

# Springer Series in Statistics

*Advisors:*

P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg,  
I. Olkin, N. Wermuth, S. Zeger

**Springer**

*New York*

*Berlin*

*Heidelberg*

*Hong Kong*

*London*

*Milan*

*Paris*

*Tokyo*

László Györfi  
Adam Krzyżak

Michael Kohler  
Harro Walk

# A Distribution-Free Theory of Nonparametric Regression

With 86 Figures



Springer

László Györfi  
Department of Computer Science and  
Information Theory  
Budapest University of Technology and  
Economics  
1521 Stoczek, U.2.  
Budapest  
Hungary  
gyorfi@inf.bme.hu

Adam Krzyżak  
Department of Computer Science  
Concordia University  
1455 De Maisonneuve Boulevard West  
Montreal, Quebec, H3G 1M8  
Canada  
krzyzak@cs.concordia.ca

Michael Kohler  
Fachbereich Mathematik  
Universität Stuttgart  
Pfaffenwaldring 57  
70569 Stuttgart  
Germany  
kohler@mathematik.uni-stuttgart.de

Harro Walk  
Fachbereich Mathematik  
Universität Stuttgart  
Pfaffenwaldring 57  
70569 Stuttgart  
Germany  
walk@mathematik.uni-stuttgart.de

Library of Congress Cataloging-in-Publication Data  
A distribution-free theory of nonparametric regression / László Györfi . . . [et al.].  
p. cm. — (Springer series in statistics)  
Includes bibliographical references and index.  
ISBN 0-387-95441-4 (alk. paper)  
1. Regression analysis. 2. Nonparametric statistics. 3. Distribution (Probability theory)  
I. Györfi, László. II. Series.  
QA278.2 .D57 2002  
519.5'36—dc21

2002021151

ISBN 0-387-95441-4      Printed on acid-free paper.

© 2002 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1      SPIN 10866288

Typesetting: Pages created by the authors using a Springer TeX macro package.

[www.springer-ny.com](http://www.springer-ny.com)

Springer-Verlag   New York Berlin Heidelberg  
*A member of BertelsmannSpringer Science+Business Media GmbH*

*To our families:*

*Kati, Kati, and Jancsi*

*Judith, Iris, and Julius*

*Henryka, Jakub, and Tomasz*

*Hildegard*

# Preface

The *regression estimation problem* has a long history. Already in 1632 Galileo Galilei used a procedure which can be interpreted as fitting a linear relationship to contaminated observed data. Such fitting of a line through a cloud of points is the classical linear regression problem. A solution of this problem is provided by the famous principle of least squares, which was discovered independently by A. M. Legendre and C. F. Gauss and published in 1805 and 1809, respectively. The principle of least squares can also be applied to construct *nonparametric* regression estimates, where one does not restrict the class of possible relationships, and will be one of the approaches studied in this book.

Linear regression analysis, based on the concept of a regression function, was introduced by F. Galton in 1889, while a probabilistic approach in the context of multivariate normal distributions was already given by A. Bravais in 1846. The first nonparametric regression estimate of local averaging type was proposed by J. W. Tukey in 1947. The partitioning regression estimate he introduced, by analogy to the classical partitioning (histogram) density estimate, can be regarded as a special least squares estimate.

Some aspects of nonparametric estimation had already appeared in belletristic literature in 1930/31 in *The Man Without Qualities* by Robert Musil (1880-1942) where, in Section 103 (first book), methods of partitioning estimation are described: "... as happens so often in life, you ... find yourself facing a phenomenon about which you can't quite tell whether it is a law or pure chance; that's where things acquire a human interest. Then you translate a series of observations into a series of figures, which you divide into categories to see which numbers lie between this value and that,

and the next, and so on .... You then calculate the degree of aberration, the mean deviation, the degree of deviation from some arbitrary value ... the average value ... and so forth, and with the help of all these concepts you study your given phenomenon" (cited from page 531 of the English translation, Alfred A. Knopf Inc., Picador, 1995).

Besides its long history, the problem of regression estimation is of increasing importance today. Stimulated by the tremendous growth of information technology in the past 20 years, there is a growing demand for procedures capable of automatically extracting useful information from massive highly-dimensional databases that companies gather about their customers. One of the fundamental approaches for dealing with this "data-mining problem" is regression estimation. Usually there is little or no *a priori* information about the data, leaving the researcher with no other choice but a nonparametric approach.

This book presents a modern approach to nonparametric regression with random design. The starting point is a prediction problem where minimization of the mean squared error (or  $L_2$  risk) leads to the regression function. If the goal is to construct an estimate of this function which has mean squared prediction error close to the minimum mean squared error, then this goal naturally leads to the  $L_2$  error criterion used throughout this book.

We study almost all known regression estimates, such as classical local averaging estimates including kernel, partitioning, and nearest neighbor estimates, least squares estimates using splines, neural networks and radial basis function networks, penalized least squares estimates, local polynomial kernel estimates, and orthogonal series estimates. The emphasis is on the *distribution-free* properties of the estimates, and thus most consistency results presented in this book are valid for all distributions of the data. When it is impossible to derive distribution-free results, as is the case for rates of convergence, the emphasis is on results which require as few constraints on distributions as possible, on distribution-free inequalities, and on adaptation.

Our aim in writing this book was to produce a self-contained text intended for a wide audience, including graduate students in statistics, mathematics, computer science, and engineering, as well as researchers in these fields. We start off with elementary techniques and gradually introduce more difficult concepts as we move along. Chapters 1–6 require only a basic knowledge of probability. In Chapters 7 and 8 we use exponential inequalities for the sum of independent random variables and for the sum of martingale differences. These inequalities are proven in Appendix A. The remaining part of the book contains somewhat more advanced concepts, such as almost sure convergence together with the real analysis techniques given in Appendix A. The foundations of the least squares and penalized least squares estimates are given in Chapters 9 and 19, respectively.

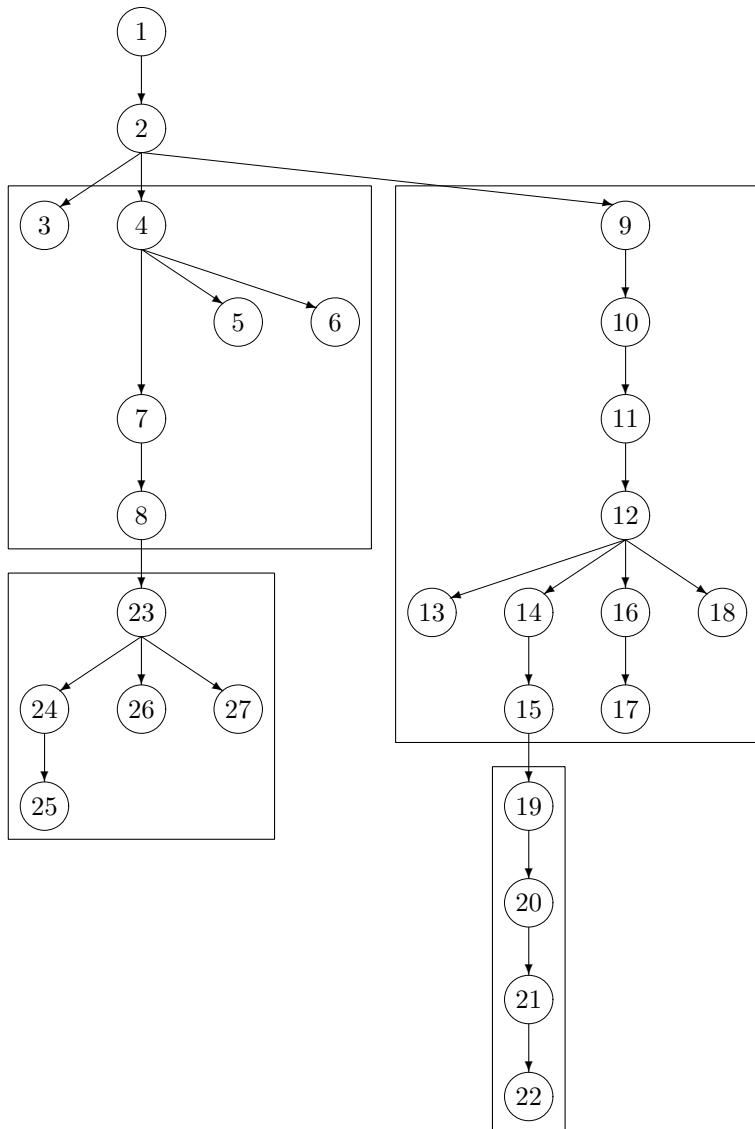


Figure 1. The structure of the book.

The structure of the book is shown in Figure 1. This figure is a precedence tree which could assist an instructor in organizing a course based on this

book. It shows the sequence of chapters needed to be covered in order to understand a particular chapter. The focus of the chapters in the upper-left box is on local averaging estimates, in the lower-left box on strong consistency results, in the upper-right box on least squares estimation, and in the lower-right box on penalized least squares.

We would like to acknowledge the contribution of many people who influenced the writing of this book. Luc Devroye, Gábor Lugosi, Eric Regener, and Alexandre Tsybakov made many invaluable suggestions leading to conceptual improvements and better presentation. A number of colleagues and friends have, often without realizing it, contributed to our understanding of nonparametrics. In particular we would like to thank in this respect Paul Algoet, Andrew Barron, Peter Bartlett, Lucien Birgé, Jan Beirlant, Alain Berlinet, Sándor Csibi, Miguel Delgado, Jürgen Dippon, Jerome Friedman, Włodzimierz Greblicki, Iain Johnstone, Jack Koplowitz, Tamás Linder, Andrew Nobel, Mirek Pawlak, Ewaryst Rafajlowicz, Igor Vajda, Sara van de Geer, Edward van der Meulen, and Sid Yakowitz. András Antos, András György, Michael Hamers, Kinga Máthé, Dániel Nagy, Márta Pintér, Dominik Schäfer and Stefan Winter provided long lists of mistakes and typographical errors. Sándor Győri drew the figures and gave us advice and help on many L<sup>A</sup>T<sub>E</sub>X-problems. John Kimmel was helpful, patient and supportive at every stage.

In addition, we gratefully acknowledge the research support of the Budapest University of Technology and Economics, the Hungarian Academy of Sciences (MTA SZTAKI, AKP, and MTA IEKCS), the Hungarian Ministry of Education (FKFP and MÖB), the University of Stuttgart, Deutsche Forschungsgemeinschaft, Stiftung Volkswagenwerk, Deutscher Akademischer Austauschdienst, Alexander von Humboldt Stiftung, Concordia University, Montreal, NSERC Canada, and FCAR Quebec.

Early versions of this text were tried out at a DMV seminar in Oberwolfach, Germany, and in various classes at the Carlos III University of Madrid, the University of Stuttgart, and at the International Centre for Mechanical Sciences in Udine. We would like to thank the students there for useful feedback which improved this book.

László Györfi,	Budapest, Hungary
Michael Kohler,	Stuttgart, Germany
Adam Krzyżak,	Montreal, Canada
Harro Walk,	Stuttgart, Germany

June 6, 2002

# Contents

<b>Preface .....</b>	vii
<b>1 Why Is Nonparametric Regression Important? .....</b>	1
1.1 Regression Analysis and $L_2$ Risk .....	1
1.2 Regression Function Estimation and $L_2$ Error .....	2
1.3 Practical Applications .....	4
1.4 Application to Pattern Recognition .....	6
1.5 Parametric versus Nonparametric Estimation .....	9
1.6 Consistency .....	12
1.7 Rate of Convergence .....	13
1.8 Adaptation .....	14
1.9 Fixed versus Random Design Regression .....	15
1.10 Bibliographic Notes .....	16
Problems and Exercises .....	16
<b>2 How to Construct Nonparametric Regression Estimates? .....</b>	18
2.1 Four Related Paradigms .....	18
2.2 Curse of Dimensionality .....	23
2.3 Bias–Variance Tradeoff .....	24
2.4 Choice of Smoothing Parameters and Adaptation .....	26
2.5 Bibliographic Notes .....	28
Problems and Exercises .....	29

<b>3</b>	<b>Lower Bounds</b>	31
3.1	Slow Rate	31
3.2	Minimax Lower Bounds	36
3.3	Individual Lower Bounds	43
3.4	Bibliographic Notes	50
	Problems and Exercises	50
<b>4</b>	<b>Partitioning Estimates</b>	52
4.1	Introduction	52
4.2	Stone's Theorem	55
4.3	Consistency	60
4.4	Rate of Convergence	64
4.5	Bibliographic Notes	67
	Problems and Exercises	68
<b>5</b>	<b>Kernel Estimates</b>	70
5.1	Introduction	70
5.2	Consistency	71
5.3	Rate of Convergence	77
5.4	Local Polynomial Kernel Estimates	80
5.5	Bibliographic Notes	82
	Problems and Exercises	82
<b>6</b>	<b>k-NN Estimates</b>	86
6.1	Introduction	86
6.2	Consistency	88
6.3	Rate of Convergence	93
6.4	Bibliographic Notes	96
	Problems and Exercises	97
<b>7</b>	<b>Splitting the Sample</b>	100
7.1	Best Random Choice of a Parameter	100
7.2	Partitioning, Kernel, and Nearest Neighbor Estimates	105
7.3	Bibliographic Notes	108
	Problems and Exercises	109
<b>8</b>	<b>Cross-Validation</b>	112
8.1	Best Deterministic Choice of the Parameter	112
8.2	Partitioning and Kernel Estimates	113
8.3	Proof of Theorem 8.1	115
8.4	Nearest Neighbor Estimates	126
8.5	Bibliographic Notes	127
	Problems and Exercises	127
<b>9</b>	<b>Uniform Laws of Large Numbers</b>	130

9.1	Basic Exponential Inequalities . . . . .	131
9.2	Extension to Random $L_1$ Norm Covers . . . . .	134
9.3	Covering and Packing Numbers . . . . .	140
9.4	Shatter Coefficients and VC Dimension . . . . .	143
9.5	A Uniform Law of Large Numbers . . . . .	153
9.6	Bibliographic Notes . . . . .	156
	Problems and Exercises . . . . .	156
<b>10</b>	<b>Least Squares Estimates I: Consistency . . . . .</b>	<b>158</b>
10.1	Why and How Least Squares? . . . . .	158
10.2	Consistency from Bounded to Unbounded $Y$ . . . . .	165
10.3	Linear Least Squares Series Estimates . . . . .	170
10.4	Piecewise Polynomial Partitioning Estimates . . . . .	174
10.5	Bibliographic Notes . . . . .	180
	Problems and Exercises . . . . .	180
<b>11</b>	<b>Least Squares Estimates II: Rate of Convergence . . . . .</b>	<b>183</b>
11.1	Linear Least Squares Estimates . . . . .	183
11.2	Piecewise Polynomial Partitioning Estimates . . . . .	194
11.3	Nonlinear Least Squares Estimates . . . . .	197
11.4	Preliminaries to the Proof of Theorem 11.4 . . . . .	203
11.5	Proof of Theorem 11.4 . . . . .	210
11.6	Bibliographic Notes . . . . .	219
	Problems and Exercises . . . . .	220
<b>12</b>	<b>Least Squares Estimates III: Complexity Regularization . . . . .</b>	<b>222</b>
12.1	Motivation . . . . .	222
12.2	Definition of the Estimate . . . . .	225
12.3	Asymptotic Results . . . . .	227
12.4	Piecewise Polynomial Partitioning Estimates . . . . .	232
12.5	Bibliographic Notes . . . . .	233
	Problems and Exercises . . . . .	234
<b>13</b>	<b>Consistency of Data-Dependent Partitioning Estimates . . . . .</b>	<b>235</b>
13.1	A General Consistency Theorem . . . . .	235
13.2	Cubic Partitions with Data-Dependent Grid Size . . . . .	241
13.3	Statistically Equivalent Blocks . . . . .	243
13.4	Nearest Neighbor Clustering . . . . .	245
13.5	Bibliographic Notes . . . . .	250
	Problems and Exercises . . . . .	251
<b>14</b>	<b>Univariate Least Squares Spline Estimates . . . . .</b>	<b>252</b>
14.1	Introduction to Univariate Splines . . . . .	252
14.2	Consistency . . . . .	267
14.3	Spline Approximation . . . . .	273

14.4	Rate of Convergence . . . . .	277
14.5	Bibliographic Notes . . . . .	281
	Problems and Exercises . . . . .	281
<b>15</b>	<b>Multivariate Least Squares Spline Estimates . . . . .</b>	<b>283</b>
15.1	Introduction to Tensor Product Splines . . . . .	283
15.2	Consistency . . . . .	290
15.3	Rate of Convergence . . . . .	294
15.4	Bibliographic Notes . . . . .	296
	Problems and Exercises . . . . .	296
<b>16</b>	<b>Neural Networks Estimates . . . . .</b>	<b>297</b>
16.1	Neural Networks . . . . .	297
16.2	Consistency . . . . .	300
16.3	Rate of Convergence . . . . .	315
16.4	Bibliographic Notes . . . . .	326
	Problems and Exercises . . . . .	328
<b>17</b>	<b>Radial Basis Function Networks . . . . .</b>	<b>329</b>
17.1	Radial Basis Function Networks . . . . .	329
17.2	Consistency . . . . .	332
17.3	Rate of Convergence . . . . .	340
17.4	Increasing Kernels and Approximation . . . . .	348
17.5	Bibliographic Notes . . . . .	350
	Problems and Exercises . . . . .	350
<b>18</b>	<b>Orthogonal Series Estimates . . . . .</b>	<b>353</b>
18.1	Wavelet Estimates . . . . .	353
18.2	Empirical Orthogonal Series Estimates . . . . .	356
18.3	Connection with Least Squares Estimates . . . . .	358
18.4	Empirical Orthogonalization of Piecewise Polynomials . . . . .	361
18.5	Consistency . . . . .	366
18.6	Rate of Convergence . . . . .	372
18.7	Bibliographic Notes . . . . .	378
	Problems and Exercises . . . . .	378
<b>19</b>	<b>Advanced Techniques from Empirical Process Theory . . . . .</b>	<b>380</b>
19.1	Chaining . . . . .	380
19.2	Extension of Theorem 11.6 . . . . .	385
19.3	Extension of Theorem 11.4 . . . . .	390
19.4	Piecewise Polynomial Partitioning Estimates . . . . .	397
19.5	Bibliographic Notes . . . . .	404
	Problems and Exercises . . . . .	405
<b>20</b>	<b>Penalized Least Squares Estimates I: Consistency . . . . .</b>	<b>407</b>

20.1 Univariate Penalized Least Squares Estimates . . . . .	408
20.2 Proof of Lemma 20.1 . . . . .	414
20.3 Consistency . . . . .	418
20.4 Multivariate Penalized Least Squares Estimates . . . . .	425
20.5 Consistency . . . . .	427
20.6 Bibliographic Notes . . . . .	429
Problems and Exercises . . . . .	429
<b>21 Penalized Least Squares Estimates II: Rate of Convergence . . . . .</b>	<b>433</b>
21.1 Rate of Convergence . . . . .	433
21.2 Application of Complexity Regularization . . . . .	440
21.3 Bibliographic notes . . . . .	446
Problems and Exercises . . . . .	447
<b>22 Dimension Reduction Techniques . . . . .</b>	<b>448</b>
22.1 Additive Models . . . . .	449
22.2 Projection Pursuit . . . . .	451
22.3 Single Index Models . . . . .	456
22.4 Bibliographic Notes . . . . .	457
Problems and Exercises . . . . .	457
<b>23 Strong Consistency of Local Averaging Estimates . . . . .</b>	<b>459</b>
23.1 Partitioning Estimates . . . . .	459
23.2 Kernel Estimates . . . . .	479
23.3 k-NN Estimates . . . . .	486
23.4 Bibliographic Notes . . . . .	491
Problems and Exercises . . . . .	491
<b>24 Semirecursive Estimates . . . . .</b>	<b>493</b>
24.1 A General Result . . . . .	493
24.2 Semirecursive Kernel Estimate . . . . .	496
24.3 Semirecursive Partitioning Estimate . . . . .	507
24.4 Bibliographic Notes . . . . .	510
Problems and Exercises . . . . .	511
<b>25 Recursive Estimates . . . . .</b>	<b>512</b>
25.1 A General Result . . . . .	512
25.2 Recursive Kernel Estimate . . . . .	517
25.3 Recursive Partitioning Estimate . . . . .	518
25.4 Recursive NN Estimate . . . . .	518
25.5 Recursive Series Estimate . . . . .	520
25.6 Pointwise Universal Consistency . . . . .	526
25.7 Bibliographic Notes . . . . .	537
Problems and Exercises . . . . .	537

<b>26 Censored Observations</b> .....	540
26.1 Right Censoring Regression Models .....	540
26.2 Survival Analysis, the Kaplan-Meier Estimate .....	541
26.3 Regression Estimation for Model A .....	548
26.4 Regression Estimation for Model B .....	555
26.5 Bibliographic Notes .....	563
Problems and Exercises .....	563
<b>27 Dependent Observations</b> .....	564
27.1 Stationary and Ergodic Observations .....	565
27.2 Dynamic Forecasting: Autoregression .....	568
27.3 Static Forecasting: General Case .....	572
27.4 Time Series Problem: Cesàro Consistency .....	576
27.5 Time Series Problem: Universal Prediction .....	576
27.6 Estimating Smooth Regression Functions .....	582
27.7 Bibliographic Notes .....	587
Problems and Exercises .....	588
<b>Appendix A: Tools</b> .....	589
A.1 A Denseness Result .....	589
A.2 Inequalities for Independent Random Variables .....	592
A.3 Inequalities for Martingales .....	598
A.4 Martingale Convergences .....	601
Problems and Exercises .....	607
<b>Notation</b> .....	609
<b>Bibliography</b> .....	612
<b>Author Index</b> .....	639
<b>Subject Index</b> .....	644