

SPRINGER TEXTS IN STATISTICS

All of Nonparametric Statistics

Larry Wasserman



Springer

SPRINGER TEXTS IN STATISTICS

All of Nonparametric Statistics



Larry Wasserman



Springer

Springer Texts in Statistics

Advisors:

George Casella Stephen Fienberg Ingram Olkin

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: Introduction to Times Series and Forecasting, Second Edition
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data—Nonparametric Regression and Response Surface Maximization, Second Edition
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Third Edition
Creighton: A First Course in Probability Models and Statistical Inference
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling, Second Edition
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lange: Applied Probability
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models

(continued after index)

Larry Wasserman

All of Nonparametric Statistics

With 52 Illustrations



Springer

Larry Wasserman
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA
larry@stat.cmu.edu

Editorial Board

George Casella
Department of Statistics
University of Florida
Gainesville, FL 32611-8545
USA

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Ingram Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

Library of Congress Control Number: 2005925603

ISBN-10: 0-387-25145-6

ISBN-13: 978-0387-25145-5

Printed on acid-free paper.

© 2006 Springer Science+Business Media, 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 Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, 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. (MVY)

9 8 7 6 5 4 3 2 1

springeronline.com

To Isa

Preface

There are many books on various aspects of nonparametric inference such as density estimation, nonparametric regression, bootstrapping, and wavelets methods. But it is hard to find all these topics covered in one place. The goal of this text is to provide readers with a single book where they can find a brief account of many of the modern topics in nonparametric inference.

The book is aimed at master's-level or Ph.D.-level statistics and computer science students. It is also suitable for researchers in statistics, machine learning and data mining who want to get up to speed quickly on modern nonparametric methods. My goal is to quickly acquaint the reader with the basic concepts in many areas rather than tackling any one topic in great detail. In the interest of covering a wide range of topics, while keeping the book short, I have opted to omit most proofs. Bibliographic remarks point the reader to references that contain further details. Of course, I have had to choose topics to include and to omit, the title notwithstanding. For the most part, I decided to omit topics that are too big to cover in one chapter. For example, I do not cover classification or nonparametric Bayesian inference.

The book developed from my lecture notes for a half-semester (20 hours) course populated mainly by master's-level students. For Ph.D.-level students, the instructor may want to cover some of the material in more depth and require the students to fill in proofs of some of the theorems. Throughout, I have attempted to follow one basic principle: never give an estimator without giving a confidence set.

The book has a mixture of methods and theory. The material is meant to complement more method-oriented texts such as Hastie et al. (2001) and Ruppert et al. (2003).

After the Introduction in Chapter 1, Chapters 2 and 3 cover topics related to the empirical CDF such as the nonparametric delta method and the bootstrap. Chapters 4 to 6 cover basic smoothing methods. Chapters 7 to 9 have a higher theoretical content and are more demanding. The theory in Chapter 7 lays the foundation for the orthogonal function methods in Chapters 8 and 9. Chapter 10 surveys some of the omitted topics.

I assume that the reader has had a course in mathematical statistics such as Casella and Berger (2002) or Wasserman (2004). In particular, I assume that the following concepts are familiar to the reader: distribution functions, convergence in probability, convergence in distribution, almost sure convergence, likelihood functions, maximum likelihood, confidence intervals, the delta method, bias, mean squared error, and Bayes estimators. These background concepts are reviewed briefly in Chapter 1.

Data sets and code can be found at:

`www.stat.cmu.edu/~larry/all-of-nonpar`

I need to make some disclaimers. First, the topics in this book fall under the rubric of “modern nonparametrics.” The omission of traditional methods such as rank tests and so on is not intended to belittle their importance. Second, I make heavy use of large-sample methods. This is partly because I think that statistics is, largely, most successful and useful in large-sample situations, and partly because it is often easier to construct large-sample, nonparametric methods. The reader should be aware that large-sample methods can, of course, go awry when used without appropriate caution.

I would like to thank the following people for providing feedback and suggestions: Larry Brown, Ed George, John Lafferty, Feng Liang, Catherine Loader, Jiayang Sun, and Rob Tibshirani. Special thanks to some readers who provided very detailed comments: Taeryon Choi, Nils Hjort, Woncheol Jang, Chris Jones, Javier Rojo, David Scott, and one anonymous reader. Thanks also go to my colleague Chris Genovese for lots of advice and for writing the $\text{\LaTeX}$ macros for the layout of the book. I am indebted to John Kimmel, who has been supportive and helpful and did not rebel against the crazy title. Finally, thanks to my wife Isabella Verdinelli for suggestions that improved the book and for her love and support.

Larry Wasserman
Pittsburgh, Pennsylvania
July 2005

Contents

1	Introduction	1
1.1	What Is Nonparametric Inference?	1
1.2	Notation and Background	2
1.3	Confidence Sets	5
1.4	Useful Inequalities	8
1.5	Bibliographic Remarks	10
1.6	Exercises	10
2	Estimating the CDF and Statistical Functionals	13
2.1	The CDF	13
2.2	Estimating Statistical Functionals	15
2.3	Influence Functions	18
2.4	Empirical Probability Distributions	21
2.5	Bibliographic Remarks	23
2.6	Appendix	23
2.7	Exercises	24
3	The Bootstrap and the Jackknife	27
3.1	The Jackknife	27
3.2	The Bootstrap	30
3.3	Parametric Bootstrap	31
3.4	Bootstrap Confidence Intervals	32
3.5	Some Theory	35

3.6	Bibliographic Remarks	37
3.7	Appendix	37
3.8	Exercises	39
4	Smoothing: General Concepts	43
4.1	The Bias–Variance Tradeoff	50
4.2	Kernels	55
4.3	Which Loss Function?	57
4.4	Confidence Sets	57
4.5	The Curse of Dimensionality	58
4.6	Bibliographic Remarks	59
4.7	Exercises	59
5	Nonparametric Regression	61
5.1	Review of Linear and Logistic Regression	63
5.2	Linear Smoothers	66
5.3	Choosing the Smoothing Parameter	68
5.4	Local Regression	71
5.5	Penalized Regression, Regularization and Splines	81
5.6	Variance Estimation	85
5.7	Confidence Bands	89
5.8	Average Coverage	94
5.9	Summary of Linear Smoothing	95
5.10	Local Likelihood and Exponential Families	96
5.11	Scale-Space Smoothing	99
5.12	Multiple Regression	100
5.13	Other Issues	111
5.14	Bibliographic Remarks	119
5.15	Appendix	119
5.16	Exercises	120
6	Density Estimation	125
6.1	Cross-Validation	126
6.2	Histograms	127
6.3	Kernel Density Estimation	131
6.4	Local Polynomials	137
6.5	Multivariate Problems	138
6.6	Converting Density Estimation Into Regression	139
6.7	Bibliographic Remarks	140
6.8	Appendix	140
6.9	Exercises	142
7	Normal Means and Minimax Theory	145
7.1	The Normal Means Model	145
7.2	Function Spaces	147

7.3	Connection to Regression and Density Estimation	149
7.4	Stein's Unbiased Risk Estimator (SURE)	150
7.5	Minimax Risk and Pinsker's Theorem	153
7.6	Linear Shrinkage and the James–Stein Estimator	155
7.7	Adaptive Estimation Over Sobolev Spaces	158
7.8	Confidence Sets	159
7.9	Optimality of Confidence Sets	166
7.10	Random Radius Bands?	170
7.11	Penalization, Oracles and Sparsity	171
7.12	Bibliographic Remarks	172
7.13	Appendix	173
7.14	Exercises	180
8	Nonparametric Inference Using Orthogonal Functions	183
8.1	Introduction	183
8.2	Nonparametric Regression	183
8.3	Irregular Designs	190
8.4	Density Estimation	192
8.5	Comparison of Methods	193
8.6	Tensor Product Models	193
8.7	Bibliographic Remarks	194
8.8	Exercises	194
9	Wavelets and Other Adaptive Methods	197
9.1	Haar Wavelets	199
9.2	Constructing Wavelets	203
9.3	Wavelet Regression	206
9.4	Wavelet Thresholding	208
9.5	Besov Spaces	211
9.6	Confidence Sets	214
9.7	Boundary Corrections and Unequally Spaced Data	215
9.8	Overcomplete Dictionaries	215
9.9	Other Adaptive Methods	216
9.10	Do Adaptive Methods Work?	220
9.11	Bibliographic Remarks	221
9.12	Appendix	221
9.13	Exercises	223
10	Other Topics	227
10.1	Measurement Error	227
10.2	Inverse Problems	233
10.3	Nonparametric Bayes	235
10.4	Semiparametric Inference	235
10.5	Correlated Errors	236
10.6	Classification	236

10.7 Sieves	237
10.8 Shape-Restricted Inference	237
10.9 Testing	238
10.10 Computational Issues	240
10.11 Exercises	240
Bibliography	243
List of Symbols	259
Table of Distributions	261
Index	263

1

Introduction

In this chapter we briefly describe the types of problems with which we will be concerned. Then we define some notation and review some basic concepts from probability theory and statistical inference.

1.1 What Is Nonparametric Inference?

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible. Usually, this means using statistical models that are infinite-dimensional. Indeed, a better name for nonparametric inference might be infinite-dimensional inference. But it is difficult to give a precise definition of nonparametric inference, and if I did venture to give one, no doubt I would be barraged with dissenting opinions.

For the purposes of this book, we will use the phrase nonparametric inference to refer to a set of modern statistical methods that aim to keep the number of underlying assumptions as weak as possible. Specifically, we will consider the following problems:

1. **(Estimating the distribution function).** Given an IID sample $X_1, \dots, X_n \sim F$, estimate the CDF $F(x) = \mathbb{P}(X \leq x)$. (Chapter 2.)

2. (Estimating functionals). Given an IID sample $X_1, \dots, X_n \sim F$, estimate a functional $T(F)$ such as the mean $T(F) = \int x dF(x)$. (Chapters 2 and 3.)
3. (Density estimation). Given an IID sample $X_1, \dots, X_n \sim F$, estimate the density $f(x) = F'(x)$. (Chapters 4, 6 and 8.)
4. (Nonparametric regression or curve estimation). Given $(X_1, Y_1), \dots, (X_n, Y_n)$ estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$. (Chapters 4, 5, 8 and 9.)
5. (Normal means). Given $Y_i \sim N(\theta_i, \sigma^2)$, $i = 1, \dots, n$, estimate $\theta = (\theta_1, \dots, \theta_n)$. This apparently simple problem turns out to be very complex and provides a unifying basis for much of nonparametric inference. (Chapter 7.)

In addition, we will discuss some unifying theoretical principles in Chapter 7. We consider a few miscellaneous problems in Chapter 10, such as measurement error, inverse problems and testing.

Typically, we will assume that distribution F (or density f or regression function r) lies in some large set $\mathfrak{F}$ called a **statistical model**. For example, when estimating a density f , we might assume that

$$f \in \mathfrak{F} = \left\{ g : \int (g''(x))^2 dx \leq c^2 \right\}$$

which is the set of densities that are not “too wiggly.”

1.2 Notation and Background

Here is a summary of some useful notation and background. See also Table 1.1.

Let $a(x)$ be a function of x and let F be a cumulative distribution function. If F is absolutely continuous, let f denote its density. If F is discrete, let f denote instead its probability mass function. The mean of a is

$$\mathbb{E}(a(X)) = \int a(x) dF(x) \equiv \begin{cases} \int a(x) f(x) dx & \text{continuous case} \\ \sum_j a(x_j) f(x_j) & \text{discrete case.} \end{cases}$$

Let $\mathbb{V}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$ denote the variance of a random variable. If $X_1, \dots, X_n$ are n observations, then $\int a(x) d\hat{F}_n(x) = n^{-1} \sum_i a(X_i)$ where $\hat{F}_n$ is the **empirical distribution** that puts mass $1/n$ at each observation X_i .