Preface

Throughout this book, the term "nonparametric" is used to refer to statistical techniques that do not require a researcher to specify a functional form for an object being estimated. Rather than assuming that the functional form of an object is known up to a few (finite) unknown parameters, we substitute less restrictive assumptions such as smoothness (differentiability) and moment restrictions for the objects being studied. For example, when we are interested in estimating the income distribution of a region, instead of assuming that the density function lies in a parametric family such as the normal or log-normal family, we assume only that the density function is twice (or three times) differentiable. Of course, if one possesses prior knowledge (some have called this "divine insight") about the functional form of the object of interest, then one will always do better by using parametric techniques. However, in practice such functional forms are rarely (if ever) known, and the unforgiving consequences of parametric misspecification are well known and are not repeated here.

Since nonparametric techniques make fewer assumptions about the object being estimated than do parametric techniques, nonparametric estimators tend to be slower to converge to the objects being studied than *correctly specified* parametric estimators. In addition, unlike their parametric counterparts, the convergence rate is typically inversely related to the number of variables (covariates) involved, which is sometimes referred to as the "curse of dimensionality." However, it is often surprising how, even for moderate datasets, nonparametric approaches can reveal structure in the data which might be missed were one to employ common parametric functional specifications. Nonparametric methods are therefore best suited to situations in which (i) one knows little about the functional form of the object being estimated, (ii) the number of variables (covariates) is small, and (iii) the researcher has a reasonably large data set. Points (ii) and (iii) are closely related be-

cause, in nonparametric settings, whether or not one has a sufficiently large sample depends on how many covariates are present. Silverman (1986, see Table 4.2, p. 94) provides an excellent illustration on the relationship between the sample size and the covariate dimension required to obtain accurate nonparametric estimates. We use the term "semiparametric" to refer to statistical techniques that do not require a researcher to specify a parametric functional form for some part of an object being estimated but do require parametric assumptions for the remaining part(s).

As noted above, the nonparametric methods covered in this text offer the advantage of imposing less restrictive assumptions on functional forms (e.g., regression or conditional probability functions) as compared to, say, commonly used parametric models. However, alternative approaches may be obtained by relaxing restrictive assumptions in a conventional parametric setting. One such approach taken by Manski (2003) and his collaborators considers probability or regression models in which some parameters are not identified. Instead of imposing overly strong assumptions to identify the parameters, it is often possible to find bounds for the permissible range for these parameters. When the bound is relatively tight, i.e., when the permissible range is quite narrow, one can *almost* identify these parameters. This exciting line of inquiry, however, is beyond the scope of this text, so we refer the interested reader to the excellent monograph by Manski (2003); see also recent work by Manski and Tamer (2002), Imbens and Manski (2004), Honoré and Tamer (2006) and the references therein.

Nonparametric and semiparametric methods have attracted a great deal of attention from statisticians in the past few decades, as evidenced by the vast array of texts written by statisticians including Prakasa Rao (1983), Devroye and Györfi (1985), Silverman (1986), Scott (1992), Bickel, Klaassen, Ritov and Wellner (1993), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Azzalini and Bowman (1997), Hart (1997), Efromovich (1999), Eubank (1999), Ruppert, Carroll and Wand (2003), and Fan and Yao (2005). However, the number of texts tailored to the needs of applied econometricians is relatively scarce, Härdle (1990), Horowitz (1998), Pagan and Ullah (1999), Yatchew (2003), and Härdle, Müller, Sperlich and Werwatz (2004) being those of which we are currently aware.

In addition, the majority of existing texts operate from the presumption that the underlying data is strictly continuous in nature, while more often than not economists deal with categorical (nominal

and ordinal) data in applied settings. The conventional frequency-based nonparametric approach to dealing with the presence of discrete variables is acknowledged to be unsatisfactory. Building upon Aitchison and Aitken's (1976) seminal work on smoothing discrete covariates, we recently proposed a number of novel nonparametric approaches; see, e.g., Li and Racine (2003), Hall, Racine and Li (2004), Racine and Li (2004), Li and Racine (2004*a*), Racine, Li and Zhu (2004), Ouyang, Li and Racine (2006), Hall, Li and Racine (2006), Racine, Hart and Li (forthcoming), Li and Racine (forthcoming), and Hsiao, Li and Racine (forthcoming) for recent work in this area. In this text we emphasize nonparametric techniques suited to the rich array of data types (continuous, nominal, and ordinal) encountered by an applied economist within one coherent framework.

Another defining feature of this text is its emphasis on the properties of nonparametric estimators in the presence of potentially irrelevant variables. Existing treatments of kernel methods, in particular, bandwidth selection methods, presume that all variables are relevant. For example, existing treatments of plug-in or cross-validation methods presume that all covariates in a regression model are in fact relevant, i.e., that all covariates help explain variation in the outcome (i.e., the dependent variable). When this is not the case, however, existing results such as rates of convergence and the behavior of bandwidths no longer hold; see, e.g., Hall et al. (2004), Hall et al. (2006), Racine and Li (2004), and Li and Racine (2004*a*). We feel that this is an extremely important aspect of sound nonparametric estimation which must be appreciated by practitioners if they are to wield these tools wisely.

This book is aimed at students enrolled in a graduate course in nonparametric and semiparametric methods, who are interested in application areas such as economics and other social sciences. Ideal prerequisites would include a course in mathematical statistics and a course in parametric econometrics at the level of, say, Greene (2003) or Wooldridge (2002). We also intend for this text to serve as a valuable reference for a much wider audience, including applied researchers and those who wish to familiarize themselves with the subject area.

The five parts of this text are organized as follows. The first part covers nonparametric estimation of density and regression functions with independent data, with emphasis being placed on mixed discrete and continuous data types. The second part deals with various semiparametric models again with independent data, including partially linear models, single index models, additive models, varying coefficient

models, censored models, and sample selection models. The third part deals with an array of consistent model specification tests. The fourth part examines nearest neighbor and series methods. The fifth part considers kernel estimation of instrumental variable models, simultaneous equation models, and panel data models, and extends results from previous Chapters to the weakly dependent data setting.

Rigorous proofs are provided for most results in Part I, while outlines of proofs are provided for many results in Parts II, III, IV, and V. Background statistical concepts are presented in an appendix.

An R package (R Development Core Team (2006)) is available and can be obtained directly from http://www.R-project.org that implements a number of the methods discussed in Part I, II, and some of those discussed in Parts III, IV, and V. It also contains some datasets used in the book, and contains a function that allows the reader to easily implement new kernel-based tests and kernel-based estimators.

Exercises appear at the end of each chapter, and detailed hints are provided for many of the problems. Students who wish to master the material are encouraged to work out as many problems as possible. Because some of the hints may render the questions almost trivial, we strongly encourage students who wish to master the techniques to work on the problems without first consulting the hints.

We are deeply indebted to so many people who have provided guidance, inspiration, or have laid the foundations that have made this book possible. It would be impossible to list them all. However, we ask each of you who have in one way or another contributed to this project to indulge us and enjoy a personal sense of accomplishment at its completion.

This being said, we would like to thank the staff at Princeton University Press, namely, Peter Dougherty, Seth Ditchik, Terri O'Prey, and Carole Schwager, for their eye to detail and professional guidance through this process.

We would also like to express our deep gratitude to numerous granting agencies for their generous financial support that funded research which forms the heart of this book. In particular, Li would like to acknowledge support from the Social Sciences and Humanities Research Council of Canada (SSHRC), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Texas A&M University Private Research Center, and the Bush School Program in Economics. Racine would like to acknowledge support from SSHRC, NSERC, the Center for Policy Research at Syracuse University, and the National Sciences

Foundation (NSF) in the United States of America.

We would additionally like to thank graduate students at McMaster University, Syracuse University, Texas A&M University, the University of California San Diego, the University of Guelph, the University of South Florida, and York University, who, as involuntary subjects, provided valuable feedback on early drafts of this manuscript.

We would furthermore like to thank numerous coauthors for their many and varied contributions. We would especially like to thank Peter Hall, whose collaborations on kernel methods with mixed data types and, in particular, whose singular contributions to the theoretical foundations for kernel methods with irrelevant variables have brought much of this work to fruition.

Many people have provided feedback that has deepened our understanding and enhanced this book. In particular, we would like to acknowledge Chunrong Ai, Zongwu Cai, Xiaohong Chen, David Giles, Yanquin Fan, Jianhua Huang, Yiguo Sun, and Lijian Yang.

On a slightly more personal note, Racine would like express his deep-felt affection and personal indebtedness to Aman Ullah, who not only baptized him in nonparametric statistics, but also guided his thesis and remains an ongoing source of inspiration.

Finally, Li would like to dedicate this book to his wife, Zhenjuan Liu, his daughter, Kathy, and son, Kevin, without whom this project might not have materialized. Li would also like to dedicate this book to his parents with love and gratitude. Racine would like to dedicate this book to the memory of his father who passed away on November 22, 2005, and who has been a guiding light and will remain an eternal source of inspiration. Racine would also like to dedicate this book to his wife, Jennifer, and son, Adam, who continue to enrich his life beyond their ken.