SPRINGER TEXTS IN STATISTICS

All of Nonparametric Statistics

Larry Wasserman

Springer

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

Library of Congress Control Number: 2005925603

ISBN-10: 0-387-25145-6
ISBN-13: 978-0387-25145-5

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springeronline.com

Springer Texts in Statistics

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)

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 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

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 $X_1, \ldots, X_n \sim F$, estimate the CDF $F(x) = \mathbb{P}(X \leq x)$. (Chapter 2.)

2. (Estimating functionals). Given an IID sample $X_1, \ldots, X_n \sim F$, 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 $X_1, \ldots, X_n \sim F$, estimate the density $f(x) = F'(x)$. (Chapters 4, 6 and 8.)

4. (Nonparametric regression or curve estimation). Given $(X_1, Y_1), \ldots, (X_n, Y_n)$ estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$. (Chapters 4, 5, 8 and 9.)

5. (Normal means). Given $Y_i \sim N(\theta_i, \sigma^2)$, $i = 1, \ldots, n$, estimate $\theta = (\theta_1, \ldots, \theta_n)$. 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 $X_1, \ldots, X_n$ are n observations, then $\int a(x) d\widehat{F}_n(x) = n^{-1} \sum_i a(X_i)$ where $\widehat{F}_n$ is the **empirical distribution** that puts mass $1/n$ at each observation X_i.

Symbol	Definition
$x_n = o(a_n)$	$\lim_{n\to\infty} x_n/a_n = 0$
$x_n = O(a_n)$	$\lvert x_n/a_n\rvert$ is bounded for all large n
$a_n \sim b_n$	$a_n/b_n \to 1$ as $n \to \infty$
$a_n \asymp b_n$	a_n/b_n and b_n/a_n are bounded for all large n
$X_n \rightsquigarrow X$	convergence in distribution
$X_n \xrightarrow{\text{P}} X$	convergence in probability
$X_n \xrightarrow{\text{a.s.}} X$	almost sure convergence
$\widehat{\theta}_n$	estimator of parameter θ
bias	$\mathbb{E}(\widehat{\theta}_n) - \theta$
se	$\sqrt{\mathbb{V}(\widehat{\theta}_n)}$ (standard error)
$\widehat{\text{se}}$	estimated standard error
MSE	$\mathbb{E}(\widehat{\theta}_n - \theta)^2$ (mean squared error)
Φ	CDF of a standard Normal random variable
z_α	$\Phi^{-1}(1-\alpha)$

TABLE 1.1. Some useful notation.

Brief Review of Probability. The **sample space** Ω is the set of possible outcomes of an experiment. Subsets of Ω are called **events**. A class of events $\mathcal{A}$ is called a σ**-field** if (i) $\emptyset \in \mathcal{A}$, (ii) $A \in \mathcal{A}$ implies that $A^c \in \mathcal{A}$ and (iii) $A_1, A_2, \ldots, \in \mathcal{A}$ implies that $\bigcup_{i=1}^\infty A_i \in \mathcal{A}$. A **probability measure** is a function $\mathbb{P}$ defined on a σ-field $\mathcal{A}$ such that $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{A}$, $\mathbb{P}(\Omega) = 1$ and if $A_1, A_2, \ldots \in \mathcal{A}$ are disjoint then

$$\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mathbb{P}(A_i).$$

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a **probability space**. A **random variable** is a map $X : \Omega \to \mathbb{R}$ such that, for every real x, $\{\omega \in \Omega :\ X(\omega) \leq x\} \in \mathcal{A}$.

A sequence of random variables X_n **converges in distribution** (or converges weakly) to a random variable X, written $X_n \rightsquigarrow X$, if

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) \tag{1.1}$$

as $n \to \infty$, at all points x at which the CDF

$$F(x) = \mathbb{P}(X \leq x) \tag{1.2}$$

is continuous. A sequence of random variables X_n **converges in probability** to a random variable X, written $X_n \xrightarrow{\text{P}} X$, if,

$$\text{for every } \epsilon > 0, \quad \mathbb{P}(\lvert X_n - X\rvert > \epsilon) \to 0 \quad \text{as } n \to \infty. \tag{1.3}$$

A sequence of random variables X_n **converges almost surely** to a random variable X, written $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| = 0) = 1. \tag{1.4}$$

The following implications hold:

$$X_n \xrightarrow{\text{a.s.}} X \quad \text{implies that} \quad X_n \xrightarrow{\text{P}} X \quad \text{implies that} \quad X_n \rightsquigarrow X. \tag{1.5}$$

Let g be a continuous function. Then, according to the **continuous mapping theorem**,

$$\begin{array}{lll} X_n \rightsquigarrow X & \text{implies that} & g(X_n) \rightsquigarrow g(X) \\ X_n \xrightarrow{\text{P}} X & \text{implies that} & g(X_n) \xrightarrow{\text{P}} g(X) \\ X_n \xrightarrow{\text{a.s.}} X & \text{implies that} & g(X_n) \xrightarrow{\text{a.s.}} g(X) \end{array}$$

According to **Slutsky's theorem**, if $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ for some constant c, then $X_n + Y_n \rightsquigarrow X + c$ and $X_n Y_n \rightsquigarrow cX$.

Let $X_1, \ldots, X_n \sim F$ be IID. The **weak law of large numbers** says that if $\mathbb{E}|g(X_1)| < \infty$, then $n^{-1}\sum_{i=1}^n g(X_i) \xrightarrow{\text{P}} \mathbb{E}(g(X_1))$. The **strong law of large numbers** says that if $\mathbb{E}|g(X_1)| < \infty$, then $n^{-1}\sum_{i=1}^n g(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}(g(X_1))$.

The random variable Z has a standard Normal distribution if it has density $\phi(z) = (2\pi)^{-1/2}e^{-z^2/2}$ and we write $Z \sim N(0,1)$. The CDF is denoted by $\Phi(z)$. The α upper quantile is denoted by z_α. Thus, if $Z \sim N(0,1)$, then $\mathbb{P}(Z > z_\alpha) = \alpha$.

If $\mathbb{E}(g^2(X_1)) < \infty$, the **central limit theorem** says that

$$\sqrt{n}(\overline{Y}_n - \mu) \rightsquigarrow N(0, \sigma^2) \tag{1.6}$$

where $Y_i = g(X_i)$, $\mu = \mathbb{E}(Y_1)$, $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and $\sigma^2 = \mathbb{V}(Y_1)$. In general, if

$$\frac{(X_n - \mu)}{\widehat{\sigma}_n} \rightsquigarrow N(0,1)$$

then we will write

$$X_n \approx N(\mu, \widehat{\sigma}_n^2). \tag{1.7}$$

According to the **delta method**, if g is differentiable at μ and $g'(\mu) \neq 0$ then

$$\sqrt{n}(X_n - \mu) \rightsquigarrow N(0, \sigma^2) \implies \sqrt{n}(g(X_n) - g(\mu)) \rightsquigarrow N(0, (g'(\mu))^2\sigma^2). \tag{1.8}$$

A similar result holds in the vector case. Suppose that X_n is a sequence of random vectors such that $\sqrt{n}(X_n - \mu) \rightsquigarrow N(0, \Sigma)$, a multivariate, mean 0 normal with covariance matrix Σ. Let g be differentiable with gradient ∇g such that $\nabla_\mu \neq 0$ where ∇_μ is ∇g evaluated at μ. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \rightsquigarrow N\left(0, \nabla_\mu^T \Sigma \nabla_\mu\right). \tag{1.9}$$

Statistical Concepts. Let $\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$ be a parametric model satisfying appropriate regularity conditions. The **likelihood function** based on IID observations $X_1, \ldots, X_n$ is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

and the **log-likelihood function** is $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$. The maximum likelihood estimator, or MLE $\widehat{\theta}_n$, is the value of θ that maximizes the likelihood. The **score function** is $s(X;\theta) = \partial \log f(x;\theta)/\partial\theta$. Under appropriate regularity conditions, the score function satisfies $\mathbb{E}_\theta(s(X;\theta)) = \int s(x;\theta) f(x;\theta)dx = 0$. Also,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \tau^2(\theta))$$

where $\tau^2(\theta) = 1/I(\theta)$ and

$$I(\theta) = \mathbb{V}_\theta(s(x;\theta)) = \mathbb{E}_\theta(s^2(x;\theta)) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right)$$

is the **Fisher information**. Also,

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1)$$

where $\widehat{\mathsf{se}}^2 = 1/(nI(\widehat{\theta}_n))$. The Fisher information I_n from n observations satisfies $I_n(\theta) = nI(\theta)$; hence we may also write $\widehat{\mathsf{se}}^2 = 1/(I_n(\widehat{\theta}_n))$.

The bias of an estimator $\widehat{\theta}_n$ is $\mathbb{E}(\widehat{\theta}) - \theta$ and the the mean squared error MSE is MSE $= \mathbb{E}(\widehat{\theta} - \theta)^2$. The **bias–variance decomposition** for the MSE of an estimator $\widehat{\theta}_n$ is

$$\text{MSE} = \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}(\widehat{\theta}_n). \tag{1.10}$$

1.3 Confidence Sets

Much of nonparametric inference is devoted to finding an estimator $\widehat{\theta}_n$ of some quantity of interest θ. Here, for example, θ could be a mean, a density or a regression function. But we also want to provide confidence sets for these quantities. There are different types of confidence sets, as we now explain.