left, likelihood cross-validation the right). 75
- 2.12 Second- and fourth-order Gaussian kernels (left and right figures, respectively). 76
- 2.13 The Rosenblatt-Parzen smooth CDF and the ECDF estimates for the eruptions data for a range of bandwidths (upper left is undersmoothed, lower right oversmoothed). 83
- 2.14 The Gaussian distribution function and quantile function. 84
- 2.15 Smooth kernel quantile estimate when  $X \sim \chi_5^2$ ,  $n = 1000$ . 85
- 2.16 Joint PDF and CDF estimates for the Old Faithful dataset. 88
- 2.17 Shannon's information function  $h(x)$ . 90
- 2.18 Shannon's entropy function  $H(X)$ , Bernoulli random variable. 91
- 2.19 Shannon's entropy function, binomial random variable. 92
- 2.20 Shannon's entropy function, Gaussian random variable. 92
- 2.21  $S_\rho$  when  $f_1 = f(x)$  and  $f_2 = g(x)$  represent a variety of distributions. We are comparing univariate densities, so  $S_\rho = \frac{1}{2} \int \left( \sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx$ . 96
- 2.22 Simulated wage distributions. The dashed line is the Gaussian, while the solid line is the  $\chi^2$ . The vertical dashed line in the figure on the right identifies the mean. The horizontal dashed lines in the figure on the right identify the respective probabilities of lying below the mean. 97

2.23	Kernel estimators of the joint density $f(x) = f(x_1, x_2)$ for the original sample (left) and for the smooth resample (right).	104
2.24	Boundary correction via data-reflection.	113
2.25	Boundary correction via transformation.	115
2.26	Boundary kernel function for a range of bandwidths.	116
2.27	Boundary correction via kernel carpentry.	118
2.28	Log-likelihood function for $\lambda \in [0, 1]$ .	122
2.29	Pointwise bootstrap bias estimate.	125
2.30	Nonparametric confidence bands; bias-corrected on the left, uncorrected on the right.	126
2.31	Asymptotic confidence bands; bias-corrected on the left, uncorrected on the right.	127
3.1	Mixed-data bivariate kernel density estimate for the joint PDF of lwage (continuous) and numdeps (ordered).	134
3.2	Mixed-data bivariate kernel density estimate for the joint CDF of lwage (continuous) and numdeps (ordered).	136
3.3	Simulated illustration of a mixed-data copula and copula density.	141
4.1	Nonparametric conditional PDF estimate for the Italian GDP panel.	150
4.2	Effect of oversmoothing on $K(z)/h$ .	152
4.3	Nonparametric conditional CDF estimates for the Italian GDP panel.	154
4.4	Parametric conditional quantile estimates for the Italian GDP panel, $\tau = (0.25, 0.50, 0.75)$ .	156
4.5	Nonparametric conditional quantile estimates for the Italian GDP panel, $\tau = (0.25, 0.50, 0.75)$ .	158
4.6	Conditional density estimates, $X$ and $Y$ joint normal random variates.	163
4.7	Univariate (unconditional) density estimate for $Y$ .	164
4.8	Conditional distribution estimates, $X$ and $Y$ joint normal random variates.	165
4.9	Univariate (unconditional) distribution estimate for $Y$ .	165
4.10	Probability of delivering a low birth weight infant as a function of mother's age and smoking status (top), and increased risk from smoking (bottom).	167
4.11	Bootstrap confidence bands for the parametric estimate of the increased probability of delivering a low birth weight infant for smokers.	169
4.12	Bootstrap confidence bands for the nonparametric estimate of the increased probability of delivering a low birth weight infant for smokers.	170



## LIST OF FIGURES

xvii

4.13	Conditional PDF estimate and smooth bootstrap resample and the associated conditional PDF estimate.	172
6.1	Linear regression estimates for two DGPs. The parametric model on the left is correctly specified and consistent, while the model on the right is incorrectly specified and inconsistent.	196
6.2	Consistent nonparametric regression (local linear) estimates for two DGPs.	197
6.3	Parametric earnings functions and marginal effects functions (U.S. Current Population Survey, 1976).	200
6.4	Comparison of counterfactual and analytic derivatives controlling for off-axis predictors.	205
6.5	The behaviour of the local constant estimator as $h$ increases (as $h \rightarrow \infty$ , $\hat{g}(x) \rightarrow \bar{Y}$ ).	208
6.6	Summary of 1,000 local constant kernel regression estimates, $n = 250$ , $Y_i = \sin(2\pi X_i) + \epsilon_i$ DGP, with three bandwidths (too large [leftmost figures], about right [middle figures], too small [rightmost figures]). The upper figures plot the DGP $g(x)$ along with the pointwise mean ( $1000^{-1} \sum_{m=1}^{1000} \hat{g}_m(x)$ ), 0.025th pointwise quantile, and 0.975th pointwise quantile. The lower figures present the pointwise squared bias ( $(E(\hat{g}(x)) - g(x))^2$ ) and pointwise variance.	218
6.7	Local constant regression, Old Faithful data.	223
6.8	Local constant estimator $\hat{g}(x)$ (left) and marginal effects function $\hat{\beta}(x)$ (right), Old Faithful data.	226
6.9	Summary of 1,000 local linear kernel regression estimates, $n = 250$ , $Y_i = \sin(2\pi X_i) + \epsilon_i$ DGP, with three bandwidths (too large [leftmost figures], about right [middle figures], too small [rightmost figures]). The upper figures plot the DGP $g(x)$ along with the pointwise mean ( $1000^{-1} \sum_{m=1}^{1000} \hat{g}_m(x)$ ), 0.025th pointwise quantile, and 0.975th pointwise quantile. The lower figures present the pointwise squared bias ( $[E(\hat{g}(x)) - g(x)]^2$ ) and pointwise variance.	230
6.10	Nonparametric local linear regression with a variety of bandwidths for the Prestige data.	231
6.11	Cross-validated local constant, local linear, and generalized local polynomial estimators.	235
6.12	Local constant estimation of a stationary univariate AR(1) time series, $Y_t = 0.9Y_{t-1} + \epsilon_t$ .	243
6.13	Percentage change in US weekly gas prices.	245
6.14	Simulated illustration of shape constrained estimation that delivers a deterministic production frontier.	248
6.15	Cross-validated joint determination of the polynomial degree and bandwidth.	252

6.16	Pointwise bootstrap bias estimate.	265
6.17	Nonparametric confidence bands; bias-corrected on the left, uncorrected on the right.	265
7.1	The function of interest $\varphi(z)$ , the conditional mean function $E(Y z)$ , and $E(U z)$ when $Z$ is not exogenous.	276
7.2	The empirical distribution function $F_n(x)$ .	278
7.3	The empirical distribution function $F_n(x)$ , the kernel smoothed distribution function $\hat{F}(x)$ , and the regularized estimate $\hat{f}(x)$ .	279
7.4	The function of interest $\varphi(z)$ , the nonparametric conditional mean estimate $\hat{g}(z)$ , and the regularized nonparametric instrumental regression estimate $\hat{\varphi}(z)$ .	284
7.5	The regularized solution path starting from the initial guess $\hat{g}(z)$ and ending with the instrumental regression estimate $\hat{\varphi}(z)$ via the stopping rule.	284
7.6	The estimated Engel curve for British family expenditure data $\hat{\varphi}(\text{logexp})$ and the conditional mean estimate $\hat{g}(\text{logexp})$ .	286
7.7	Nonparametric local linear instrumental regression with a linear DGP and the linear parametric IV estimate.	288
9.1	Fan and Yao's (1998) and Chen et al.'s (2009) conditional variance function estimators.	312

# Preface

In the early 20th century, the pioneering statistician Sir R. A. Fisher (1890–1962) set in motion what is known today as the *classical parametric Fisherian* approach by casting statistical estimation as a problem involving a *finite* number of parameters. However, parametric models provide only an approximation to the underlying data generating process and may therefore be biased and inconsistent. Models that seek to describe the data generating process in a statistically consistent manner are more involved, since the unknown components in such models are functions that fully characterize the underlying joint distribution of a data sample. *Nonparametric* methods are suitable for the estimation of an unknown function that belongs to a very broadly defined class of functions, and in this context, the number of parameters involved is said to be of *infinite* dimension. Although the complexity of nonparametric estimators often exceeds that of their more rigid parametric counterparts, they offer practitioners alternative approaches that can reveal features present in a data sample that might otherwise remain undetected.

Interest in nonparametric methodology has grown considerably over the past few decades, stemming in part from vast improvements in computer hardware and the availability of new software that allows practitioners to take full advantage of these numerically intensive methods. The earliest work on nonparametric *kernel* estimation of *probability density functions* dates back to the early 1950s (Fix and Hodges, 1951), on kernel estimation of *regression functions* to the 1960s (Watson, 1964), and on kernel estimation of *probability mass functions* to the 1970s (Aitchison and Aitken, 1976). There exist a variety of books that are devoted to nonparametric estimation and inference, although most of them appear to have been written with an audience of advanced graduate students and researchers in mind, and their focus is often on one very specific aspect of the field (e.g., density estimation). A list of notable contributions would include

- Prakasa Rao (1983; Prakasa Rao, 2014) (devoted to large sample properties of various nonparametric estimators)
- Devroye and Györfi (1985) (devoted to the  $L_1$  approach to nonparametric estimation)

- Silverman (1986) (devoted to density estimation and related topics)
- Härdle (1990) (devoted to applied nonparametric regression)
- Scott (1992) (devoted to density estimation and high-dimensional visualization)
- Wand and Jones (1995) (devoted to an accessible treatment of kernel density estimation and regression)
- Fan and Gijbels (1996) (devoted to local polynomial estimation)
- Simonoff (1996) (devoted to smooth density estimation, regression, and ordered categorical data)
- Bowman and Azzalini (1997) (devoted to the application of kernel methods in S-plus)
- Hart (1997) (devoted to nonparametric smoothing and lack-of-fit tests)
- Bosq (1998) (devoted to the theory of kernel methods for dependent data)
- Horowitz (1998) (devoted to semiparametric econometric methods)
- Pagan and Ullah (1999) (first broad treatment of nonparametric econometrics)
- Fan and Yao (2003) (devoted to time series modeling)
- Yatchew (2003) (devoted to applied semiparametric methods using a differencing technique)
- Ruppert et al. (2003) (devoted to semiparametric modeling)
- Härdle et al. (2004) (devoted to nonparametric and semiparametric modeling)
- Wasserman (2006) (devoted to brief accounts of many modern topics in nonparametric inference)
- Li and Racine (2007) (devoted to nonparametric and semiparametric modeling with an emphasis on categorical covariates)
- Tsybakov (2009) (devoted to construction of optimal estimators, minimax optimality and adaptivity)
- Ahamada and Flachaire (2010) (devoted to an accessible introduction to nonparametric and semiparametric econometrics)
- Henderson and Parmeter (2015) (devoted to an accessible treatment of nonparametric econometrics)
- Politis (2015) (devoted to a transformation-based approach to model free inference)
- Hansen (2018) (devoted to econometrics but with chapters for kernel regression and density estimation)

In Li and Racine (2007), our aim was to provide a rigorous and comprehensive treatment of nonparametric econometric methodology, with an emphasis on mixed categorical and continuous data settings, intended for advanced graduate students and researchers looking to keep abreast of this rapidly growing field. The accompanying R (R Core Team, 2018) package, titled *np*, (Hayfield and Racine, 2008) was intended to facilitate the implementation

in applied research settings of many of the methods that we discussed. We are grateful for the constructive criticism and helpful feedback that we have received about these projects, and we owe an enormous debt to the scores of researchers whose work made them possible.

In this book, we are aiming our attention squarely at advanced undergraduate students, intermediate graduate students, and faculty who wish to explore this exciting field, although not necessarily at the level of theoretical rigour that was found in our previous treatment. We take a more *organic* approach than existing treatments of the subject, and present a unique sequence of topics that are not collectively found elsewhere. We begin with a simple estimator that is standard fare in introductory statistics courses, namely the sample proportion, which is a nonsmooth nonparametric estimator of an unknown probability. This serves as preliminary motivation for the progressive introduction of kernel-smoothing, density estimation, conditional density estimation, and the estimation of more general conditional moments such as the conditional mean (regression), variance, and related objects. Proof concepts are illustrated *once* when each unique case is first encountered, whereas proofs that are of a similar nature to those already treated are either relegated to exercises or accompanied by citation info so that the interested reader may find them in existing treatments. Our approach emphasizes the plug-in principle that is the essence of most nonparametric methods. This involves identifying a fundamental statistical object (e.g., a conditional mean), expressing the object in terms of unknown density or distribution functions, and then plugging in *smooth* and *consistent* estimates of these unknowns. Special attention is also given to smoothing parameter selection and to the statistical properties of the estimator that results.

Our treatment of nonparametric estimation evolves along the lines of what one might encounter in an introductory statistics course, closely following the conventional sequence of topics. That convention is to first introduce discrete probability (i.e., mass) functions in Chapter 1 and then proceed to the study of continuous probability density functions in Chapter 2. However, one chapter that is conspicuously absent from introductory courses is a chapter on probability distributions with mixed discrete and continuous features (such problems are known to be “parametrically awkward” (Aitchison and Aitken, 1976, page 419)). In a nonparametric framework, modeling such objects isn’t awkward at all, and hence we fill this gap in Chapter 3 with a treatment of mixed discrete and continuous probability density functions and their cumulative counterparts. Moreover, it will be seen that we can subsequently tackle in a seamless manner *any* statistical object that is defined over mixed discrete and continuous data. Along the way, we will also cover nonparametric estimation of smooth quantile functions and copula functions. We then consolidate and fix notation by means of a parsimonious representation of the mixed-data multivariate product kernel. This then

allows us to plunge into a range of methods for estimation and inference including nonparametric regression, nonparametric modeling of volatility, as well as methods for stationary time series.

We assess *pointwise* and *global* estimation error via the mean square and integrated (summed) mean square error, respectively. The pointwise error of estimation at a given point  $x$  is the difference between an estimate of the statistical object of interest and the object itself. For instance, we might compute the difference between the empirical CDF  $F_n(x)$  and the unknown CDF  $F(x)$ . Pointwise error is a simple measure that is useful for the construction of confidence intervals. The *integrated* mean square error (or the *summed* mean square error in the context of discrete support random variables) measures the overall error of estimation and is useful as a criterion for bandwidth selection. *Uniform* error is another metric that is computed as the maximal difference between the estimate and the object, i.e.,  $\sup_x |F_n(x) - F(x)|$ . It is typically approached using empirical process theory (Prakasa Rao, 2014). Uniform error is useful for placing bounds on other types of error and establishing *simultaneous* or uniform confidence bands. In this book, we consider only the first two types of error (pointwise and global) and direct the reader whose interest lies in uniform error to other more advanced treatments.

We emphasize how kernel estimators can be interpreted as *shrinkage* estimators (Stein, 1956), as demonstrated in Kiefer and Racine (2009) and Kiefer and Racine (2017). From this perspective, the local constant, local linear, and other variants of local polynomial kernel estimators can be improved; for a broad class of data-generating processes (the class of *analytic* functions), these estimators are able to achieve the rate of convergence that is associated with correctly specified parametric models. Theoretical underpinnings for this result, which is achieved through joint selection of the polynomial degree and bandwidth vectors, can be found in Hall and Racine (2015). Although this approach requires a solution to a mixed-integer problem, its implementation is now feasible in R, and this represents an exciting advance in the area of local polynomial estimation of statistical objects. The interpretation of kernel methods as shrinkage estimators is underscored wherever appropriate in each chapter. Simulations and practical exercises reveal that the performance of this estimator may be superior to that of alternative approaches that are based on ad hoc selection of the polynomial order. Our perspective on kernel estimators, as seen through the lens of shrinkage estimators, is quite novel and, to the best of our knowledge, is not found elsewhere.

The computational run time of various routines in the R package `np` (Hayfield and Racine, 2008) can be reduced through their ability to exploit the power of multiple processors (see the R package `npRmpi`) and through their incorporation of algorithmic enhancements such as the use of trees. That being said, kernel methods are computationally intensive relative to

many of their parametric peers; however, patience in this regard often pays dividends.

R code for all examples in this book is sourced from an R Markdown script and can be studied and modified by readers (this document is composed in R Markdown and uses R bookdown extensions (<https://bookdown.org/yihui/bookdown>) (Xie, 2017)). Each chapter ends with a *Practitioner's Corner* that provides a set of commented examples in R that can be refined by the reader to suit their needs. A solutions manual is available to instructors along with  $\text{\LaTeX}$  Beamer PDF formatted slides authored in R Markdown that can be modified and tailored to an instructor's needs.

In this book, we derive results only for the *notationally parsimonious* case involving univariate data (or univariate conditioning/conditioned variables). Where appropriate, we present results for the multivariate case and draw attention to the salient differences between the two; however, for a thorough theoretical treatment of the multivariate cases, we simply direct the interested readers to Li and Racine (2007) and other sources. It is our conjecture that essentially all of the intuition underlying nonparametric kernel methods can be distilled from the univariate case, at least from the theoretical perspective. However, from the applied perspective, we impose no such limitations, and emphasize cases involving multivariate (and often mixed multivariate) data throughout.

We also touch upon a number of practical aspects of nonparametric kernel methods such as *kernel carpentry* (i.e., the construction of kernel functions with certain useful properties), and provide empirical examples to illustrate these concepts. We encourage the use of tools that facilitate reproducible research.

This book would not exist without the legacy (and ongoing) contributions of an incredibly talented global network of academics harbouring a wide array of research interests in the field of nonparametric statistics and econometrics. If you are reading this and have contributed to this exciting field, please take a virtual bow and accept our heartfelt thanks.

I would like to thank an abbreviated cast of characters, without whom this project would not exist. Qi Li, a co-author on a range of projects, has been an ongoing source of guidance, support, and encouragement. Tristen Hayfield and Zhenghua Nie, co-authors on the R packages `np` and `crs`, respectively, have helped craft user-friendly and computationally efficient implementations of the procedures that are detailed in this book. Nick Kiefer, a co-author, was the first to open my eyes to the interpretation of kernel estimators as shrinkage estimators. Peter Hall, a co-author whose acumen, friendship, and wisdom are sorely missed, made enduring contributions to the field and left a rich legacy that will surely last for generations. I would also like to thank but not implicate John Kealey, a former Ph.D. student who painstakingly pored through this book and polished its many rough edges, along with the students

and faculty who attended a graduate course at McMaster University in the Fall of 2017 and who reported numerous typos in early drafts of this book (Alyssa, Anthony, Camille, Francis, James, Joaquin, Karen, Mark, Richard, Yuyan, and Zvezdomir). And last but certainly not least, I am indebted to my wife Jennifer and son Adam, who endured far too many months of my seven-day-a-week obsession with this project.

This book is dedicated to the memory of our kind, gentle, generous, and irreplaceable colleague, Peter Gavin Hall AO FAA FRS (November 21, 1951—January 9, 2016), an Australian researcher who worked in the areas of probability theory and mathematical statistics. Peter was described by the American Statistical Association as one of the most influential and prolific theoretical statisticians in the history of the field. It is fitting that The School of Mathematics and Statistics Building at The University of Melbourne was renamed the Peter Hall Building in his honour on December 9, 2016.



# Glossary of Notation

Object	Brief Definition
$\beta(x)$	marginal effects function (derivative or finite difference of $g(x)$ )
$\hat{\beta}(x)$	kernel smoothed marginal effects function (derivative or finite difference of $\hat{g}(x)$ )
$C(u_x, u_y)$	bivariate copula function
$f(x)$	probability density function
$\hat{f}(x)$	kernel smoothed probability density function
$f(y x)$	conditional probability density function
$\hat{f}(y x)$	kernel smoothed conditional probability density function
$F(x)$	cumulative distribution function
$F_n(x)$	empirical cumulative distribution function
$\hat{F}(x)$	kernel smoothed cumulative distribution function
$F(y x)$	cumulative conditional distribution function
$\hat{F}(y x)$	kernel smoothed cumulative conditional distribution function
$\gamma$	vector of bandwidths and smoothing parameters for $q$ continuous, $r$ unordered, and $s$ ordered covariates
$G((x - X_i)/h)$	continuous support univariate cumulative probability density kernel function
$G_\gamma(X_i, x)$	mixed-data multivariate cumulative probability density kernel function
$g(x)$	conditional mean function
$\hat{g}(x)$	kernel smoothed conditional mean function
$h$	bandwidth for continuous covariate
$K((x - X_i)/h)$	continuous support univariate probability density kernel function
$K_\gamma(X_i, x)$	mixed-data multivariate probability density kernel function
$\lambda$	smoothing parameter for discrete covariate

Object	Brief Definition
$l(X_i, x, \lambda)$	unordered discrete support univariate probability mass kernel function
$L(X_i, x, \lambda)$	ordered discrete support univariate probability mass kernel function
$\mathcal{L}(X_i, x, \lambda)$	ordered discrete support univariate cumulative probability mass kernel function
$M(x)$	conditional mode function
$\hat{M}(x)$	kernel smoothed conditional mode function
$p(x)$	probability mass function
$p_n(x)$	empirical probability mass function (sample proportion)
$\hat{p}(x)$	kernel smoothed probability mass function
$q_\tau$	unconditional quantile function (inverse CDF)
$\hat{q}_\tau$	kernel smoothed unconditional quantile function
$q_\tau(x)$	conditional quantile function (inverse conditional CDF)
$\hat{q}_\tau(x)$	kernel smoothed conditional quantile function